Proceedings of the
Seventh ISCA workshop on
Speech and Language Technology in Education
2017

August 25–26, 2017

Djurö, Stockholm, Sweden
Foreword

Welcome to the seventh ISCA workshop on Speech and Language Technology in Education (SLaTE). The workshop has been held bi-annually since 2007, with an additional occasion in 2010, when the workshop Second Language Studies: Acquisition, Learning, Education and Technology was co-organised by SLaTE. However, the SLaTE history goes further back, over the Interspeech special session in 2006, and the InSTIL/ICALL symposium in 2004, to the first meeting Speech technology in language learning (STiLL), held at the island of Marholmen, Sweden in 1998.

When SLaTE is now organised by KTH Royal Institute of Technology, we return to its roots, and much of the original STiLL concept is reused: The workshop is again held in the beautiful Stockholm archipelago, with all meals eaten together and accommodation on-site, and with several presenters and reviewers from the first meeting returning.

On a personal note, STiLL was my very first international workshop experience as a new PhD student and I have very positive recollections of the meeting, with its convenient convenient size, its plenary setting and the communal meals, which allowed all attendees to easily interact with each other. I am therefore very happy to reuse the same concept for SLaTE 2017, with some updates in accommodation comfort and presentation topics (although some of the 1998 paper titles still seem strikingly relevant...). I hope that SLaTE 2017 attendees will be remembering this workshop as fondly in 20 years time.

Continuing on the personal note, I am happy to welcome you to the beautiful island of Djurö, where I spent most of my summers as a child and teenager. Djurö is located 50 km from the Stockholm city centre, and even though it is easily reached by bus or car, it is very far from the city centre in terms of atmosphere. The two islands Djurö-Vindö have 2,000 year-round inhabitants, but about ten times more during the summers, when Stockholmers go out to their summer cottages and boats. The conference site Djuroöniäset is located by one of the gateways to the outer archipelago, and the views from the hotel will be filled by sail boats on a beautiful day (and if you are lucky you might also spot one of the resident seals). The surrounding pine woods are ideal for walks or runs and the water is marvellous (being a mixture of salty and fresh water), albeit somewhat cool for more southern visitors. Which is where the sauna comes in! Enjoy!

These proceedings contain the 33 scientific papers accepted for presentation at SLaTE 2017, and in addition four demo papers. All submissions were assigned three reviewers and all but a handful also received more than two reviews. We are very grateful to all of the reviewers listed overleaf for providing their valuable time. We would also like to express our gratitude to Helmer Strik, Chair of the ISCA SIG SLaTE, for all help and advice during the planning of the workshop.

The papers are included in order of presentation, grouped by different SLaTE themes, including the two special sessions: the CALL shared task challenge and Robot-Assisted Language Learning.

The CALL shared task challenge aims at allowing quantitative comparisons between methodologies used by different researchers on the same task and has been addressed by seven different teams. Robot-Assisted Language Learning (RALL) has emerged during the last decade as a situated interaction alternative to conversational CALL, with different types of more or less antropomorphic robots being used to stimulate language practice, leading to numerous research questions on what the robots should look like and how they should behave in order to provide effective training.

RALL is also the theme of the presentation of the invited speaker, Tony Belpaeme, Professor of Cognitive Systems and Robotics at Plymouth University and leader of the European project Second language tutoring using social robots (L2Tor). We are very proud to have professor Belpaeme as keynote speaker and are grateful to ISCA for the financial support that made it possible to invite him.

We wish you an enjoyable SLaTE workshop.

Olov Engwall, Chair of SLaTE 2017
Review committee

- Abeer Alwan, UCLA
- Paul Baxter, University of Lincoln
- Pierre Badin, GIPSA-lab, Grenoble
- Kay Berkling, DHBW, Karlsruhe
- Jared Bernstein, Analytic Measures Inc.
- Anne Bonneau, LORIA, Nancy
- Lei Chen, Educational Testing Service (ETS)
- Catia Cucchiarini, Radboud University, Nijmegen
- Rodolfo Delmonte, University of Venice
- Maxine Eskenazi, Carnegie Mellon University (CMU)
- Keelan Evanini, Educational Testing Service (ETS)
- Björn Granström, Royal Institute of Technology (KTH)
- Diego Guiliani, Fondazione Bruno Kessler (FBK), Povo
- Rebecka Hincks, KTH Royal Institute of Technology
- Florian Höning, FAU, Erlangen
- Denis Jouvet, LORIA, Nancy
- Tatsuya Kawahara, Kyoto University
- James Kennedy, Plymouth University
- Jacques Koreman, NTNU, Trondheim
- Iolanda Leite, KTH Royal Institute of Technology
- Diane Litman, University of Pittsburgh
- Dominic Massaro, University of California – Santa Cruz
- Nobuaki Minematsu, University of Tokyo
- Berd Mobius, Saarland University
- Omar Mubin, University of Western Sydney
- Patti Price, PPrice.com
- Manny Rayner, University of Geneva
- Preeti Rao, Indian Institute of Technology, Bombay
- Martin Russell, University of Birmingham
- Stefan Steidl, FAU, Erlangen-Nurnberg
- Helmer Strik, Radboud University, Nijmegen
- Isabel Trancoso, IST Lisbon
- Febe de Wet, CSIR, Pretoria
- Preben Wik, Furhat Robotics
## Contents

### Foreword

iii

### Review committee

iv

### Scientific papers

<table>
<thead>
<tr>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was it something I said? Facial Expressions in Language Learning</td>
<td>1</td>
</tr>
<tr>
<td>John Sloan and Julie Carson-Berndsen</td>
<td></td>
</tr>
<tr>
<td>A First Visit to the Robot Language Café</td>
<td>7</td>
</tr>
<tr>
<td>José Lopes, Olov Engwall and Gabriel Skantze</td>
<td></td>
</tr>
<tr>
<td>Measuring Effect of Repetitive Queries and Implicit Learning with Joining-in-type Robot Assisted Language Learning System</td>
<td>13</td>
</tr>
<tr>
<td>Albara Khalifa, Tsuneo Kato and Seiichi Yamamoto</td>
<td></td>
</tr>
<tr>
<td>Emotion and Memory Model for a Robotic Tutor as a Social Partner in a Learning Environment</td>
<td>18</td>
</tr>
<tr>
<td>Muneeb Ahmad, Omar Mubin, Suleman Shahid and Joanne Orlando</td>
<td>36</td>
</tr>
<tr>
<td>Evaluating the Efficiency of Synthetic Voice for Providing Corrective Feedback in a Pronunciation Training Tool Based on Minimal Pairs</td>
<td>25</td>
</tr>
<tr>
<td>Cristian Tejedor-García, David Escudero, César González-Ferreras, Enrique Cámara-Arenas and Valentín Cardenoso-Payo</td>
<td></td>
</tr>
<tr>
<td>Investigation of teacher-selected sentences and machine-suggested sentences in terms of correlation between human ratings and GOP-based machine scores</td>
<td>47</td>
</tr>
<tr>
<td>Yutaka Yamauchi, Jianwei Yue, Kayoko Ito and Nobuaki Minematsu</td>
<td>53</td>
</tr>
<tr>
<td>Towards a Tool for Automatic Spelling Error Analysis and Feedback Generation for Freely Written German Texts Produced by Primary School Children</td>
<td>65</td>
</tr>
<tr>
<td>Ronya Laarmann-Quante</td>
<td></td>
</tr>
<tr>
<td>Transfer Learning based Non-native Acoustic Modeling for Pronunciation Error Detection</td>
<td>71</td>
</tr>
<tr>
<td>Richeng Duan, Tatsuya Kawahara, Masatake Dantsuji and Hiroaki Nanjo</td>
<td>85</td>
</tr>
<tr>
<td>Effects of Educational Context on Learners’ Ratings of a Synthetic Voice</td>
<td>91</td>
</tr>
<tr>
<td>Neasa Ní Chiarráin and Ailbhe Ní Chasaide</td>
<td></td>
</tr>
<tr>
<td>Linguistic Factors Affecting Evaluation of L2 Korean Speech Proficiency</td>
<td></td>
</tr>
<tr>
<td>Seung Hee Yang and Minhwa Chung</td>
<td></td>
</tr>
<tr>
<td>Automatic Characterisation of the Pronunciation of Non-native English Speakers using Phone Distance Features</td>
<td></td>
</tr>
<tr>
<td>Konstantinos Kyriakopoulos, Mark Gales and Kate Knill</td>
<td></td>
</tr>
<tr>
<td>Mispronunciation Diagnosis of L2 English at Articulatory Level Using Articulatory Goodness-Of-Pronunciation Features</td>
<td></td>
</tr>
<tr>
<td>Hyuksu Ryu and Minhwa Chung</td>
<td></td>
</tr>
<tr>
<td>Overview of the 2017 Spoken CALL Shared Task</td>
<td></td>
</tr>
<tr>
<td>Claudia Baur, Cathy Chua, Johanna Gerlach, Manny Rayner, Martin Russell, Helmer</td>
<td></td>
</tr>
<tr>
<td>Strik and Xizi Wei</td>
<td></td>
</tr>
<tr>
<td>Spoken CALL Shared Task system description</td>
<td></td>
</tr>
<tr>
<td>Andrew Caines</td>
<td></td>
</tr>
<tr>
<td>The CSU-K Rule-Based Pipeline System for Spoken CALL Shared</td>
<td></td>
</tr>
<tr>
<td>Nico Axtmann, Carolina Mehrret and Kay Berkling</td>
<td></td>
</tr>
<tr>
<td>The University of Birmingham 2017 SLaTE CALL Shared Task Systems</td>
<td></td>
</tr>
<tr>
<td>Mengjie Qian, Xizi Wei, Peter Jancovic and Martin Russell</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Using an Automated Content Scoring Engine for Spoken CALL Responses: The ETS submission for the Spoken CALL Challenge</td>
<td>97</td>
</tr>
<tr>
<td>Keelan Evanini, Matthew Mulholland, Eugene Tsuprun and Yao Qian</td>
<td></td>
</tr>
<tr>
<td>Deep-Learning Based Automatic Spontaneous Speech Assessment in a Data-Driven Approach for the 2017 SLaTE CALL Shared Challenge</td>
<td>103</td>
</tr>
<tr>
<td>Yoo Rhee Oh, Hyung-Bae Jeon, Hwa Jeon Song, Byung Ok Kang, Yun-Kyung Lee, Jeon-Gue Park and Yun-Keun Lee</td>
<td></td>
</tr>
<tr>
<td>Syntactic and semantic features for human like judgement in spoken CALL</td>
<td>109</td>
</tr>
<tr>
<td>Ahmed Magooda and Diane Litman</td>
<td></td>
</tr>
<tr>
<td>A System for Assessing Children Readings at School</td>
<td>115</td>
</tr>
<tr>
<td>Stefano Artuso, Luca Cristofoletti, Daniele Falavigna, Roberto Grettter, Nadia Mana and Gianluca Schiavo</td>
<td></td>
</tr>
<tr>
<td>Automatic Assessment of Children’s L2 Reading for Accuracy and Fluency</td>
<td>121</td>
</tr>
<tr>
<td>Kamini Sabu, Prakhar Swarup, Hitesh Tulsiani and Preeti Rao</td>
<td></td>
</tr>
<tr>
<td>Improving fluency of young readers: introducing a Karaoke to learn how to breath during a Reading-while-Listening task</td>
<td>127</td>
</tr>
<tr>
<td>Erika Godde, Gérad Bailly, David Escudero, Marie-Line Bosse, Maryse Bianco and Coriandre Vilain</td>
<td></td>
</tr>
<tr>
<td>Modeling Discourse Coherence for the Automated Scoring of Spontaneous Spoken Responses</td>
<td>132</td>
</tr>
<tr>
<td>Xinhao Wang, Keelan Evanini, Klaus Zechner and Matthew Mulholland</td>
<td></td>
</tr>
<tr>
<td>CARAMILLA - Speech Mediated Language Learning Modules for Refugee and High School Learners of English and Irish</td>
<td>138</td>
</tr>
<tr>
<td>Emer Gitmartin, Jaebok Kim, Alpha Diallo, Yong Zhao, Neasa Ni Chiarain, Ketong Su, Yuyun Huang, Benjamin Cowan and Nick Campbell</td>
<td></td>
</tr>
<tr>
<td>An attention based model for off-topic spontaneous spoken response detection</td>
<td>144</td>
</tr>
<tr>
<td>Andrey Malinin, Kate Knill, Anton Ragni, Yu Wang and Mark Gales</td>
<td></td>
</tr>
<tr>
<td>Empirical Evaluation of the Communicative Effectiveness of an Automatic Speech-to-Speech Translation System</td>
<td>150</td>
</tr>
<tr>
<td>Xinhao Wang and Keelan Evanini</td>
<td></td>
</tr>
<tr>
<td>Detecting listening difficulty for second language learners using Automatic Speech Recognition errors</td>
<td>156</td>
</tr>
<tr>
<td>Maryam Sadat Mirzaei and Kourosh Meshgi</td>
<td></td>
</tr>
<tr>
<td>Predicting Clinical Evaluations of Children’s Speech with Limited Data Using Exemplar Word Template References</td>
<td>161</td>
</tr>
<tr>
<td>Gary Yeung, Amber Afshan, Kaan Ege Ozgun, Kantapon Kaeutip, Steven M. Lulich and Abeer Alwan</td>
<td></td>
</tr>
<tr>
<td>Deep Context Model for Grammatical Error Correction</td>
<td>167</td>
</tr>
<tr>
<td>Chuan Wang, Ruobing Li and Hui Lin</td>
<td></td>
</tr>
<tr>
<td>Studies of a Self-Administered Oral Reading Assessment</td>
<td>172</td>
</tr>
<tr>
<td>Jared Bernstein, Jian Cheng, Jennifer Balogh-Ghosh and Elizabeth Rosenfeld</td>
<td></td>
</tr>
<tr>
<td>Speech perception training as a serious game in the EFL classroom</td>
<td>177</td>
</tr>
<tr>
<td>Mariza Hommel</td>
<td></td>
</tr>
<tr>
<td>Menusigne: A Serious Game for Learning Sign Language Grammar</td>
<td>181</td>
</tr>
<tr>
<td>Manny Rayner, Irene Strasly, Nikolaos Tsurakis, Johanna Gerlach and Pierrette Bouillon</td>
<td></td>
</tr>
<tr>
<td>Demo papers</td>
<td>187</td>
</tr>
<tr>
<td>Lärka: an online platform where language learning meets natural language processing</td>
<td></td>
</tr>
<tr>
<td>Ildiko Pilan, David Alfter and Elena Volodina</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Authors</td>
</tr>
<tr>
<td>---------------------------------------------------------------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>New Features and Effectiveness of Suzuki-kun, the First and Only Prosodic Reading Tutor of Tokyo Japanese</td>
<td>Nobuaki Minematsu and Daisuke Saito</td>
</tr>
<tr>
<td>Development and Maintenance of Practical and In-service Systems for Recording Shadowing Utterances and Their Assessment</td>
<td>Junwei Yue, Daisuke Saito, Nobuaki Minematsu, Yutaka Yamauchi and Kayoko Ito</td>
</tr>
<tr>
<td>A pipeline for automatic assessment of foreign language pronunciation</td>
<td>Aku Rouhe, Reima Karhila, Heini Kallio and Mikko Kurimo</td>
</tr>
</tbody>
</table>

**Author Index** 191
Was it something I said? Facial Expressions in Language Learning

John Sloan, Julie Carson-Berndsen

ADAPT Centre, School of Computer Science, University College Dublin, Ireland

Abstract

This paper describes an experiment to evaluate facial expressions of an animated avatar as a means of providing feedback to non-native English learners on language production. The aim of the study was to ascertain whether native and non-native English speakers interpret and respond differently to facial expressions and whether such expressions have a role to play as feedback in emerging language learning technologies. Native English speakers and non-native English learners took part in an experiment where facial expressions were presented as a response to their textual input sentences and were asked for their interpretation of the change in expression (or otherwise). Furthermore, it was investigated to what extent non-native learners subsequently altered their language behaviour in line with their perceived interpretation of the expression. The majority of non-native learners attributed a change in facial expression, where the avatar looked away, to errors in language production in the preceding sentence and they reduced the syntactic complexity of the following sentence accordingly. The results underpin the potential of facial expressions as a feedback mechanism for language learning and the insights will now be deployed in an effective and engaging personalised e-learning language platform.

Index Terms: second language acquisition, feedback, facial expressions, avatars

1. Introduction

Corrective feedback in second language acquisition (SLA) has been shown to improve learners’ retention of vocabulary and grammar [2]. Recasts, clarification requests and comprehension checks are among the verbal feedback forms which have proven most effective [3,4], however, there has been no real consensus on which of these mechanisms produces the optimal learning outcome. In addition, the impact of non-verbal, paralinguistic feedback has received very little attention [2]. With the growth of Computer Assisted Language Learning (CALL), the question of when and how to provide feedback to learners is of immediate concern, especially when the effectiveness of current CALL systems is limited [5]. While e-learning can provide learners with an increase in content, interactions and a variety of methods of study, the distance between learners and instructors (or indeed other learners) can lead to a shallow experience with inaccurate, ineffective or even a lack of feedback.

One area of CALL which has received a lot of recent interest is the use of robots which interact with learners in the same physical space. These robots can be controlled remotely [6] or automatically [7]. Advantages offered by this form of interaction include increased engagement and a richer educational experience. Improvements in performance on tasks involving puzzle-solving in adults [8] and prime number learning in children [9] have been demonstrated. However, the body of research in this area is still small and is based on short-term studies. Furthermore, several negative effects have also been identified. These include deception, privacy issues, lack of accountability and, particularly in the case of children, attachment to the robot, loss of human contact and impeded social skill development [9,10].

In contrast to physically embodied robots, screen technologies are widely used to deliver CALL content. The recent global proliferation of internet and mobile technology (43% own a smartphone, 76% report using the internet [11]) means through the screen, CALL is readily accessible. Over the past decade, the most popular language learning applications have reached tens of millions of registered users [12]. For SLA, screen-based technologies offer advantages over other interactive media due to the privacy afforded to the learner in a task where making errors is common. Success in acquiring a second language requires not only considerable time and effort, but also the ability to learn from these errors. However, Cross Linguistic Interference (CLI) can lead to the repeated production of ‘basic’ mistakes which often results in anxiety, as the learner fears appearing foolish in front of peers [13]. In the early stages of language learning, anxiety can be low, but may grow through negative learning experiences emanating from negative perceptions of others [14,15]. Educational gains from short-term interaction with robots might not continue in the long-term if users’ frequent errors are salient to others in the shared physical space.

Using facial expressions as the means of providing feedback through a screen offers an opportunity to reach large numbers of users while increasing engagement and pedagogical effectiveness. Current, popular screen-based
CALL applications (e.g. ‘Duolingo’, ‘Babbel’) typically provide feedback through simple video/audio signals which are intuitively understood by the user: green vs. red colours, a progress bar moving forward vs. backward, pleasing ping sounds vs. harsh klaxons. Facial expressions offer a potentially more powerful medium to provide feedback, as many expressions can be intuitively understood [16] and we have an innate preference to pay attention to faces and face-like stimuli [17]. Additionally, recent human-computer interaction studies have shown our lexical and gestural alignment with human-like avatars is comparable to that with real humans [18,19]. Through the medium of a screen, which can be kept relatively private, and the inaudible nature of facial expressions, the user is afforded a sense of anonymity [20] while engaged in the interaction. These factors could aid retention rates in language learning applications where dropout rates are high, e.g. 54% use Babbel for less than a month [21].

This paper presents an experiment to evaluate the use of an avatar’s facial expressions as a potential feedback mechanism for language learning. Based on first-hand experience of the first author teaching Korean-learners of English for many years, a personalised e-learning platform is under development which aims to enhance non-native learner engagement and performance, and overcome anxiety issues associated with errors. The aim is to create an environment where language learners can practise producing samples of language while receiving immediate, accurate feedback which serves the function of guiding the learner to produce well-formed utterances. The platform currently uses the Mery avatar [1] with animations which have been created specifically for the purpose of providing expressive feedback. The animations run in modern web browsers at 101ps in 400x320px resolution. This produces smooth, realistic animations with the result akin to an animated movie. The avatar was chosen in preference to a life-like model due to unnatural movements of life-like models causing feelings of unease in users [22], the perceived intensity of expressions in animated avatars are higher than life-like models [23], and the avatar offers more flexibility in the creation of animations.

The remainder of this paper is structured as follows. Section 2 focusses on the facial expressions presented to the participants and the hypotheses being tested. The experiment to evaluate the facial expressions is described in section 3. Section 4 presents the results, with the discussion following in Section 5. Finally, Section 6 draws some conclusions and outlines future work.

2. Facial Expression Feedback Hypothesis

The research presented in this paper represents the first step in the investigation of the interpretation and response of non-native learners (NNS) and native speakers (NS) to changes in facial expression of an animated avatar. The sequence of expressions (e1 to e6), as depicted in Figure 1, are the respective responses to six sentences (s1 to s6) input by the user to the system. The focus will be on reactions to one specific, clear and distinct change in expression - from smiling and looking straight ahead to slightly frowning and looking down and to the side (e2 → e3 in Figure 1). The change from eye-contact to averting eye-contact is based on task-oriented studies where gaze aversion has been shown to correlate with the cognitive difficulty of the task [24,25]. Therefore, the avatar looking away following an input sentence indicates an increased cognitive load.

The broad hypothesis to be tested is drawn from observances of NNS in interaction with NS, namely:

There is a difference between NS and NNS in how they interpret and respond to facial expressions of a NS interlocutor.

NNS, particularly proficient speakers, often appear to reduce the length and complexity of utterances when met with a NS facial expression which would be typically described as confused. This may be due to a recognition by the NNS that their production is causing additional cognitive effort on the part of the NS which must be reduced.

In the experiment presented in the next section, the animated avatar was used in place of a NS interlocutor and participants were informed that the avatar could read and understand English sentences. The broad hypothesis can then be refined to refer to facial expressions of this avatar. This hypothesis was broken down into four sub-hypotheses.

- **H1:** There is a difference between NS and NNS in the emotion attributed to the avatar’s e2 → e3 change in expression.
- **H2:** There is a difference between NS and NNS in the reason given for this change in expression.
- **H3:** There is a difference between NS and NNS in the complexity of sentence produced following an e2 → e3 change in expression.
- **H4:** In the complexity of sentence produced following an e2 → e3 change in expression, there is a difference between NNS who are sensitive to small changes in the avatar’s expression and NNS who are not.

The experiment to test these sub-hypotheses is now described.

3. Experiment

Two groups were selected for the experiment: native English speakers, and intermediate and above non-native learners of English. Participants (N=57) were gathered through postings on internet cafes for English language learners and snowball sampling. The experiment consisted of 2 parts: questions on emotions and English, and typing sentences to the avatar. All participants were aged 18+ and successfully passed part 1 of the experiment; this evidenced accurate recognition of images of human facial expressions and a command of English sufficient to partake in part 2. The sample contained 36 NNS and 21 NS, with participants indicating their nationality and L1 prior to beginning the experiment. Five participants were excluded from analysis: 2 incorrectly responded to questions in part 1 and 3 did not fully complete the sentences in part 2.

3.1. Experiment Design

A web application was developed for the experiment and hosted at emotionandlanguage.ucd.ie. Upon signing up for the research, participants were sent to the website where they were asked to complete a series of tasks. These tasks involved answering multiple choice questions, typing sentences and watching the avatar’s subsequent expressions. These tasks were split into 2 parts which each took approximately 5-10 minutes to complete. Participants completed the experiment at a time and place of their choosing.
3.1.1. Part 1: Questions on English and emotions

Participants answered 15 multiple choice questions before interacting with the avatar. The first 10 showed a photograph of a person expressing an emotion with a choice of 4 words to describe the image. Participants chose one of the 4 options.

Figure 2: Question on facial expressions and emotion.

The purpose of the 10 expression questions was threefold: draw participants’ attention to facial expressions, ensure NNS participants understood the vocabulary used to describe expressions and emotions, and confirm that participants gave typical responses to facial expressions. Some of the expression images were shown alongside only one word which matched (e.g. a photo of a smiling girl with the choices: ‘angry’, ‘happy’, ‘sad’, ‘tired’), while others contained multiple (e.g. a man frowning and rubbing his eyes with the choices: ‘excited’, ‘tired’, ‘sad’, ‘worried’). Five further questions on vocabulary and grammar were then presented. These ensured that NNS participants understood the English needed to give reliable answers to the questions in part 2.

3.1.2. Part 2: Sentences and Expressions

Our 4 sub-hypotheses were tested in part 2. Participants were introduced to the avatar, and after a short practice session, entered 6 sentences, one-by-one, about their favourite holiday. Regardless of linguistic content, the avatar changed expression after each sentence in the sequence shown in Figure 1. Participants were given no information as to the reason for the avatar’s expression changes. Our focus was on the participants’ response to the avatar’s change (e2 → e3) after they entered sentence s3.

Figure 3: View of the main interface where participants input sentences and monitor the avatar’s expression (e2).

For each sentence input, participants were shown the avatar, the prompt and a text box with cursor focus where they could type (Figure 3). Two seconds after a sentence was entered, the avatar began to move its head and eyes to simulate reading (Figure 4). Note that the movements of the avatar between all of the expressions are smooth.

Figure 4: View of the avatar ‘reading’ s3.

After ‘reading’ the sentence, the avatar either returned to looking straight ahead, or, in the case of interest after s3, looked down and to the left (Figure 5). Depending on the length of the input sentence, each animation lasted between 7 and 12 seconds. The facial expression changed in five of the six movements, with only the final sentence (s6) eliciting no change.

Figure 5: View of the e3 expression after ‘reading’ s3.

In this case, after the avatar ‘read’ s3 and changed expression, participants were asked to answer two questions: the first (H1) involved describing the avatar’s emotion, and the second (H2), the reason behind this change of expression. The questions appeared 4 seconds after the change in expression and covered the avatar and the typed sentence on the screen as depicted in Figure 6.

Figure 6: Question box for H2 which appeared after the e2 → e3 expression change.

3.1.3. Change in complexity

H3 and H4 were tested by comparing the complexity of sentences produced immediately before and after e3 → e4. This involved scoring the complexity of s3 and s4 for both NS and NNS. While measures of syntactic complexity as development markers in L1 children, such as Developmental Sentence Scoring [26] and Index of Productive Syntax [27], and the extension of such systems to measure L2 proficiency [28,29] have been developed, there is no single reliable and valid measure for both NS and NNS complexity. In addition,
the NNS in this experiment came from a variety of countries and native languages (L1), which compounds the problem, e.g. for an NNS of a language which shares the head-first structure of English, a relative clause may be considerably easier to form than for an NNS whose native language is head-last.

With this in mind, it is still possible to formulate a general score of complexity (c-score) which correlates with human judgement. The formula used for this experiment is based on a small number of widely used factors which have been identified as useful markers in the literature and recent research on automatic syntactic complexity analysis [30]. These are: length of sentence, number of t-units (main clause with attached subordinate clauses [31]) per sentence, amount of subordination and sophistication of Noun Phrases (NP). For this experiment, the complexity score of each sentence was calculated using the following formula:

\[ c\text{-score} = tuM(L / 3 + 2sC + cNP) \]  

(1)

where

- \( L \) = length of sentence in words
- \( sC \) = number of subordinate clauses
- \( cNP \) = number of complex NP (NP including head noun and an adjective or prepositional phrase)
- \( tuM = t\text{-unit multiplier} \) (1 – \( \ln(tu / 3) \)) where \( tu \) is the number of t-units

The weightings of each variable were determined based on theoretical ideals with testing and evaluation on sentences produced by NNS from a previous study. Therefore, subordinate clauses, which are often difficult to accurately produce for NNS were assigned a higher weighting than complex noun phrases. The additional length from higher t-units per sentence was balanced through ‘\( tuM \)’, which decreases exponentially from 1 in the order [1, 0.77, 0.63, 0.54 ...]. An example calculation on 2 contrived sentences is shown below (\( sC \) is indicated with ‘\{\}’, \( cNP \) with ‘\_’, and t-units are split with ‘\|’):

\[
\begin{align*}
\text{s3: } & "(\text{Because heavy rain was falling everyday}, my sister, \text{who hadn't brought a raincoat}, didn't enjoy the holiday.}.” \\
\text{s3 c-score:} & \quad 1*(19/3 + 0 + 1) = 11 \\
\text{s4: } & "\text{We stayed inside and didn't travel, but she went out on the third day and we had some fun.”} \\
\text{s4 c-score:} & \quad 0.63*(19/3 + 0 + 1) = 4.62 \\
\text{s3-s4 change in c-score:} & \quad 4.62-11 = -6.38
\end{align*}
\]

To assess the validity of the c-score metric, all s3 and s4 sentences were independently scored for complexity by an expert with an MA in Linguistics and 10+ years experience teaching ESL. The scorer was informed of the factors taken into consideration for complexity, but not told of the weighting or method of calculation. The c-scores and the independent annotator’s scores were normally distributed and exhibited a strong correlation \( r(55) = .59, p < .001 \).

3.1.4. Sensitivity

H4 investigates two populations of English learners: those with high levels of attention or sensitivity to facial expressions, and those with lower levels. We split our sample of English learners into ‘sensitive’ (\( N=11 \)) and ‘not sensitive’ (\( N=25 \)) groups based on the accuracy with which they recognised the avatar’s changes of expression \( \text{e4 } \rightarrow \text{ e5 } \) and \( \text{e4 } \rightarrow \text{ e6 } \). \( \text{e4 } \rightarrow \text{ e5 } \) displays a small change in expression, with a slight lowering of the edges of the mouth and eyebrows; \( \text{e5 } \) and \( \text{e6 } \) exhibit the exact same expression. Only participants who correctly identified both changes were assigned to the ‘sensitive’ group. The change in \( \text{s3 } \rightarrow \text{ s4 } \) complexity between the groups was measured as an independent variable, ‘Sensitivity’.

4. Results

In this section, results for each of the sub-hypotheses (H1 to H4) are presented in turn.

H1: Change in expression

A Fisher’s exact test [32] was performed to examine the relationship between NS and NNS in their identification of emotion in the avatar’s expression after s3. No difference was found between the groups (\( p=.6 \)). Both rated the expression of the avatar most frequently as ‘confused’ followed by ‘thoughtful’ as depicted in Figure 7.

![Figure 7: Identification of emotion in expression e2 → e3.](image)

H2: Reason for change in expression

A significant difference was found in the reasons given for the avatar’s change in expression after s3 (\( p=.002 \)). Figure 8 shows the majority of NNS (53.8%) attributed the change to their mistakes in vocabulary or grammar in s3. In contrast, NS were unsure of the reasons for the expression change, selecting ‘don’t know’ more frequently than NNS (47.6%).

![Figure 8: Reasons given for e2 → e3 change in expression.](image)
H3: Change in complexity

NNS significantly changed the complexity of their sentences after s3 ($M=-1.46$, $SD=2.85$) in comparison to NS ($M=15$, $SD=2.45$), $t(45.94)=2.24$, $p=.03$. NNS showed a propensity to reduce sentence complexity in s4, while NS did not, as shown in Figure 9.

![Boxplot of change in sentence complexity (c-score) from s3 to s4.](image)

**Figure 9:** Boxplot of change in sentence complexity (c-score) from s3 to s4.

H4: Change in complexity by sensitivity

No difference was found between 'sensitive' ($N=11$, $M=-1.77$, $SD=2.87$) and 'not sensitive' ($N=25$, $M=1.33$, $SD=2.89$) in their change in sentence complexity after s3, $t(19.3)=-0.42$, $p=.7$. This is shown in figure 10.

![Boxplot of change in learner s3-s4 sentence complexity by sensitivity to s4-s6 expression changes.](image)

**Figure 10:** Boxplot of change in learner s3-s4 sentence complexity by sensitivity to s4-s6 expression changes

5. Discussion

The results of the study indicate that there is a difference in how NS and NNS interpret and respond to an e2 → e3 change in facial expression of an animated avatar as feedback to their typed sentences. NNS tend to view the change as resulting from their mistakes in vocabulary or grammar of their previous sentence, while this is absent in NS. The effect of this difference in interpretation is manifest in s4, where NNS reduce the sentence complexity, with the sample of NS showing no such reduction. No statistically significant difference was observed between the NS and NNS groups in the emotion attributed to this change in expression. Furthermore, there was no change in complexity of 'sensitive' vs. 'not sensitive' NNS.

There are several limitations of the experiment that could be addressed in future studies. The sample of NNS included a mix of nationalities and L1. There may be cultural differences and varying CL1 effects between these groups in their response to facial expressions [33]. Studying independent populations who share nationality and L1 would allow us to identify any such differences. Three participants selected ‘happy’ to describe the e2 → e3 change in expression, which possibly indicates a misunderstanding, a lack of concentration, or a browser issue at the time of participation (despite prior testing). As participants undertook the experiment online at a time and place of their choosing, we cannot be certain as to the conditions under which the tasks were completed. Inviting participants into the lab and giving instructions in person may increase concentration and understanding in this regard, but it would impact privacy and anonymity.

A dissatisfaction with the limitations of the screen in creating ‘deeper’ interactions has, in part, spurred research into using robots in CALL. However, some of the robots presently used in language learning are unable to display changes in facial expression (e.g. Nao [7]). This paper makes a valuable contribution to the field of emerging language learning technologies by demonstrating that there is a great deal of potential in using expressions as a mechanism to provide feedback with currently available screen technologies. This is also a feature which could be incorporated into developing robotic systems (e.g. Furfat [34]) in the CALL field.

6. Conclusions and Future Work

The results presented in the previous section offer intriguing possibilities for the potential of using an avatar’s change in facial expressions to guide language learners toward advantageous learning strategies. The e2 → e3 expression could be used to indicate to learners that the sentence they have just produced is not well-formed. The facial expression does not explicitly inform the user that what they produced is incorrect in the same way a red cross or harsh klaxon sound does, it only indicates that they are causing the listener some difficulty in understanding. It is then up to the user to respond, and as evidenced in this study, the reaction tends towards a reduction in sentence complexity. The extent to which this is manifest in the L2 production of learners with different abilities, cultural backgrounds and L1 is an avenue of research which has the potential to provide fine-grained insights into the language learning process. The growing corpus of learner sentences, errors and corrections will also allow for a gradual increase in automation of corrections. What will initially be a manual process involving a NS quickly assessing sentences and selecting an appropriate emotional response or correction, will develop into a semi-automated process where NS judgement and correction is needed only for sentences which contain structures/ambiguities not easily categorised by the corpus-based language model. On the learner side, by supplementing the e2 → e3 ‘error’ expression with positive changes for well-formed sentences, intermediate language learners may be guided by facial expressions towards advanced levels and more confident production. This is the premise which underpins the novel, personalised e-learning platform currently under development at UCD.
7. Acknowledgements

This study is supported by the Science Foundation Ireland ADAPT Centre for Digital Content Technology Research (Grant 13/RC/2106). We would like to thank Frances Galliford for scoring the sentences used in the analysis and Benjamin R. Cowan for his comments on the paper.

8. References

A First Visit to the Robot Language Café

José Lopes, Olov Engwall, Gabriel Skantze

KTH Royal Institute of Technology, Sweden
jdlopes@kth.se, engwall@kth.se, skantze@kth.se

Abstract

We present an exploratory study on using a social robot in a conversational setting to practice a second language. The practice is carried out within a so called language café, with two second language learners and one native moderator; a human or a robot; engaging in social small talk. We compare the interactions with the human and robot moderators and perform a qualitative analysis of the potentials of a social robot as a conversational partner for language learning. Interactions with the robot are carried out in a wizard-of-Oz setting, in which the native moderator who leads the corresponding human moderator session controls the robot. The observations of the video recorded sessions and the subject questionnaires suggest that the appropriate learner level for the practice is elementary (A1 to A2\(^1\)), for whom the structured, slightly repetitive interaction pattern was perceived as beneficial. We identify both some key features that are appreciated by the learners and technological parts that need further development.

Index Terms: conversational L2 practice, human-robot interaction, computational paralinguistics

1. Introduction

Developing conversational skills in a new language can certainly be a challenging task for an adult second language (L2) learner. Other skills, such as vocabulary, grammar and written and oral comprehension, can be readily practiced, in the L2 classroom, or with self-study exercises and by passive exposure. In contrast, it is difficult to allocate an adequate amount of time for individual practice of conversational skills in an L2 classroom with many students, and learners’ progress is therefore very much dependent on their practice outside the classroom. For adult immigrants learning Swedish, this has been shown to be difficult, for two reasons:

Firstly, Sweden has welcomed a number of refugees over the past decades that is very large in proportion to the number of inhabitants (e.g., in 2007, the town of Södertälje, with 80,000 inhabitants, welcomed more Iraqi immigrants than the US and Canada together), and many of these immigrants live in communities where they mainly interact with compatriots. Secondly, almost the entire Swedish population is proficient in English as an L2, with 86% self-reporting that they can hold a conversation in English [1], and it is hence extensively used as lingua franca for social conversations.

However, even if English is widely accepted for social interactions, integration in society in general and on the job market in particular is very dependent on the mastery of Swedish. To complement the traditional classroom teaching, many societal agencies (NGAs, churches, universities) therefore arrange Language Cafés to help learners practice conversation.

A Language Café is an open gathering to which native and non-native speakers convene spontaneously to engage in open-

\(^1\)European Common Language Reference Framework

topic social conversations for an hour or so, with the explicit aim of maintaining the conversation in the target language, even in case of linguistic problems. The native speakers are often given the extra responsibility of initiating the conversations and assisting the L2 learners with linguistic problems, but they nevertheless have the role as equal conversational partner, rather than as a teacher. The language cafés therefore constitute a relaxed setting, in which errors, hesitation and slow interaction are allowed to a far greater extent than in everyday interaction. However, there is often a shortage of native speakers for the language cafés, in general, and in immigrant communities and at asylum residences in particular. Technology-enhanced learning solutions are therefore an attractive alternative to be able to provide more L2 learners with practicing opportunities.

Over the last decades, computer-assisted language learning software has been developed to practice communication skills. Systems such as SPELL [2], DEAL [3], The Tactical Language and Culture Training System [4] and Dansksimulatoren [5] allow learners to practice spoken interaction in 3D virtual reality environments, using respectively, task scenarios such as ordering meals at the restaurant, bargaining at a flea market, interacting with Iraqi citizens as a US army soldier, and exploring a Danish community as an immigrant. While these systems may provide valuable training time, the on-screen virtual setting fails to practice real world skills such as multimodal turn-taking and communication signals.

Within a newly initiated research project on collaborative robot-assisted language learning (RALL), we have therefore started to explore how a social anthropomorphic robot could be used as native speaker in a language café setting with two L2 learners of Swedish. Language café interactions can vary greatly depending on the learners’ level of Swedish, but the concept of mostly meeting with new persons of unknown background and linguistic level, entails that the topics of the conversations are often rather limited and repetitive. They focus on personal aspects (occupation, family, hobbies, food and travel preferences) or comparisons between different countries, cultures and languages. While this may perhaps be a downside for the returning visitor to a language café, it makes the use of a robot as a native conversation leader more achievable: since the expected number of topics is limited, the robot’s dialogue system can control the interaction more and have better chances of predicting answers from the conversational partners. This is important as the input from the learners can be expected to contain a substantial amount of errors of vocabulary, grammar and pronunciation, which makes the task for the automatic speech recognition (ASR) challenging if the topics are wider. The language café setting comes with an additional benefit for RALL, since the two peers may support each other, both to correct erroneous utterances that could not be understood by the robot, and to monitor if a non- or mis-understanding by the robot of something that the peer says is due to actual linguistic problems or in fact to technological shortcomings of the system. This is
also of importance, as erroneous feedback from a CALL system, either explicit or in terms of misrecognition of a correctly produced utterance, may be detrimental for learning [6].

Developing a robot that can participate in a multi-party, multimodal, open-domain conversation is, to our knowledge, a new challenge in RALL. In this paper we present a first step in that direction recording recorded a corpus of language café interactions, both with a human native speaker and a (wizard-of-Oz controlled) Furhat robot head leading the conversation.

We are aiming to address four main questions in this exploratory study: 1) to compare the interactions in the two conditions in general, and the effect of the order in which they interacted with the human or the robot moderator. 2) to assess if there is an ideal level of the learners for this type of exercise. 3) to find out technological improvements required for a successful autonomous interaction of the robot in this setting. 4) on the meta level, to investigate to what extent the experience of the wizard (confederate versus naïve) influenced the interaction.

2. Related Work
Educational robot tutors for language learning is still a very recent research field, which has emerged during the last decade (c.f. [7] and [8] for surveys). Previous work in robot-assisted language learning has almost exclusively targeted school and pre-school children and have mainly focused on the motivational aspects of language learning with robots. Examples show that children’s English speaking skills as well as confidence and motivation increased significantly through interaction with robot tutors [9], and that robots can be used in game settings with children to teach a foreign language [10].

The robots have either employed a video screen to allow for telepresence display of a human teacher, or have been toy-like robots, such as Philips iCat [10] or the yellow snow-man Keepon [11]. The recently launched European project L2tor uses humanoid Nao robots that interact with pre-school children one-to-one to teach them English, Dutch or German [12]. Nao robots can use human-like body gestures, but can display neither extra-linguistic human facial expressions conveying emotion and turn-taking signals (e.g., smile, eye and eyebrow movements) nor lip movements to give linguistic information, which is fundamental in spoken L2 learning. The anthropomorphic robot Furhat (c.f. Section 3) used in the current study has clear benefits in that it can mimic a human interlocutor.

3. Experimental Setup
The experimental setup for this data collection is shown in Figure 1. The full scene was recorded with one digital video camera and in addition each participant was recorded individually, using a head-mounted microphone for the audio and a dedicated GoPro camera to record facial expressions. The individual audio and video recordings will be used for future training of detection of linguistic errors, hesitation and motivational state. The robot’s actions were also logged for further analysis.

Three participants were invited to each experiment session; one fluent Swedish speaker and two L2 learners of Swedish; and were told that they were going to have a language café type of interaction. For those who were not familiar with language cafés, they were instructed to engage in a social conversation on any topic of their choice. The session consisted of two 15-minute interactions, either starting with the L1 speaker as moderator and then switching to Furhat as moderator, or vice versa. The moderator was instructed to lead the conversation and to help the L2 learners when they so require, but the L2 speakers were also free to initiate topics of their interest and to assist each other with linguistic difficulties.

3.1. The Furhat social robot
The robotic head used, Furhat [13], can display a wide variety of human face gestures, as it uses a computer-animated face projected on a 3D mask (c.f. Figure 1). As the neck is fitted with a motor-servo, Furhat can also nod and turn the head, which has been shown to be important in three-party interactions [14]. The robot is controlled by the IrisTk framework [15], which interfaces to components for text-to-speech synthesis, facial animation, automatic speech recognition and interaction event tracking.

Figure 1: Experimental Setup and the Furhat robot.

3.2. Wizard-of-Oz control
The Furhat robot was controlled by the same native speaker who acted as moderator in the setting with a human moderator. Whenever the moderator was not a confederate, (s)he was instructed separately before the session and was given a (very) short introduction to the wizard interface. In this interaction, the wizard was sitting in an adjacent, sound isolated control room, and could observe the language café scene with Furhat and the two learners through a VoIP connection with sound and video. The wizard could control the robot’s interaction in the following manner:

Spoken output: Through 1) Buttons with pre-generated greetings, questions and responses that are commonly occurring in language cafés sessions; or 2) ASR input allowing the wizard to provide other utterances. These were first recognised using Google Cloud ASR and presented on the wizard’s screen. When the recognized text corresponded to the wizard’s intention, (s)he pressed a button and the utterance was synthesized and spoken by Furhat. While waiting for the utterance recognised, Furhat would signal the intention to speak by uttering a filled pause when the ASR was activated by the wizard.

Non-verbal communication (controlled by buttons): 1) Head turns to address a particular person to indicate whom a question was directed to. This active action by the wizard overrules the automatic direction-sensitive turning of the robot’s head towards the current speaker. 2) Non-verbal feedback, e.g., “Mhm”, to backchannel when someone else is speaking.
4. Data Collection

The subjects for the experiment were recruited from the KTH language cafés, on-line recruitment platforms and confederates, and were told that they were going to be part of an experiment that aimed to collect data to improve a robot that can participate in language cafés. For the L2 learners we recommended a level above A2, but everyone who was interested in practicing Swedish was welcome. Each L2 participant was allowed to take part in the experiment once, whereas the L1 speaker could act as moderator several times (and three of them did).

Once both conversations were concluded the participants were asked to fill in a short questionnaire; one for the L2 speakers and one for the L1 moderator. The questions to the L2 speakers were (unless otherwise stated, answers were on the scale 1-5): 1) Have you ever been to a language café before? [Yes/No] 2) Self-rated proficiency in Swedish. [Beginner, Elementary, Intermediate, Proficient] 3,5) I feel that the [human/robot] moderator understood what I was saying. 4,6) I understood what the [human/robot] moderator was saying. 5,7) I feel that the [human/robot] moderator helped me in the interaction. 8) Compared to the interaction with the human moderator, that with the robot was [Score: Much worse (1) to Much better (10)]. 9) What features need to be improved in the robot for it to be useful in a language café setting? [Speech understanding, Speech synthesis, Response time, Interaction style, Choice of topics, Other] 10) What features did you like in the interaction with the robot moderator? 11) Additional comments about the practice. The questions to the L1 moderator focused on the perceived differences between being on-site and mediated through Furhat, and on suggestions for improvement. These questions were of a more qualitative nature and are hence analyzed together with the general initial observations in Section 5.2.

We performed 11 experiment sessions with a total of 22 L2 Swedish speakers (7 female and 15 male; average age 29.1 years) and 6 different moderators (3 female and 3 male; average age 36.2 years). For their participation, subjects received a cinema ticket as compensation. Of the 11 experiment sessions, 6 started with a human moderator (condition 1) and 5 started with Furhat (condition 2); 8 dialogues had a male moderator and 3 a female moderator. One of the moderators was a non-native, but fluent speaker of Swedish. She and one other recruited moderator had previous experiences of moderating language café sessions. The 22 participants had 14 different mother tongues; some of them were bilingual and all were proficient speakers of English (which influenced strategies employed to handling linguistic uncertainty, as discussed below).

5. Analysis

The number of varying parameters (L2 level of subjects, session order, wizard experience, familiarity with the language cafés, subject and moderator gender) is too large to allow for a rigorous statistical analysis of the impact of each parameter separately. However, some observations may nevertheless be made, to guide the future development, based on the user questionnaires and/or the video recordings.

5.1. Questionnaires

Table 1 summarizes the comparisons of the questionnaire results. Unsurprisingly, in general, the paired t-test showed there were significant differences in the way the participants felt understood by the human moderator compared to the robot moderator and how they got help from him/her. However, there were no significant differences in the way they could understand the robot and human moderators. When comparing the interaction with the robot moderator to that with the human dito, the average score was 3.9 on a scale from 1 ("Much worse") to 10 ("Much better"), indicating that the interaction with the robot was perceived as inferior to that with the human, but not severely so.

We will divide our analysis in three dimensions that could possibly affect the interaction and experience of it: 1) the self-reported language proficiency, to see if a user proficiency level is more suitable than others, 2) if the users have previous experience of language cafés, to see if this influenced the expectations and 3) if the wizard is a confederate or a recruited subject, to see if previous experience of wizarding tasks affects the overall outcome of interaction.

5.1.1. Self-reported proficiency

The levels that could be reported in the questionnaire varied from Beginner to Proficient, and 4 participants reported that they were beginners, 6 that their knowledge of Swedish was elementary, 11 intermediate and 1 proficient (who was discarded in this analysis). Table 1 shows how the questionnaire answers related to the proficiency level.

The scores related to proficiency level can be analyzed with paired t-tests and summarized as: while intermediate learners preferred the interaction with the human moderator (comparison score 3.6, significant differences in the feeling of being understood by the moderator, understanding and being helped by the moderator), elementary level learners (comparison score 4.2, non-significant differences in understanding the moderator and being helped by him/her) and in particular beginners (comparison score 5.0, no significant differences in scores for being understood, understanding and being helped by the moderator2) were relatively more positive towards the robot moderator.

Comparing between different proficient levels, the statistical test (one-way ANOVA) showed no significant difference between levels with respect to feeling understood by the moderator (robot or a human). Differences were significant when it comes to understanding the robot moderator ($F(3) = 0.733, p = 0.546$) and understanding the human moderator ($F(3) = 9.465, p < 0.001$), with more proficient participants understanding the moderators better. The post-hoc Tukey revealed an effect between Elementary and Beginner ($p < 0.05$) and Intermediate and Beginner ($p < 0.001$). We found significant differences in the way subjects felt helped by the human moderator ($F(3) = 3.846, p < 0.05$), with an effect between Beginner and Intermediate (post-hoc Tukey test, $p < 0.05$). Regarding the feeling of being helped by the robot moderator, there was no significant level difference between levels ($F(3) = 2.624, p = 0.08$). Finally, no significant differences between the interaction with robot or human were found between different levels ($F(3) = 1.412, p = 0.272$).

5.1.2. Previous experience in language cafés

Of the 22 L2 Swedish speakers, 15 reported that they had already been to a language café, whereas the remaining 7 answered it was the first time for them.

The results can be summarized as: Experienced language

2Note however that the fact that the differences were not significant for the beginners might also be due to the fact that there were fewer (four) participants at this level.
Table 1: Average scores for feeling understood by the moderator (scale: 1-5), understanding the moderator (scale: 1-5), feeling helped by the moderator (scale: 1-5) and comparing the interaction (scale: 1-10) with a human (Hum) and robot (Rob) moderator w.r.t. proficiency level, language café experience and wizard experience. Statistical differences are reported as ns=non significant, * for p < 0.05, ** for p < 0.01 and *** for p < 0.001, using paired t-tests

<table>
<thead>
<tr>
<th></th>
<th>General</th>
<th>Proximity level</th>
<th>Language café experience</th>
<th>Wizard experience</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hum</td>
<td>Rob</td>
<td>Hum</td>
<td>Rob</td>
</tr>
<tr>
<td>Understanding</td>
<td>4.1</td>
<td>4.1</td>
<td>4.7</td>
<td>4.7</td>
</tr>
<tr>
<td>Helped by</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Comparison</td>
<td>3.9</td>
<td>3.9</td>
<td>4.1</td>
<td>3.6</td>
</tr>
</tbody>
</table>

5.1.3. Wizard experience

The human moderator (and hence wizard for the robot) was more familiar with the wizard interface in 5 of the sessions (being the second author in 4 and a colleague in 1), was a returning recruited subject moderator in 2 of the sessions and was a first time recruited subject wizard in the remaining 4 sessions.

With a confederate wizard, reported that the human moderator was significantly better, at understanding them (p < 0.05), for understanding (p < 0.05) and helping them (p < 0.001) than the robot. With a recruited wizard, the subjects found that the human moderator was better at understanding them (p < 0.01) and helping them (p < 0.01), but that the robot moderator was easier to understand (non-significant difference).

Regarding how subjects felt understood by or understanding the robot, there was no significant difference between wizard categories. The confederates were rated significantly easier (p < 0.05) to understand. In both scenarios (human or robot), the difference between how the users felt the moderator could help them with respect to the wizard categories was not significant, even though the confederate (human or robot) was rated higher.

The score when comparing the interaction with the robot to that with the human is higher for the recruited moderators (score=4.5 vs. 3.2), and ratings were similar in the robot moderator setting (except the non-significant difference that conferee-controlled robot was perceived as more helpful). A tentative conclusion is hence that wizard experience did not influence how the subjects rated the interaction with the robot.

5.1.4. Features to be improved

The questionnaire suggested a list of features that may need improvement for a successful language café interaction: speech understanding, speech synthesis, interaction style, response time and choice of topics. Participants also had the possibility of reporting features to be improved that were not in the list provided. These are discussed below, together with our observations of the video recordings.

Table 2 shows that almost all of the participants thought that the response time should be improved, and in addition a majority that the related interaction style also needed improvement. The term interaction style leaves room for interpretation, but our understanding of this is the fact that the robot addressed one subject at a time and then turned to the other subject with the same or a similar question, rather than engaging with both of them at the same time.

Table 2: Ratio of participants stating that feature needs improvement.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Response</th>
<th>Interaction style</th>
<th>Understanding of speech</th>
<th>Speech synthesis</th>
<th>Choice of topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>81.8%</td>
<td>54.5%</td>
<td>45.5%</td>
<td>40.9%</td>
<td>36.4%</td>
</tr>
</tbody>
</table>

5.2. Observations from recordings and user comments

From the analysis of the video recordings, the following qualitative observations can be made.

**Interaction**

The set-up with the robot generally entices more pairwise interaction between one learner and the robot, rather than between peers. This is a caveat for more advanced learners, but may be appreciated by beginner or elementary level learners.

Quite naturally, the learners’ L2 level influenced the interaction to a simpler interaction with learners at lower level. Some beginner subjects in fact responded that they preferred the interaction with the robot, since it was more structured.

Interaction between peers increased with language café experience.

The order of the sessions with robot or human moderator influenced the interaction. When the robot session was first the average score for comparing both interaction was also higher (4.5 vs 3.4), although the difference was not significant according to the Welch two-samples t-test performed.

There was a variety of features that participants appreciated in the robot, for instance, the more structured interaction and the fact that they could repeat the question whenever he was unable to understand it.
Interestingly, one participant mentioned that she felt that she "was not wasting a native speaker’s time", which is something that many people might feel whenever practicing a new language in such scenarios.

**Speech technology requirements**

Concurring with the questionnaire, observations indicate that the response times for the robot were too long in general, and in particular when the wizard choose to correct an ASR misrecognition of the wizard input. The L2 learners were in fact surprisingly patient when the robot’s response time was long (up to several seconds), but the interaction clearly suffered from long response times.

The speech synthesis quality was appreciated for the clarity and simplicity of the robot’s utterances, but in the future its speed needs to be dynamically adjustable to facilitate understanding for learners with a lower L2 level and emphasis needs to be specifiable in the synthesized utterances, since this is fundamental for implicit linguistic feedback and clarifications.

Many subjects were annoyed with the robot’s filled pause, which was uttered when ASR on the wizard’s spoken input was being performed. This was particularly true when the same filled pause was repeated if the wizard had to try again to get the sentence correctly recognised before the robot could speak. Filled pauses should hence be used less frequently and with more variations. Previous studies have shown that Furhat can utilize different multi-modal signals (such as gaze aversion and facial expressions) to effectively claim the floor when responses are delayed [14].

**Furhat robot**

The robot’s automatic ability to turn the head towards the current speaker (c.f. Section 3) was found inconsistent by several subjects according to the questionnaire. Observations further indicate that this may in part have been due to timing conflicts between the automatic turning and the wizard’s manual request, so that the head turned in the opposite direction to the wizard’s intention. It is apparent that the automatic decision of whom to turn the head towards can not be based only on acoustic level, but needs to take interaction state and modeling of the robot’s intentions for the following interaction into account.

Related to the above, subjects also requested improvement in gaze, presumably to get an improved experience of the robot actually looking at the person that it is interacting with.

In some sessions, there were facial animation issues, resulting in that the robot’s lips were not moving when speaking or were moving when not speaking. This likely influenced the subjects’ perception of the interaction, and in the questionnaire, some participants requested more facial expressions.

A background story for the robot is needed. Learners were returning questions to ask robot about personality, preferences or technical details, based on role play, interest in the technology or to play the system.

A robot personality can improve the interaction by introducing jokes, irony and interaction style, as this increases learner engagement in the role play with the robot. This is also something that subjects commented on as being appreciated.

**Wizard-of-Oz**

A more structured scripted interaction sequencing is required to improve interaction flow and shorten response times (in particular for less experienced wizards). In the questionnaire, wizards requested more buttons with more sentences, backchannels and the possibility to control the robot’s facial movements. When observing the wizards in action, it seemed that they spent time browsing the different tabs and buttons in the interface, unsure of what could be said with pre-generated sentences, thus prolonging response time compared to if fewer alternatives, which were relevant to the current interaction state, had been presented in a state-chart based interface that would change over time.

Compared to being on-site moderator, the wizards were less comfortable being mediated (shift of 1 on the scale 1–5, judged that they understood the participants almost as well (86% responded 4 or 5), and that the participants understood the robot almost as well (71% answering 4 or 5), but that it was much more difficult to assist the learners with linguistic problems (72% on 1 or 2) and that the most difficult was to respond quickly to direct questions to Furhat.

Wizard experience affected the interaction, but the possibility for spoken wizard input using ASR clearly facilitated the wizard task for inexperienced wizards, and as seen above, subjects did not rate the interactions with the robot controlled by naive wizards lower. Figure 2 shows that most wizards employed the ASR input option substantially more than pre-generated utterances. It is noteworthy to observe firstly that the two sessions with lower than 50% spontaneous utterances were from sessions when the moderator started in the robot setting and secondly that the two sessions with almost exclusive spontaneous input were for the moderator with the most experience at leading the normal language cafés at KTH.

![Figure 2: Incremental ratio of spontaneous wizard utterances for different wizards and sessions. The dashed red line indicates equal use of pre-generated and spontaneous utterances.](image)

**6. Discussion**

As expected, the interaction with the human moderator was significantly better than the one with the robot moderator in almost every dimension, except for understanding the moderator, where the difference was non significant and some participants actually felt that the robot was easier to understand. Another interesting fact is that even if we can assume that the understanding by the robot was at human level, since it was controlled by the wizard, there was a significant difference in the extent to which participants felt understood. We attribute this difference to the possibility of long response times, and to some extent to inconsistencies in the robot’s turning of the head, which may have given the impression that the robot was not understanding.
The results reported in Section 5.1.2 show that in general participants who had never been to a language café before were significantly more enthusiastic about the interaction with the robot than those that usually attend language cafés. This might be due to the fact that the latter may have some expectation about what a language café should be like and they were disappointed with the interaction with the robot. This is something that should be taken into account, since it seems that the target audience was not entirely satisfied with the interaction.

Most results related to different factors in Section 5 were not significant, partly due to the fact that the number of subjects in each category is low and several factors are varied simultaneously. However, as a qualitative interpretation, we conclude that the intermediate speakers probably had a too high level for the robot language café to provide appropriate training, at least in the current state. The self-reported intermediate were in fact rather fluent in Swedish, and even though they made occasional mistakes in grammar or vocabulary, they rarely hesitated or asked for support and help, unlike the subjects at beginner and elementary level, who on the other hand appreciated the simpler, structured practice. Rather than A2 level being the lower threshold, as we initially targeted, we therefore conclude that learners who have reached at least A1 level may be more appropriate for future development of the robot language café, and that above B1 constitutes the upper limit above which interactions with the robot is not fruitful, until the system has reached a state where it is able to engage in more complex discussions.

We would have expected that interactions where a confederate with more experience using the wizard interface would have been smoother than those with a recruited wizard, but the results presented in Section 5.1.3 on the contrary showed that the average score for interaction comparison was higher with a recruited subject as wizard. However, as the confederates were rated higher than the recruited moderators in the human moderator setting, it is unclear whether the difference was that the recruited wizards were judged to perform better or if it was rather the relative difference compared to the human moderator setting that was smaller, because the confederate was also more at ease with being human moderator. More data would be needed to understand how the wizard experience affects the moderator’s interaction.

7. Conclusions and Future Work

This paper presents a preliminary analysis of a data collection in a language café setup with a robot. We have described the setup and reported an analysis both from the questionnaires filled out after the interaction and from observations of video data. Despite the wizard-of-oz setup used, which assumes perfect understanding capabilities and language generation, there were significant differences between the participants perception of the human and the robot moderators. The response time from the robot in the wizard setup is the most urgent feature to be improved in the near feature. From the data and also from the observations of the video recordings, the appropriate target level to interact with the robot in a language café scenario should be in the range A1 to B1. Participants with such level will create chances for the robot to incorporate strategies used to help participants such as completing phrases or correcting mistakes automatically. In addition, conversations with this target group will more likely have a more limited range of topics, thus being easier to process from the speech understanding point of view. On the other hand, their pronunciation can be an obstacle for the use of speech recognition. Another challenging feature to implement is handling of the quite commonly used strategy to ask how a given English word is said in L2. This requires the ability to handle code switching between Swedish and English.

8. Acknowledgements

This work is supported by the Swedish Research Council grant 2016-03698 Collaborative robot-assisted learning of Swedish as a second language.

9. References

Measuring Effect of Repetitive Queries and Implicit Learning with Joining-in Type Robot Assisted Language Learning System

AlBara Khalifa, Tsuneo Kato and Seiichi Yamamoto
Doshisha University, Japan
albara.khalifa@gmail.com

Abstract

Computer assisted language learning (CALL) becomes more realistic and motivating for learners through introduction of humanoid robots. A robot assisted language learning (RALL) system is expected to provide an immersive environment for a second language (L2) learner to prepare for real face-to-face communication. We are developing a joining-in type RALL system using two humanoid robots, one playing the role of a teacher and the other playing the role of an advanced peer learner. The interaction between the two robots and learner is designed to smoothly switch between two learning modes, that is, tutoring and implicit learning, for effective language learning. In this paper, we measured the effect of implicit learning with repetitive queries quantitatively with 37 learners divided into two groups with and without interaction for implicit learning. The experimental results showed that the repetitive queries of specific grammatical expressions consistently improved the correct use of the expressions, and the improvement was significantly greater when the peer learner robot presented an answer for implicit learning compared with when there was no assistance from the robot.

Index Terms: robot assisted language learning, implicit learning, repetitive training

1. Introduction

Today’s globalization has made communication in a second language (L2) an everyday matter for a large number of people. Computer assisted language learning (CALL) is expected to be supplementary for self-learning of L2. In accordance with technical advances in automatic speech recognition (ASR) and natural language processing (NLP), CALL systems are further expected to support various aspects of language learning, such as pronunciation, stress and accents, lexical choices, idioms and expressions, and grammatical rules. To train general oral skills for communicating in L2, various dialogue-based CALL systems have been proposed, such as SCILL [1], SPELL [2], DEAL [3], POMY [4], and DISCO [5].

To make such a dialogue-based CALL system more attractive and realistic, and to let learners prepare for the real face-to-face communication, robot-assisted language learning (RALL) systems have been proposed [6], [7], [8], [9].

A RALL system has a physical presence that a learner must be aware of while responding to the system. Physical presence was reported to be effective in increasing cognitive learning gains [10]. A RALL system introduces nonverbal modalities such as gestures, nodding, and face tracking into the interaction. Those modalities raise the level of experience closer to a real communication, letting the learner imagine reality. Previous studies, most of which were directed toward children, reported that the introduction of robots enhanced learners’ interest, motivation, and engagement [8].

In terms of a learning effect, one-on-one tutoring by a skilled instructor is believed to be the best way to learn a L2 [11], [12]. On the other hand, a learner should be exposed to various learning styles like in a classroom. Though a classroom often has the problem of having too many students to give each of them enough chances to communicate with the teacher, a student repeats after the teacher, answers questions, receives correction sometimes, and also learns by viewing the interaction between other students and the teacher. Occasionally, students are asked to collaborate on a more complex task and present their thoughts or ideas on the task. Such a combination of tutoring from a teacher and implicit learning from peer learners is considered effective in learning various aspects of communication in L2. An experimental study reported the effect of accompanying an assistive robot with a human teacher in a L2 classroom for children [13]. Children with the teacher and robot learned and retained more vocabulary than children with only the teacher. On the other hand, another study examined how the social behaviours of a social robot affect child second language learning [14]. Though children showed significant improvement between pre- and post-test in both conditions, the difference of high and low verbal availability made no additional gain.

The combination of tutoring from a teacher and implicit learning from peer learners can be simulated with robots in a clear manner. That is, a learner receives tutoring when a robot asks a question or corrects the learner directly, and the learner learns implicitly when a robot similarly interacts with another robot. Giving some examples, the interaction between robots can provide hints to make a response or simply show a model answer. Robots can even entertain or relax the learner to facilitate spontaneous speaking. Furthermore, it is possible to measure the effect of tutoring and implicit learning by automatically evaluating learner responses through the interaction with robots.

In terms of feasibility, effective implicit learning helps not only the learners but also ASR of L2 speech. Generally, recognizing L2 speech is a challenge even for state-of-the-art ASR engines because L2 speech contains various levels of pronunciation, lexical, syntactic, and semantic errors. For these challenges, implicit learning helps regulate learner utterances. The interaction between robots is expected to encourage learners to use expressions similar to those used by robots.

On the basis of this concept, we previously proposed a novel joining-in type RALL system that uses two humanoid robots for Japanese to learn English [15], [16], [17]. One robot plays the role of a teacher and the other plays the role of an advanced peer learner. Though it is not easy to realize truly flexible learning, a typical form of implicit learning is to borrow a useful expression from what a peer learner uses. We designed several scenarios involving two robots and a learner that let the learner listen to the interaction between the robots and learn useful expressions from the interactions. In this paper, we designed
a new scenario to measure the effect of implicit learning quantitatively focusing on three English grammatical expressions in conjunction with the effect of repetition, because repetition is essential in measuring the effect of learning [18]. We conducted an experiment with 37 learners to verify the learning effect with the scenario.

2. Joining-in type robot assisted language learning system

2.1. System configuration

The system is configured with two humanoid robots. The robots are set on a table, forming a triangle with a learner sitting at the table. A snapshot is shown in Figure 1. We use two NAO robots. NAO is equipped with the basic functionalities of speech output, sound production, automatic speech recognition, face tracking, and gesture making. One robot plays the role of a teacher, and the other plays the role of an advanced peer learner. When a robot speaks, it turns its face to whom it is addressing to make it clear for the learner. When a robot speaks to the other robot, the addressed one turns its face to the addressing one, too. Two robots are distinguishable with their voice, and they do not show any visual cues such as light when they speak.

The robots are currently operated in Wizard-of-Oz method because the system is under development and we are collecting a learner corpus on how learners respond to questions and how they behave with the RALL system for further development of a fully automated system. The robots are controlled remotely by an experimenter hidden from the learner through an Ethernet connection. We used recorded speech for speech output from the robots because the scenario is mostly fixed with a limited variation in responses.

2.2. Scenario with focus on specific expressions

After a couple of prototypes [16, 17], we designed a 10-minute long scenario focusing on three specific English grammatical expressions. A part of the dialogue along with the scenario is listed in Table 1. The scenario has a natural dialogue flow to let learners concentrate on the dialogue. A variety of questions that are prompted with the specific English grammatical expressions and expected to be answered with the expressions are asked repeatedly. Specifically, a total of fifteen kinds of questions are asked to the learner with the expectation that the learner will answer them by using three types of expressions that they have been prompted to focus on. That is, five kinds of questions are prepared for each expression. The three expressions are answering of negative questions, passive voice and a causative verb, all of which are difficult for Japanese learners to use correctly. The five questions for each grammatical expression are listed in Table 2, Table 3 and Table 4. To measure the effect of simple repetitive training, the scenario can be repeated.

The scenario has a flexibility of responses which can be switched depending on the learner’s input. For example, the robot is able to repeat the question, or present a sample answer followed with repeating the question if a learner could not respond to the robot, or just pass the current question and continue the scenario in case the learner could not answer at all.

3. Design of experiment for measuring effect of repetitive queries and implicit learning

3.1. Setting of learner groups

To measure the effect of repetitive queries and implicit learning quantitatively, all participants are divided into two groups: a control group and an experimental group. The baseline performance without implicit learning is measured with the control group, and the effect of implicit learning is measured as the difference of the control group and experimental group. Participants of both groups experience the same scenario consisting of fifteen questions but with a different query order. The order is set as follows. To the control group, the teacher robot always asks a question to a human learner first. To the experimental group, the teacher robot asks a question to the peer learner robot first, and the peer learner robot responds to the question in such a way that the human learner can refer to the

<table>
<thead>
<tr>
<th>speaker</th>
<th>listener</th>
<th>utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>L</td>
<td>When is your birthday?</td>
</tr>
<tr>
<td>L</td>
<td>R1</td>
<td>My birthday is November twenty seven.</td>
</tr>
<tr>
<td>R1</td>
<td>L</td>
<td>Ok.</td>
</tr>
<tr>
<td>R1</td>
<td>L</td>
<td>What were you given for birthday present last year?</td>
</tr>
<tr>
<td>L</td>
<td>R1</td>
<td>An iTunes card.</td>
</tr>
<tr>
<td>R1</td>
<td>L</td>
<td>I see.</td>
</tr>
<tr>
<td>R1</td>
<td>R2</td>
<td>When is your birthday?</td>
</tr>
<tr>
<td>R2</td>
<td>R1</td>
<td>My birthday is May third.</td>
</tr>
<tr>
<td>R1</td>
<td>R2</td>
<td>Ok.</td>
</tr>
<tr>
<td>R1</td>
<td>R2</td>
<td>What do you think your mother will be given by your father for mother’s day?</td>
</tr>
<tr>
<td>R2</td>
<td>R1</td>
<td>I think my mother will be given a necklace by my father.</td>
</tr>
<tr>
<td>R1</td>
<td>R2</td>
<td>That’s great.</td>
</tr>
<tr>
<td>R1</td>
<td>L</td>
<td>What do you think your mother will be given by your father for mother’s day?</td>
</tr>
<tr>
<td>L</td>
<td>R1</td>
<td>I think my mother will be given a flower.</td>
</tr>
</tbody>
</table>

Table 1: Part of scenario involving two robots and learner. (L: learner, R1: teacher robot, R2: peer learner robot)
Five questions expected to be answered with passive voice.

<table>
<thead>
<tr>
<th>question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
</tr>
<tr>
<td>2nd</td>
</tr>
<tr>
<td>3rd</td>
</tr>
<tr>
<td>4th</td>
</tr>
<tr>
<td>5th</td>
</tr>
</tbody>
</table>

Table 4: Five questions expected to be answered with causative verb.

<table>
<thead>
<tr>
<th>question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
</tr>
<tr>
<td>2nd</td>
</tr>
<tr>
<td>3rd</td>
</tr>
<tr>
<td>4th</td>
</tr>
<tr>
<td>5th</td>
</tr>
</tbody>
</table>

Table 5: Query order of each expression in scenario for control and experimental groups.

<table>
<thead>
<tr>
<th>query</th>
<th>control group</th>
<th>experimental group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>L first, R2 second</td>
<td>L first, R2 second</td>
</tr>
<tr>
<td>2nd-4th</td>
<td>L</td>
<td>R2 first, L second</td>
</tr>
<tr>
<td>5th</td>
<td>L first, R2 second</td>
<td>L first, R2 second</td>
</tr>
</tbody>
</table>

L: learner, R2: peer learner robot

correct use of a certain grammatical rule, but needs a criterion for every grammatical rule.

3.2.2. BLEU score

As a general measure, we evaluate learner utterances with the bilingual evaluation understudy (BLEU) [19] score. BLEU is a popular index for evaluating the quality of machine translation (MT). BLEU score is given by equation (2).

\[ BLEU = BP \cdot \exp \left( \sum_{n=1}^{N} \frac{1}{N} \log p_n \right) \] (2)

where \( p_n \) is the precision of \( n \)-grams in a learner utterance that is determined through comparison with reference sentences. \( N \) is usually set at 4, and \( BP \) is a brevity penalty, a coefficient for correction. A question from a teacher robot is used as the reference sentence when the teacher robot asks a question to the human learner first, and a sample answer presented by the peer learner robot is used as the reference when the teacher robot asks a question to the peer learner robot first.

BLEU is a general measure, though it is not perfect for evaluating if the learner utterance uses a specific expression correctly or not.

4. Experiments

4.1. Participants and experimental setup

We collected 37 participants between the ages of 18 and 24. They were Japanese university students who had acquired Japanese as their L1 and had learned English as L2. A total of 24 participants out of 37 had taken TOEIC test before. Their scores ranged between 320 and 980, with an average score of 620.

The participants were divided into the control and experimental groups, by referencing their TOEIC scores as a counterbalance. Eighteen participants were assigned to the control group, and the other nineteen were assigned to the experimental group. The average TOEIC score of the control group was 616.8, and
The average of the experimental group was 622.5. Though the averages did not cover all the participants, the general level of English skill is expected to be comparable between the groups as far as we see from the TOEIC scores available.

The experiment was conducted at a laboratory with recording equipment. Two video cameras placed in front of and behind the learner captured video. Audio was recorded with a headset microphone on the learner as well as the microphones on the video cameras. Information on a learner’s point of gaze was captured with a glasses-type eye-tracking system for future analysis.

After wearing the headset microphone and the glasses-type eye-tracking system on and receiving a brief instruction on how to respond to the robots, every participant performed the 10-minute long scenario. To measure the effect of simply repeating the scenario, every participant was asked to repeat the scenario twice with a 5-minute break in between. Hence, every participant was asked to respond to the system with a specific expression ten times, and the performance was measured in the first and fifth questions of the 1st and 2nd rounds.

4.2. Basic statistics of collected data

Every participant had 19 chances to respond to the robots including the 15 questions that were expected to be answered with the expressions that had been prompted to focus on each round. A total of 1,253 utterances were collected from the 37 participants. The average number of words in a learner utterance was 4.5. The total size of vocabulary was 691. The average time required for completing one round of the scenario was 10 minutes.

4.3. Analysis on effect of repetitive queries and implicit learning

Figure 2 shows the overall ratios of using the target expressions correctly for the 1st and 2nd round of the control group and experimental group. The effect of repetitive queries was obvious for both groups, and the degree of improvement was greater for the experimental group than for the control group. The improvement for the experimental group was 15 points while that for control group was 10 points.

Figure 3 shows the ratios of correct use of the causative verb, passive voice, and answering of negative questions. The causative verb was not practical to use orally. Some of the participants seemed confused between the causative verb and the perfect tense of “have”. The ratio of correct use stayed at a low level even after the repetitive queries, but the ratio for the experimental group showed a small improvement in the 2nd round. The passive voice of verbs was also difficult for the participants to use correctly. The ratios in the 1st round had a gap between the control group (17%) and experimental group (8%), but the improvement was 9 points for the control group while that was 16 points for the experimental group in the 2nd round. Answering negative questions was the easiest among the three expressions. Though the ratio of correct use was around 15% in the 1st round, it improved soon after repetitive queries. The improvement was 17 points for the control group and 18 points
for the experimental group. Though the correct ratio itself is different between the expressions, every expression showed improvement in the 2nd round compared with the 1st round, and the improvement was greater for the experimental group.

Figure 4 shows the overall BLEU scores for the 1st and 2nd rounds of the control group and experimental group. The scores improved in the 2nd round for both groups, and the improvement was greater for the experimental group. Figure 5 shows the BLEU scores of the causative verb, passive voice, and answering of negative questions. The magnitude relation of the scores of the two groups was not always equal to that of the ratios of correct use shown in Figure 3, but the effect of repetitive queries was obvious, and the increase was greater for the experimental group compared with the control group.

5. Conclusions

We prototyped a joining-in type robot assisted language learning (RALL) system using two humanoid robots operated in Wizard-of-Oz method. We designed an extended scenario in which learners were asked similar questions expected to be answered with specific expression forms a number of times, and measured the effect of repetitive queries and implicit learning quantitatively with 37 participants collected. Specifically, we divided participants into two groups: a control group without implicit learning and an experimental group with implicit learning, and evaluated their answers with an expression-dependent measure and BLEU score. Both improved consistently as repetitive queries were made, but the improvement was greater in the case where learners responded to the system after repeating answering similar questions with hearing sample answers by the peer learner robot.

The next step is to verify the effect of repetitive queries and implicit learning on memory over a longer term. Furthermore, after collecting enough of a learner corpus to train a language model for the L2 learner speech, we will develop a fully automated system based on ASR. We will attempt to automatically generate dialogue scenarios from a variety of topics in CogInf oCom [15].

6. Acknowledgements

This research is supported by a contract with MEXT, number 15K02738.

7. References

Emotion and Memory Model for a Robotic Tutor in a Learning Environment

Muneeb Imtiaz Ahmad, Omar Mubin, Suleman Shahid, Joanne Orlando

1The MARCS Institute, Human-Machine Interaction Group, Western Sydney University, Australia
2School of Computing, Engineering and Mathematics, Western Sydney University, Australia
3Department of Computer Science, Lahore University of Management and Sciences, Pakistan
4School of Education, Western Sydney University, Australia

Abstract

In this paper, we present an emotion and memory model for a social robot. The model allowed the robot to create a memory account of a child’s emotional events over 4 individual sessions. The robot then adapted its behaviour based on the developed memory. We tested our model through using the NAO robot. The robot was programmed to teach vocabulary to children during the popular game 'Snakes and Ladders'. We conducted a between-subject study with 24 children at a primary school to check the validity of our model. We also evaluated the impact of robot’s positive, negative, and neutral emotional feedback of the NAO robot on children vocabulary learning. Three groups of children (8/group) interacted with the robot for four different times during three weeks. Our results showed that the condition where the robot displayed positive emotional responses had a significant effect on the child’s learning performance as compared to the two other conditions: negative feedback and neutral feedback. In addition, our model helped children in improving their vocabulary.


1. Introduction

One of the growing interests of the social robotics research community in the last decade has been towards the use of social robots in Education. Robots have been employed as a tool to teach computer programming skills in the past [1]. However, due to the introduction of Humanoid robots, new opportunities have opened to use robots in education in ways other than as a tool to teach concepts from various subjects. Most recently, Mubin et al. [2] reported a survey on the applicability of robots in Education. They emphasised the need for an adaptation mechanism that will enable the robot to adapt its behaviour according to the characteristics of the user/student. The applications of such adaptive social robots can be found in education [3] [4]. Most of these studies evaluated short-term interaction [5] and did not capture childrens real interaction with a robot and long term engagement with a robot which is essential to understand the role of robots in the future educational landscape.

We face many technical and social challenges during long-term interaction with social robots [6]. One of them is about decrease in children’s interest in the interaction over time. The reasons for this decrease in interest are robot’s repetitive behaviour [7] and loss of novelty factor [8]. It is emphasised to implement various autonomous adaptation mechanisms for a social robot to overcome the aforementioned effect. These mechanisms can be based on user’s emotions, memory, or personality [5] [9] [8]. The autonomous adaptation mechanism for a social robot can be implemented using different approaches. It can either be through combining machine learning algorithms to chose the robot’s behaviour or by following cognitive models that describe how humans create memory or how emotions are regulated in diverse situations and later applying them for social robots. For instance: Belpaeme et al. [10] proposed a model for adaptive strategies for sustainable long-term social interaction based on the theories in cognitive sciences. Trafton et al. [11] presented a cognitive architecture named ACT-R/E (Adaptive Character of Thought-Rational / Embodied) that enables the robot to predict what a user will do in a certain scenario through understanding previous knowledge about the user. Leite et al. [8] designed an emphatic model for an iCAT robot capable of playing chess with children. As stated earlier, we find limited research on social robots that have been used as partners with students in a learning environment during long-term interactions [12]. These afore-listed models are designed for specific purposes in HRI. We therefore, find a vacuum for a robotic model that can be employed to facilitate personalised learning.

In this paper, we present an emotion and memory model for a social robot. The model enables the robot to create a memory of user’s emotional events and selects an appropriate behaviour accordingly. The model is based on the theory on how humans create memories of an emotional event/episode [13]. We conducted a 4 week long-term children-robot interaction (cHRI) study to evaluate our model. We programmed the NAO robot to play snakes and ladders game designed to teach vocabulary from the Robot Interaction Language (ROILA) [14] to children in a playful and interactive way. We implemented three types of robot’s emotional responses based on positive, negative and neutral emotional events happening during a snakes and ladders game play. Our research question is about studying the effect of robot’s positive, negative and neutral emotional response on a child’s long-term learning performance in a vocabulary learning task. To the best of our knowledge, this effect has not been studied during the long-term child-robot interaction.

In this study, we choose to focus on vocabulary learning as the interaction task because it is one of the essential components of language learning [15]. Vocabulary learning helps improve listening, reading, writing and speaking skills [16]. Moreover, we choose the artificial language ROILA for this study because it was created based on the rules, syntax, and principles of the major natural languages of the world, which will allow us to mitigate the confounding factor of children having different linguistic backgrounds; it will always be an influence but perhaps less so in the case of ROILA [14] because it has no connection with other languages and dialects spoken. We choose games because the significance of play and interaction in education has
been well described [17] [18].

2. Emotion and Memory Model

Our model is grounded in the process of formation of emotional memories as described by [13]. It is a well understood fact that humans create memories of both positive and negative emotional experiences. This type of memory is usually stored in two different parts of the human memory system: 1) explicit memory that refers to conscious memory and 2) implicit memory that refers to unconscious memory. In the literature, the memory of the different emotional experiences is categorised as Emotional memory (implicit memory) and memory about emotions (explicit memory). In general, emotional events are processed in human sensory systems. They are later transmitted to the temporal lobe or to the amygdala in order to form either an explicit memory or an implicit memory. Simultaneously, the memory is retrieved in case of occurrence of a cue. The cue is again processed by the sensory system that later leads to retrieval of both explicit or implicit memories. We have utilised the process of creation and retrieval of the memory about the external emotional event as described by [13] and created a model for robots to create memory about different emotional situations and behave accordingly.

It is necessary to define positive and negative emotional events in order to explain our model for the robot. Positive emotional events are described as events when goals are achieved or no immediate problems are encountered towards achieving the goal. Negative emotional events are registered as impediments towards a plan and causing loss to achieve a certain goal. On the other hand, Neutral events are situations that do not significantly threaten an outcome in either positive or negative ways [19].

Levine et al. [20] presented a review on emotion and memory research and showed that different types of information are remembered under various emotional states. As a user’s emotional state is directly related to an emotional situation. A positive or negative situation would refer to positive or negative states. It is therefore, necessary to understand the information that should be stored in an emotional state or at an emotional situation. According to [20], humans store a broad range of information from general knowledge and the environment to their memory. Depending on the type of negative emotional state (sad, fearful, or anger) during an emotional situation, humans store different types of information. For example; Sadness may lead to remember about the outcomes and consequences of goal failure. Anger may lead to store information about goals or agents obstructing goal fulfillment. Lastly, fear may lead to storing information about the source of threat and means of avoiding the threat [20].

Based on the general understanding of how humans store emotional information in their memory, we have designed our emotion and memory model, as shown in figure 1, to enable a robot to create a memory of user emotional events and behave accordingly. The purpose of the model is to facilitate different kinds of learning such as concepts from science or mathematics during children-robot-game interactions. Our model has four different modules 1) Inputs, 2) Emotional Event Calculation (EEC), Memory Mechanism Generation (MMG) and 4) Behaviour Selection Unit (BSU). The model has three input types: 1) Game event, 2) user emotional state, and 3) Learning outcome that amalgamates to create an emotional event during the interaction. Based on the type of game event (positive or negative or neutral), user emotional state (happy, sad, angry, fear, surprise, neutral), and learning state, we calculate the type of emotional event in the EEC module. EEC module transmits this information to the MMG module. Based on the type of event and following the type of information remembered under various emotional situations, we send this information to our Memory Processing Unit. We create the memory for the robot in this unit. In addition, in the case of an occurrence of same event type during same circumstances, we update our database with the new event and send the information back to the MMG module. The MMG module later transmits the information to the BSU, that is responsible for selecting an appropriate behaviour or response. Lastly, the robot displays the behaviour.

![Figure 1: Emotion and Memory model](image_url)

2.1. System Description

To test the applicability of our model, we implemented a scenario where the NAO robot plays snakes and ladder game with a child during an one-to-one interaction. In this section, we present our modified version of the snakes and ladders game. We also discuss the mechanism we used to calculate a type of an emotional event along with the type of information stored in our system. Lastly, we give information on selection of the robot’s behaviours under different situations.

**Snake and Ladders Game:** We modified the game to facilitate vocabulary learning [21]. We updated rules of the Snakes and Ladders game as shown in figure 2 and also introduced stars on the game board. On every snake appearance in the game, NAO robot was programmed to teach a new ROILA word to the child. In the first iteration, the child was asked to go back to the tail of the snake. However, if the same snake reappeared and if the child stated that the word correctly that was taught to him/her on that number, the child could bypass the snake. The same process was repeated on each snake. In the case of a ladder, the child would take the ladder. On every star, positive or negative number appeared on the dice suggesting the player to move forward or backwards. Lastly, the child was declared the winner when he/she reached the 100 mark.
Applying the Model in the Snakes and Ladders Game:
We, in order to realise on the emotional event type, categorised both positive and negative type of game events based on children reactions coded in our previous study on various game events [22]. In the previous study, we coded for the significant game events such as appearance of a snake, ladder or positive/negative star near or away from the 100 mark, continuous sixes on the dice, continuous wastage of turns near 100, and winning or losing the game. We calculated the emotional state of the player through automatic face scans as described in our previous research [22]. We stored six different emotions (happy, sad, fear, surprise, angry, and neutral) values after every 10 seconds of the interaction. On each significant game event, we calculated current emotional state by taking an average of the last six emotional states stored in our system. Lastly, the learning state referred to the outcome of the words taught during the interaction. In table 1, we briefly present our list of selected emotional events calculated on the basis of three inputs of our model along with the type of information stored during these events in order to create robot’s memory. We also include the behaviour for the NAO robot. For instance, considering the definition of a negative emotional event, a snake near 100 will be rated as a negative event because it hampers the child from winning the game or thwarts the child from achieving the final goal. The information will be stored about the number of times a snake is encountered near 100 and the emotional state of the user. Similarly, a ladder near 100 will be considered a positive event because it is helping the user achieve the end goal. The information about the ladder will be stored in this regard. The behaviour selection of the robot uses the memory of previous emotional events to generate context-aware verbal and non-verbal response either independently or simultaneously. As our purpose was to confirm our model’s applicability, we used decision making statements to chose robot’s behaviour. We created a database of robot’s behaviour consisting of all plausible emotional events during the snakes and ladder game. On each event, the robot displayed the most appropriate behaviour by retrieving it from the database. In the table 1, for understanding, we only enlist a few behaviours.

3. Research Method
Our research tried to explore two aspects. Firstly, we wanted to understand how well our model for the robot performed in terms of teaching vocabulary to children in a long-term interaction. Secondly, we find different claims in literature with respect to the effect of emotions on human memory. A body of research shows that emotions enhance memory in tone, while other claims that emotions enhance central information at the cost of peripheral details [20]. Considering these claims, it would be interesting to find answers to the following Research Questions (RQs) in the context of children-robot interaction:

<table>
<thead>
<tr>
<th>Learning Outcome</th>
<th>NAO’s Behaviour on the Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>correct word</strong></td>
<td>Positive Condition: I am delighted to know you got this correct, I am so happy that in the &lt;SESSION NO&gt; session, you got it wrong but this time your answer is correct. Happy or Joy Gestures &lt;br&gt; Negative Condition: Ok, this is fine that you have got it correct, You remembered it today, but it took you &lt;NO OF ATTEMPS&gt; to learn &lt;WORD NAME&gt;. &lt;br&gt; Neutral Condition: This is fine. let’s check the next one.</td>
</tr>
<tr>
<td><strong>incorrect word</strong></td>
<td>Positive Condition: It is alright, I am hopeful you can get it right next time. &lt;br&gt; Negative Condition: I am so sad, that you didn’t remember &lt;WORD NAME&gt;, I am feeling disappointed that you don’t remember &lt;WORD NAME&gt;, we learnt it in the &lt;SESSION NO&gt; session, Sad Gesture &lt;br&gt; Neutral Condition: this is Ok, you can get next one right.</td>
</tr>
<tr>
<td><strong>total correct words in the test</strong></td>
<td>Positive Condition: I am so fond of you, you are performing consistently well, I am so proud of you, you are performing consistently well in all you tests, Clapping, bow down Gesture &lt;br&gt; Negative Condition: you are doing well, but I am still sad that you answered &lt;NUMBER OF WORDS&gt; incorrectly, Disappointment Gesture &lt;br&gt; Neutral Condition: This is alright, I think, lets hope for the best next time.</td>
</tr>
</tbody>
</table>

Table 2: Robot’s varying behaviour on Children’s learning outcome.

Our interaction session had three main phases. In the first phase, NAO robot asked about the words to be taught during the game. In the first session, NAO asked about the words with an assumption that children didn’t know the word. The NAO responded by saying “We will learn about the <WORD NAME>...
3.2. Setup and Materials

We were assigned a quiet room at the school that was divided into two parts with a divider as shown in figure 3. On the left side, one of the researchers was controlling the speech recognition capabilities of the robot. On the right side, the child interacted with the NAO robot placed on the table along with a Samsung tablet. We used the NAO robot designed and developed by Aldebaran robotics. It is a humanoid robot measuring 58 cm in height with 25 degrees of freedom.

3.3. Participants

We conducted our between-subject study with 24 children (12 males, 12 females) aged between 10-12 at a school. Each group comprised 8 children with equal ratio for the gender. None of the participants had previously interacted with a robot.

3.4. Procedure

Our study was setup as a long-term between-subject evaluation that lasted for three weeks. The study was conducted individually with one child at a time. Each child played the snakes and ladders game with the NAO robot 4 times for 4 days (one session per day), for a total of 96 sessions (24 child * 4 sessions). Each group of children played the game on a tablet in one of the three conditions (robot’s displaying positive, negative, neutral emotional expression on child’s learning outcome) for three school weeks. We conducted our sessions on the 1st, 5th, 10th and 14th day respectively. Each session lasted for approximately 24 minutes and had five steps: 1) a 4-minute pre-test, 2) a 10-minute game playing session, 3) a 4-minute break, and 4) a 4-minute post-test. The facilitator used a stopwatch to maintain time consistency throughout the sessions.

**Introduction:** NAO introduced itself and communicated with the child through a high-level dialogue. The dialogue involved inquiring about their day and activities that they are performing today.

**Pre-test:** The robot initiated the session through asking about unknown words from the ROILA language. The robot asked about six words during the first session. The rationale for selecting 6 words in each session came from a pilot study conducted with 5 participants. In the following sessions, six new words were added to the test. Therefore, 6, 12, 18 and 24 words were tested for in the first, second, third and fourth session respectively. The pre-test was an auditory-visual word identifica-

<table>
<thead>
<tr>
<th>Game Event / Learning State</th>
<th>Emotional State</th>
<th>Event Type</th>
<th>Information NAO’s Behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snake near 100</td>
<td>Happy</td>
<td>Negative</td>
<td>First Session: A snake near 100, I can see you are feeling &lt;USER EMOTION&gt;, its looks you want to learn more. Other Sessions: You had a snake near 100 during &lt;SESSION NO&gt; session, you looked &lt; USER EMOTION &gt;.</td>
</tr>
<tr>
<td></td>
<td>Smiling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Surprise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snake near 100</td>
<td>Angry</td>
<td>Negative</td>
<td>First Session: This is &lt;SNAKE OCCURRENCE&gt; time you are on a snake today, but you look &lt;USER EMOTION&gt; about it, let’s learn another word. Other Sessions: You looked &lt; USER EMOTION &gt; in session &lt; SESSION NO &gt; on a snake near 100, Are you not enjoying vocabulary learning today.</td>
</tr>
<tr>
<td></td>
<td>Sad</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ladder near 100</td>
<td>Angry</td>
<td>Positive</td>
<td>First Session: Why are you looking &lt;USER EMOTION&gt;, you are moving towards 100. Other Sessions: You also had a ladder near 100 in the &lt;SESSION NO&gt; session, I am happy to see you progressing well.</td>
</tr>
<tr>
<td></td>
<td>Sad</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ladder near 100</td>
<td>Happy</td>
<td>Positive</td>
<td>First Sessions: You look &lt;USER EMOTION&gt;, It is good to see you are playing well. Other Session: You also had a ladder near 100 in the &lt;SESSION NO&gt; session, you are extremely lucky.</td>
</tr>
<tr>
<td></td>
<td>Smiling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Surprise</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Taxonomy for Emotional Event type and type of Information stored based on the Emotional Events

shortly”. From the second to the fourth session, the robot provided positive, negative and neutral feedback through combining gestures and dialogue on the learning performance of each child. In the second phase, NAO robot played snakes and ladders game with the child. During the game, the child was taught six different words in each session. We coded a “fixed/predetermined” pattern of turns for both child and robot during the game for every session for all the participants. The aforementioned model was applied during the gameplay to create the memory of emotional events during the game. This memory was utilised after the first session. In the last phase, NAO tested the vocabulary learning today.

We used 24 vocabulary words from the Robot Interaction Language (ROILA) taken from the first two chapters of the book on ROILA [14].

Figure 3: Setup: A Child playing snakes and ladders with NAO.
Table 3: Mean values of children’s learning outcome during sessions

<table>
<thead>
<tr>
<th>Session Nr.</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.7500</td>
<td>.98907</td>
</tr>
<tr>
<td>2</td>
<td>5.3750</td>
<td>.76967</td>
</tr>
<tr>
<td>3</td>
<td>5.7500</td>
<td>.60792</td>
</tr>
<tr>
<td>4</td>
<td>5.0000</td>
<td>.65938</td>
</tr>
</tbody>
</table>

The visual used in the test to represent a word was identical to the one used in the game playing sessions.

Game Play: Each child played the snakes and ladders game following a pre-defined pattern of dice outcomes for four times. Each child faced a snake inside the game six different times. On each snake, a new word was taught to the child. Therefore, in each session, six new words were taught to the participants.

Post-Test: After a 4-minute break, the child participated in the post-test to determine the accuracy of words learned during the session. The same procedure as pre-test was repeated in this phase, however, in the case of a mistake, the robot mentioned the correct answer during the feedback. The post-test was identical to the pre-test, containing the same words. We chose identical words for both pre and post-test to maintain the consistency of test results. All of the test results and mistakes were logged in the database.

3.5. Measurements

We measured the following Dependent variables: 1) total number of words learned during all the sessions (total number of words remembered in the last post-test on the last day), 2) immediate retention of new words during the session (total number of words remembered in the post-test of every session) and 3) retention of old words across sessions (total number of words remembered during the pre-test taught in the previous sessions).

4. Results

We conducted Kolmogorov-Smirnov test to ensure that the generated data was normally distributed before conducting Analysis of Variance (ANOVA). The results showed that the data was normally distributed for the learning outcomes of the children.

In order to check our model, we checked for the immediate retention of words learnt during each gameplay in all the sessions for all the participants. We conducted a repeated measure ANOVA with the session as the within-subjects factor with four levels using immediate retention of words learnt per session as a Dependent Variable (DV). Results showed that there was a significant effect of session (F(2, 19) = 5.7, p < 0.02) of the type of robot’s emotional feedback on the child’s learning outcome. The mean retention rate across conditions were as follow: Condition 1) M: 22.25, S.D.: 1.03510, Condition 2) M: 20.25, S.D.: 1.16496 and Condition 3) M: 20.50, S.D.: 1.60357. We performed a Bonferroni posthoc check to further examine this significant difference. We found that robot with a positive emotional feedback had the better effect on child’s overall learning outcome as compared with negative (p < 0.04) and neutral feedback (p < 0.02). The neutral feedback was significantly preferred (p < 0.04) as the second choice and negative was the last choice.

To check the effect of robot’s feedback on children learning of words across sessions, we conducted a repeated measure ANOVA with the session as the within-subjects factor with three levels and type of emotional feedback as the between-subject factor using retention of words learnt during the first session across sessions. We found a significant effect (p < 0.04) of robot’s emotional feedback on the retention of words learnt during the first session. The Bonferroni posthoc check showed that positive emotional feedback from the robot has a better effect on child’s learning (p < 0.04) as compared with negative feedback. We also conducted repeated measure ANOVA with the session as the within-subjects factor with two levels and type of emotional feedback as the between-subject factor using retention of old words learnt during the second session in the third and fourth session. We found a significant positive effect (p < 0.02) of robot’s emotional feedback on the retention of words learnt during the first sessions. The Bonferroni posthoc check showed that robot with a positive emotional feedback has the significant effect on child’s overall learning outcome (p < 0.03) as compared to the negative feedback. For the fourth session words, we conducted a one-way between-subject analysis of variance (ANOVA) with robot behaviour type as the IV and using a total number of words retained during the fourth session as DV. Our results showed that there was a significant effect (p < 0.001) of robot’s emotional expression on the child’s learning outcome. The Bonferroni posthoc check showed that robot with both positive and negative emotional feedback has the significant effect (p < 0.03) on child’s overall learning outcome as compared to the neutral feedback.

5. Discussions and Conclusions

Our emotion and memory model for the NAO robot resulted in positive findings as the immediate retention of vocabulary for all the children was satisfactory during all the sessions. We witnessed a difference in children’s learning outcomes within sessions. We conjecture that the reason might be due to NAO’s novelty effect during the first session as children might be excited to see the robot. Similarly, a slight decline in the fourth session on the learning performance can also be grounded in the varying levels of children’s interests in the interaction. As it is known that children learning can be affected due to the fall in their interest [7].

We also found that the positive emotional feedback provided by the robot on a child’s learning performance did have a significant effect on the short-term and long-term children’s vocabulary learning. Our hypothesis was accepted as children were able to retain the most number of words during robot’s positive emotional feedback condition. In addition, the neutral response on child’s feedback was also found to have an influence on child’s learning. Moreover, the negative emotions were least regarded in terms of children’s long-term learning
performance. We conjecture that when the robot positively empathises with the child during the interaction, it creates a positive effect on the child’s development in general. In the past, the role of empathetic robotic behaviour has been appreciated during a playful interaction as it was able to sustain children’s interest in a long-term child-robot interaction [8]. We also speculate that the reason children were able to better retain words during positive emotional feedback is due to the positive emotional state of the children in response to the robot’s reaction. It is shown in literature that humans store different kinds of information during different emotional states [20]. In addition, positive emotional feedback is conducive to feeling confident and successful in the learning process [24]. It would have facilitated enhanced learning. Therefore, we believe that a positive reaction of a robot created a positive emotional state of the child. It in return, made the child perform better as compared with the negative or neutral reaction of the robot. One of our findings also showed that from the third to fourth session, the negative feedback was preferred over neutral feedback. We speculate that although positive feedback is highly desirable, but negative feedback is also needed during the learning process. As it is shown in previous literature that negative feedback in terms of criticism may positively encourage student engagement and attention on learning task [25]. Therefore, negative feedback or directive critique can be useful in certain situations.

In general, our findings highlight the need for robotic tutoring systems where the robot takes a positive role and appreciates children. In addition, we also found a little evidence of using negative feedback such as displaying sadness during feedback may also lead towards improved learning. In summary, it indicates towards implementing positive adaptive roles for the robotic tutor. The role can be a helper, a friend or a buddy that may also provide criticism at times during cHRI as also indicated by the teachers in one of the past studies [26].

5.1 Limitations and Future Work

One may argue on the number of sessions when it comes to being categorised as “long-term”. However, our selection of a number of sessions is based on the findings of our previous study [22], where children’s novelty factor diminished from the third session.

We understand that the number of repetitions of words as a part of feedback was not constant between the conditions. It might have created a potential confound in the results. In future, we will try to take it into account as a part of our data analysis.

We also understand that some participants would speak more than one language and/or have a greater level of capacity for learning another language. Our study did not select children on this basis.

In the future, we intend to test the generality of our model with multiple learning tasks (maths, science) during different games. We also intend to utilise our findings based on the feedback in future studies.

6. References


Evaluating the Efficiency of Synthetic Voice for Providing Corrective Feedback in a Pronunciation Training Tool Based on Minimal Pairs

Cristian Tejedor-García¹, David Escudero-Mancebo¹, César González-Ferreras¹, Enrique Cámara-Arenas², Valentín Cardeñoso-Payo¹

¹Department of Computer Science ²Department of English Philology
Universidad de Valladolid
cristian@infor.uva.es

Abstract

Feedback is an important concern in Computer-Assisted Pronunciation Training (CAPT), inasmuch as it bears on a system’s capability to correct users’ input and promote improved L2 pronunciation performance in the target language. In this paper, we test the use of synthetic voice as a corrective feedback resource. A group of students used a CAPT tool for carrying out a battery of minimal-pair discrimination-production tasks; to those who failed in production routines, the system offered the possibility of undergoing extra training by using synthetic voice as a model in a round of exposure exercises. Participants who made use of this resource significantly outperformed those who directly repeated the previously failed exercise. Results suggest that the Text-To-Speech systems offered by current operating systems (Android in our case) must be considered a relevant feedback resource in pronunciation training, especially when combined with efficient teaching methods.

Index Terms: Synthetic Voice, L2 corrective feedback, non-native speech recognition, minimal pairs, CAPT.

1. Introduction

In cybernetics, the notion of feedback describes a process by which the effect of an action is sent back to the system so that it can decide what the next step should be [1]. In this particular sense, adaptive systems behave just like experienced teachers, who are able to adapt to their students’ learning styles and improvise situated teaching strategies that make the most of their students’ potential [2, 3]. Within the field of speech technology, the number of experiments with CAPT tools that incorporate automatically generated corrective feedback has been increasing over the last few years [4, 5, 6]. Although most of the feedback usually consists of a right/wrong answer and a score, some new methods involving the use of pedagogical [7], visual [8, 9] and exaggerated [10] speech are currently being developed. In our experiment, a system for training L2 pronunciation can also determine particular difficulties of its users and propose specific exercises by way of feedback in order to improve users’ performance.

The training protocol implemented by our tool is partially based on the Native Cardinality Method (NCM) [11, 12] and other related training programs [13, 14, 15]. The basic dynamics consists of the iteration of exposure-discrimination-production cycles. We use Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) technology within a gamified environment, as described in [16, 17, 18, 19]. On this occasion, we present the results of a controlled experiment for which an adaptation of the tool was required. Pedagogical content was reduced to the training of six pairs of vowels (including some of most challenging for L2 Spanish speaker) through minimal-pair based techniques. In phonology, a pair of words is considered minimal when they differ in only one phoneme, as in bet-bed; primarily devised as a technique for elucidating the phonological system of unknown languages, minimal-pairs have been used for increasing phonemic awareness in second language teaching for more than half a century [20]. For the adapted version of the tool we reduced the gaming component and made it mandatory for participants to watch a set of videos containing articulatory instructions for each vowel and carefully designed exposure cycles, closely following here the native cardinality approach [19]. The training program included a fixed number of compulsory discrimination and production exercises within each working session. When participants produced wrong pronunciations they were forced to repeat the production task. At this point, however, and as a feedback, the system offered the possibility of extra practice through specific exposure exercises. Participants could either try again with the failed exercise, or follow the feedback training recommendations generated by the system. In this paper we report on the benefits of the second choice.

In the exposure exercises used as feedback training, a TTS system was used to generate model performances of particular minimal pairs. We have elsewhere discussed the limitations of TTS systems as language learning tools [21]. However, over the last years the quality of these systems has improved significantly due to the availability of ever larger corpora, on the one hand, and statistical parametric [22] and deep neural methods [23] that can process both superficial and hidden information of these corpora, on the other. In this paper we show that a significant number of users objectively improve their discrimination and pronunciation skills when Android TTS tools are used in feedback exercises.

In section 2 below, we expound the pedagogical basis of the training program (subsection 2.1), the working dynamics and the interface of our CAPT tool (subsection 2.2) and the experimental procedure deployed (subsection 2.3). Results are presented in section 3. The paper ends with discussions and conclusions.

2. Method

2.1. The pedagogical fundamentals of the training sessions

As described in the introduction section, the training protocol proposed follows partially the NCM [11, 12]. Carefully designed exposure activities mediate the inductive discovery of the L2 phonemes from first-hand perceptive experience. When this experience is integrated and memorized, success at recognition and identification through discrimination exercises con-
firms and deepens acquired knowledge of L2 phonemes. The last step consists in rematerializing (producing) the mentally acquired phonemes. Recent research has emphasized the importance of getting the learners to notice their own errors [24, 25, 26]. In producing the L2 sounds at the final stages of our experiment protocol, learners are no longer imitating an externally presented model, but trying to build the sound by accommodating to a mental representation of it, already acquired at the previous stages. In this way, students are also expected to detect mismatch between mental and physical forms; they should be able to self-diagnose accuracy, and know when self-correction is in order.

The notion of different learners learning differently, according to individual styles and abilities, has been gaining relevance among researchers in the field over the last years [27]. In fact, many students manage to jump from perceptive memory to accurate production by dint of sheer intuition, while others welcome explicit articulatory instructions. The topic of whether explicit instruction in phonetics actually assists improvement remains rather controversial [28]. In any case, each training session is prefaced by a brief theory video informed by the NCM approach that aims at providing, above all, the perceptive induction-oriented experience mentioned above. However, for deduction-minded students, the videos also incorporate instructions in the NCM style; that is, they indicate the kind of transformations we must practice upon an L1 sound in order to turn it into an L2 sound. The wording in the videos is intentionally synthetistic and intuitive terms – ‘pronounce Spanish /e/, and now try to give it a little bit of /a/ flavor’. Both articulation and perception cues are used. In this sense we try to address different learning styles.

2.2. Interface of the application

We have developed an Android application that is technologically similar to the prototypes used in previous work [16, 17, 18, 19]. In this version, we have eliminated several game elements and turned it into a strictly guided pedagogical tool. The first screenshot in figure 1 corresponds to the application’s launch screen, as it appears when login is completed. It displays the six proposed lessons and the accumulated score in each of them. Each lesson addresses a vocatic contrast that students of English as a Foreign Language (EFL) tend to find challenging. In each 60-minute session, two lessons were completed. The first session dealt with vowels /a:, æ, 2/ between user and application, and they could end in either success or failure. After three consecutive failures, the system executes a correction is in order.

The notion of different learners learning differently, according to individual styles and abilities, has been gaining relevance among researchers in the field over the last years [27]. In any case, each training session is prefaced by a brief theory video informed by the NCM approach that aims at providing, above all, the perceptive induction-oriented experience mentioned above. However, for deduction-minded students, the videos also incorporate instructions in the NCM style; that is, they indicate the kind of transformations we must practice upon an L1 sound in order to turn it into an L2 sound. The wording in the videos is intentionally synthetistic and intuitive terms – ‘pronounce Spanish /e/, and now try to give it a little bit of /a/ flavor’. Both articulation and perception cues are used. In this sense we try to address different learning styles.

The second mode (figure 1, fourth screenshot) presents the participants with the task of producing the words of a minimal pair with as much precision as possible. Here we rely on Google automatic speech recognition for Android (Google’s ASR) to offer an n-best list of probable results for each utterance. In our tool, the challenge is overcome only when the first item of the n-best list coincides with the target word. The words to be pronounced in this version of the tool constitute a close list whose items have been selected and supervised by an expert, ensuring that they are all recognized by ASR, and that homophones are adequately processed. Each pronunciation token contained a minimal pair to be read, each word separately. Five attempts per word were allowed in order not to discourage users. After three consecutive failures, the system executes a feedback response that allows the user to listen to a synthesized version of the problematic word.

The aim of the Mixed mode (figure 1, seventh screenshot) is to further consolidate acquired knowledge and skills. In this mode, Discrimination and Pronunciation tasks alternate summing up a total of nine tokens.

All lessons followed the same structure, that is, in each of them the five modes are consecutively undertaken, in the same order in which they have been described. Figure 1 shows the

<table>
<thead>
<tr>
<th>Mode</th>
<th>THE</th>
<th>EXP</th>
<th>DIS</th>
<th>PRO</th>
<th>MIX</th>
</tr>
</thead>
<tbody>
<tr>
<td># Task-tokens</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>
algorithm process of a lesson. The lesson is completed only if and when users get at least 60% in all their mode’s scores. When the score in any mode remains below 60% after a fixed number of attempts our tool executes optional corrective feedback, as shown in figure 2.

2.3. Protocol and testing population

Ten EFL students participated in a three-day experiment, consisting of three 60-minute sessions separated by at least 48 hours. The experiment was carried out under the supervision of a member of the research team, in a classroom equipped with computers. Each participant used a pair of headphones, a microphone and a computer where our tool was installed. All student-computer interaction as well as all events and audio recordings were automatically monitored and stored for later analysis.

Participants were recruited from the same English course at the Language Center of the University of Valladolid. Their certified level of English proficiency was in all cases B1-B2 of the Common European Framework of Reference for Languages (CEFR). In order to ensure the homogeneity of the group, participants took an initial test. On this occasion we particularly interested in working with students with low or null previous knowledge and training in English phonetics. Table 2 gives the relevant demographic details concerning participants.

Table 2: Participant demographic profiles.

<table>
<thead>
<tr>
<th>Total</th>
<th>Female</th>
<th>Male</th>
<th>Age: 15-25</th>
<th>Age: 26-45</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

3. Results

Table 3 summarizes the use of the tool in the training sessions. The second row registers the total time spent by user in each training mode. Participants spent a significant portion of time viewing the videos of the Theory mode: 31.32 minutes (27.42% of the time). However, more time was spent as a whole in meeting the challenges posed by the other modes: 82.91 minutes (72.58 % of the total). Users dedicated less time to Discrimination mode activities than to other task-types. This confirms the fact that discrimination activities, albeit methodologically essential, are generally easier than production tasks for most users.

Table 3: Time spent and total events by person in each mode. THE, EXP, DIS, PRO and MIX correspond to Theory, Exposure, Discrimination, Pronunciation and Mixed modes, respectively. NA stands for not applicable. In. and Ex. mean intrinsic and extrinsic. Listening types use TTS. Production types use ASR.

<table>
<thead>
<tr>
<th>Mode</th>
<th>THE</th>
<th>EXP</th>
<th>DIS</th>
<th>PRO</th>
<th>MIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (min)</td>
<td>31.32</td>
<td>16.93</td>
<td>5.48</td>
<td>41.47</td>
<td>19.03</td>
</tr>
<tr>
<td># In.listening</td>
<td>NA</td>
<td>3570</td>
<td>695</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td># Ex.listening</td>
<td>NA</td>
<td>1469</td>
<td>299</td>
<td>1479</td>
<td>632</td>
</tr>
<tr>
<td># Productions</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>4415</td>
<td>1741</td>
</tr>
<tr>
<td># Recordings</td>
<td>NA</td>
<td>902</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

The third and fourth rows in table 3 are directly connected to the use of the TTS synthesizing system. The third row shows the number of intrinsic TTS listening events within each training mode. Intrinsic listening events are those that are presented by the system as part of a task-type, and therefore necessary for the completion of EXP, DIS or MIX modes. The computation of intrinsic listening events includes, of course, those that happen when the user has to repeat a task-token after one or more failures. On the other hand, extrinsic listening events are those that are either requested by the user (in EXP, DIS, MIX modes) or accepted when offered by the system as optional feedback (in PRO, MIX modes). As expected, the Exposure mode registers the most intense use of the TTS system: 3570(In) + 1469(Ex) = 5039 TTS events. This is, first, because the main challenge in EXP lies precisely in listening attentively and as often as necessary for subtle contrasting sound features; and, second, because
the EXP mode is revisited both mandatorily and upon request whenever the participants stumble against a difficult sound.

On the other hand, the amount of extrinsic listening registered in Table 3, a total of 3879 events, gives us a hint as to the use of TTS feedback required by users. TTS feedback was relatively little used in DIS mode (299 times). Although EXP and PRO modes register a similar amount of TTS listening requests (1469 and 1479 respectively), an adequate interpretation of these values must take into account its relation to the number of recording within EXP and productions fed into the ASR in the PRO mode. In the EXP mode, users recorded their speech 902 times with 1469 extrinsic listening events: 1.63 listening events for each recording. With far more production (4415) than extrinsic listening events (1479), the situation in the PRO mode is totally reversed: there is one extrinsic listening event every 2.99 productions fed into ASR.

Considering all the values registered in Table 3, except those in the time row, the experiment involved a total of 15370 events. A total of 8.5 events of listening, production or record per minute and per user were registered, generating a rich information pool on TTS and ASR interactions collected during the three sessions of the experiment.

Table 4 shows the number of times each mode was practiced. We could envisage several scenarios here: (1) a mode was passed (grade 60% or higher) with a single round, (2) a mode was passed after repetition with or without feedback, and (3) a mode was not passed with or without feedback.

The asymmetry of the modes is evident. Mode DIS was the easiest one: It was passed 51 out of 60 times only one round (83.33%). The PRO mode was the most difficult, with 61 repetitions and a 58.33% success in the first round. When repeated, only in two occasions was it overcome without the help of designed feedback. The teaching of efficient vowel production is the final goal of our CAPT tool. In a world without panaceas, it was only to be expected that our teaching tool would attain, like any other tool, a partial success.

Nevertheless, the experiment shows significant differences (Mann–Whitney U test with 99% confidence level) between following and not following the corrective feedback offered by the tool. Particularly, resorting to feedback made a clear difference in relation to the most difficult mode. Without feedback, only a 10% of success was registered at the PRO mode. In the DIS and MIX modes, success reached a 100% rate after feedback.

Table 4: Effectiveness of following the feedback suggested after failure in a mode versus not following it. DIS, PRO and MIX are Discrimination, Pronunciation and Mixed modes, respectively. Numbers between square brackets correspond to [passed, failed].

<table>
<thead>
<tr>
<th>Proposed modes</th>
<th>Completed modes</th>
<th>Mode repetitions</th>
<th>Mode repetitions with feedback</th>
<th>Mode repetitions without feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIS</td>
<td>60</td>
<td>51</td>
<td>10</td>
<td>6 [6,0]</td>
</tr>
<tr>
<td>PRO</td>
<td>60</td>
<td>35</td>
<td>61</td>
<td>40 [23,17]</td>
</tr>
<tr>
<td>MIX</td>
<td>60</td>
<td>43</td>
<td>34</td>
<td>12 [12,0]</td>
</tr>
</tbody>
</table>

4. Discussion and Conclusions

A controlled and guided protocol in a pedagogical CAPT tool with corrective feedback helps users to improve their pronunciation of isolated L2 words. Following the training cues offered by calculated feedback clearly makes a difference. Furthermore, without this feedback, some of the participants would have been probably unable to complete some of the task-tokens and modes.

A voice synthesis system used in the generation of pronunciation models proves useful in the process of helping students to improve their discrimination and production skills. The quality of the sound generated by the TTS was highly valued both by users and teachers; it is fully functional in the exposure and discrimination task-types, and satisfactorily orienting in the production mode. Although we have only tested the Android TTS, we are convinced that others like Microsoft TTS, Nuance TTS or Apple TTS would lead to similar results. The use of TTS systems is, in any case, not only possible but also advisable in CAPT tools.

Participants in the experiment were able to perform a large number of controlled learning tasks, to an extent that seems virtually impossible in the classroom. In this sense, our tool constitutes an adequate complementary resource in L2 pronunciation training. It allows for autonomous learning outside the classroom, and it might help students to achieve better results in a relatively short time.

To conclude, a final observation seems in order. Although we have focused mainly on the benefits of TTS feedback, it is important to notice that our pedagogical approach was specifically concerned with teaching students to produce the target sounds directly from the mental representations acquired through previous training; in other words, the easier Theory, Exposure and Discrimination modes were designed to guarantee a flawless single-round success in the Production and Mixed modes. First-round success rates were lower in Production than in any other mode, but still, the 58% mentioned above, after less than 3 hours of training, constitutes by no means the kind of result that disallows optimism. On the other hand, we have also emphasized the need to adapt to different learning styles. In this sense, TTS feedback manages to rescue those students for whom the method seems to be somewhat less effective. In this way, the users of our CAPT tool do not get permanently stuck in specific modes, the learning process becomes more dynamic and flexible and, in the end, better global results are achieved.

5. Acknowledgements

We would like to thank the Ministerio de Economía y Competitividad y Fondos FEDER – project key: TIN2014-59852-R Videojuegos Sociales para la Asistencia y Mejora de la Pronunciación de la Lengua Española.

6. References


Investigation of teacher-selected sentences and machine-suggested sentences in terms of correlation between human ratings and GOP-based machine scores

Yutaka Yamauchi\textsuperscript{1}, Junwei Yue\textsuperscript{2}, Kayoko Ito\textsuperscript{3}, Nobuaki Minematsu\textsuperscript{2}

\textsuperscript{1}Tokyo International University, Japan \hspace{1em} \textsuperscript{2}The University of Tokyo, Japan \hspace{1em} \textsuperscript{3}Kyoto University, Japan

yyama@tiu.ac.jp, \hspace{1em} jyue.mine@gavo.t.u-tokyo.ac.jp, \hspace{1em} ito.kayoko.6m@kyoto-u.ac.jp

Abstract

This study investigated relationships between teacher-selected stimulus sentences and machine-suggested ones in terms of the correlation between human ratings and GOP-based machine scores. In assessing shadowed speech consisting of 55 sentences recorded by 125 Japanese learners of English, it was examined which sentence combinations out of the 55 sentences could maximize the correlation between automatic scores and human ratings. A veteran teacher selected 10 sentences based on criteria such as sentence length, grammar, pronunciation, and prosodic features. The shadowed speech of the 10 sentences were manually rated by two native speakers of English focusing on pronunciation, prosody and lexical access (whether shadowing is done adequately or not after each word or phrase is identified). The same shadowed utterances were automatically assessed by the DNN-based GOP procedure. A significantly high correlation ($r=0.738$, $p<.01$) was found between manual ratings and automatic scores. Then groups of sentences were selected out of the original 55 sentences by greedy search (full search) so that correlation between their DNN-GOP scores and manual ratings of the selected 10 sentences could be maximized, and the top-ranked five combinations of groups of sentences were listed up. The number of shared sentences between the 10 teacher-selected sentences and machine-suggested ones in the top-ranked five combinations was only one, which was much smaller than expected, and thus the teacher’s strategies in selecting stimulus sentences turned out to be inappropriate in terms of maximizing correlation. Examining the sentence sets suggested by the machine revealed that some particular sentences frequently appeared in the sets and seemed to have something to do with maximizing correlation. Hence, the present study compared the 10 sentences selected by the teacher and his selection criteria with the sentence sets suggested by the machine and discussed what kind of sentence should be chosen to improve reliability in automatic assessment.

Index Terms: shadowing, automatic assessment, DNN-based GOP, teacher-selected stimulus sentence, machine-suggested stimulus sentence

1. Introduction

Shadowing is a simultaneous oral reproduction task. In shadowing learners are requested to listen to and comprehend model utterances usually read by a native speaker and simultaneously reproduce them orally as quickly and accurately as possible [1]. Learners have to quickly reproduce the utterances. When they try to shadow, the native speaker's sound images still remain in the learners' auditory memory. Therefore, they can easily imitate pronunciation, rhythm and intonation and then gradually get accustomed to speaking the target language with accurate pronunciation and prosodic features like native speakers.

Shadowing practice is expected to not only enhance listening comprehension and speaking skills, but also promote learners’ language processing to become more accelerated and automatized. As the model utterances are spoken at a faster rate, learners have to speed up in decoding auditory input information, comprehending the message and orally reproducing what they have heard. As a result, their language processing is thought to be changed from controlled into automatized. Shadowing is thought to include complex perception-production interaction and automatic semantic and syntactic processing [2-4]. Recent studies have reported the effectiveness of shadowing practice in second language (L2) learning in terms of listening comprehension skills [5-6], pronunciation, intonation and fluency [7-8] and overall proficiency [9].

2. Assessing shadowed utterances

Although shadowing practice seems to be prominent in developing L2 aural and oral skills, a serious problem lies in how to assess this performance objectively. In many cases the evaluator has to listen to the recorded shadowed utterances repeatedly while checking the script and calculate the ratio of the number of syllables or words correctly reproduced to the total number of syllables or words in the target passage [10]. This procedure is too time- and energy-consuming for teachers to implement this task in daily classroom activities.

To reduce the rater’s burden, an automatic evaluation system was developed by our research group using HMM-based phoneme posteriors called GOP (Goodness of Pronunciation) [11]. In this system a computer can automatically compare learners’ shadowed utterances with sounds expected from word strings in model utterances using an acoustic model stored in the PC at a phoneme level. This system can analyze and evaluate shadowed utterances and give scored feedback to the learner [12]. A significantly high correlation ($r=0.82$, $p<.01$) between GOP scores obtained from this system and manual scores by veteran language instructors was found. The GOP scores were also observed to be highly correlated with overall proficiency scores ($r=0.84$, $p<.01$) measured by TOEIC (Test of English for International Communication), one of the most popular standardized proficiency tests by ETS (Educational Testing Service). Thus the validity of the automatic evaluation system was confirmed [13].
3. Stimulus sentence selection

Selecting a limited number of proper sentences as stimulus sentences out of original passages is crucial to save time and energy in evaluating shadowed utterances for manual scoring. It also makes automatic assessment more effective. This study aims to investigate relationships between teacher-selected stimulus sentences and machine-suggested ones in terms of the correlation between human ratings and GOP-based machine scores. In other words, it is examined (1) how sentence combinations out of original sentences can maximize correlation between automatic scores and human ratings, (2) how teacher-selected sentences are different from ones suggested by the machine and (3) what kind of characteristics are observed in sentences forming combinations for maximizing correlation.

4. Speech data collection

Shadowed speech consisting of 55 sentences recorded by 125 Japanese learners of English were collected. An online recording site was developed for this data collection. The participants were requested to shadow the 55 model utterances without viewing any manuscripts or transcripts. Each model utterance was shadowed four times, and the fourth recorded utterance was used for data analysis. Four-time repetition was thought to get the learners accustomed to this task and lead to their best performances. Prior to shadowing and recording, they were asked to view an instruction page and practice shadowing and recording on the web.

5. Teacher-selected sentences

A veteran instructor teaching English over 30 years selected 10 sentences out of the original 55 sentences based on criteria such as sentence length, grammar, pronunciation and prosodic features as in Table 1. As sentence length becomes longer and more grammatical expressions like an embedded relative clause are included in stimulus sentences, syntactic and semantic difficulty increases in sentence processing. A sentence with a tag question or an alternative question sentence both require learners to choose proper intonation out of rising and falling options. As a result such complex sentences are expected to effectively differentiate good shadowers from those who are not. After the ten sentences were listed up, another veteran instructor checked these sentences based on her teaching experience and agreed that these sentences were appropriately selected as stimuli to differentiate learners’ performances.

The shadowed utterances of the 10 sentences were manually rated by two native speakers of English focusing on three criteria: pronunciation, prosody and lexical access (whether shadowing is done adequately or not after each word or phrase is identified). The manual rating was conducted with the five-point Likert scale ranging from 1 (worst) to 5 (best); a full mark was 15, and the worst score was 3 in total.

The same shadowed utterances were automatically assessed by the DNN (Deep Neural Network)-based GOP procedure. The reason for employing the DNN-based GOP in the present study is that DNN-based acoustic models have been reported to have better accuracy than traditional acoustic models, as long as a very large amount of data is provided for machine training [14-15].

A significantly high correlation ($r=0.738$, $p<0.01$) was found between manual ratings and automatic scores of the 10 sentences selected by the teacher.

6. Machine-suggested sentences

To select N sentences that can maximize the correlation between human ratings and machine scores, all the possible combinations of N sentences out of the 55 ones should be examined, and then the best combination can be detected. However, if we use ten for N, the number of possible combinations is almost infinite. Although it may not be theoretically impossible, implementing greedy search (full search) on such a large scale seems to be fairly impractical. For this reason we took three for N. However, we did not use only the top one combination but used the top 5, 10, 100 and 1000 combinations for further analysis in this investigation as shown in Tables 2 and 4.

Table 1: Mean scores of attributes of 10 teacher-selected sentences

<table>
<thead>
<tr>
<th>Ten teacher-selected sentences</th>
<th>sentence length</th>
<th>sentence complexity</th>
<th>T-unit length</th>
<th>prosody</th>
<th>automatic score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>word number</td>
<td>syllable number</td>
<td>T-unit number</td>
<td>word number</td>
<td>syllable number</td>
</tr>
<tr>
<td>I’m studying photography too. Shall we exchange some recent photos we’ve taken and discuss them on the internet?</td>
<td>18</td>
<td>26</td>
<td>2</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>The boy said that it had already been broken before he and his friend went to the house</td>
<td>18</td>
<td>20</td>
<td>1</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>It was you who kicked the door open, wasn’t it?</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Did you just want to have a bit of fun, or were you trying to get some money?</td>
<td>8</td>
<td>19</td>
<td>1</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>Then on February fourthteenth, two hundred seventy-six AD, a man named Valentine was killed by the Romans because of his Christian beliefs.</td>
<td>12</td>
<td>37</td>
<td>1</td>
<td>12</td>
<td>37</td>
</tr>
<tr>
<td>All of a sudden Valentine’s day became a big holiday for who made and sold cards.</td>
<td>17</td>
<td>24</td>
<td>1</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>The hospital doctors thought the men had been poisoned, but couldn’t work out what was wrong with them.</td>
<td>18</td>
<td>23</td>
<td>1</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>The reason the three men were taken to the hospital is because the puffier fish is also very poisonous.</td>
<td>19</td>
<td>29</td>
<td>1</td>
<td>19</td>
<td>29</td>
</tr>
<tr>
<td>Your brain still works perfectly however, so you know you are dying, but you can’t speak or do anything about it.</td>
<td>21</td>
<td>28</td>
<td>3</td>
<td>7</td>
<td>9.3333</td>
</tr>
<tr>
<td>Most people who die from eating fugu these days are people who have tried their hand at preparing the fish themselves.</td>
<td>21</td>
<td>28</td>
<td>1</td>
<td>21</td>
<td>28</td>
</tr>
</tbody>
</table>

Mean score: 16.7 | 24.4 | 1.3 | 13.9 | 21.2 | 1.7 | 0.278
Three sentences were selected out of the original 55 sentences by greedy search (full search) so that correlation between the DNN-GOP scores and manual ratings of the selected 10 sentences could be maximized. The top five combinations of three sentences are listed up in Table 2. The correlation values of the top five combinations ranged from 0.803 to 0.793, all of which outperformed the correlation ($r=0.738$, $p<0.01$) between human ratings and DNN-based GOP scores of the 10 sentences from the teacher.

Based on the comparison of Table 1 with Table 2, only one sentence “THE BOY SAID THAT IT HAD ALREADY BEEN BROKEN BEFORE HE AND HIS FRIEND WENT TO THE HOUSE.” was identical. Just one out of the 10 sentences selected by the teacher was included in the top five combinations of three sentences for maximizing correlation. The same tendencies were observed in the top 10 and 100 combinations of three sentences. As just described, it was found that the number of shared sentences between teacher-selected sentences and machine-suggested ones was much smaller than expected, and thus the teacher’s strategies in selecting stimulus sentences turned out to be inappropriate in terms of maximizing correlation.

7. Characteristics of machine-suggested sentences

7.1 Tendencies in machine-suggested sentence sets

Examining sentence sets suggested by the machine showed that some particular sentences frequently appeared in the sets and seemed to have something to do with maximizing correlation. In Table 2, for instance, the sentence beginning with “HELLO CAROL.” showed up five times, and the sentence starting with “THE POLICE OFFICER ASKED” appeared three times. The same tendency was observed in the bottom (worst) five combinations of three sentences for maximizing correlation as shown in Table 3. The sentence “HI, MY NAME IS AKIRA.” showed up five times. The reason why “HELLO CAROL. I SAW YOUR HOMEPAGE AND LIKE IT A LOT. YOUR PHOTO WAS REALLY SOMETHING.” was treated as one sentence was that these three were presented to participants at one time in a shadowing session, and they were required to shadow these three utterances all together. On the other hand, “HI, MY NAME IS AKIRA.” in Sentence 1 in Table 3 was presented to the participants, and they were asked to shadow only this sentence at a given time.

Table 2: Top 5 combinations of three sentences to maximize correlation

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>Sentence 2</th>
<th>Sentence 3</th>
<th>sentence length</th>
<th>sentence complexity</th>
<th>T-unit length</th>
<th>prosody</th>
<th>automatic score</th>
<th>DNN-GOP</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 HI, MY NAME IS AKIRA.</td>
<td><strong>HE FINDS A BOY STANDING NEARBY.</strong></td>
<td>THE POLICE OFFICER ASKED, “WHY WERE YOUR FINGERPRINTS FOUND ALL OVER THE DOOR?”</td>
<td>9.0</td>
<td>11.3</td>
<td>1.0</td>
<td>9.0</td>
<td>11.33</td>
<td>0.3</td>
<td>0.358</td>
</tr>
<tr>
<td>2 HI, MY NAME IS AKIRA.</td>
<td><strong>WHENEVER IT IS ATTACKED, THE FISH PUFFS UP ITS BODY TO OVER TWICE ITS NORMAL SIZE.</strong></td>
<td>FUGU IS SAID TO BE SO DELICIOUS THAT IT HAS EVEN STARTED TO BE IMPORTED INTO HONG KONG AND THE UNITED STATES.</td>
<td>14.3</td>
<td>19.3</td>
<td>1.0</td>
<td>14.3</td>
<td>19.33</td>
<td>0.3</td>
<td>0.350</td>
</tr>
<tr>
<td>3 HI, MY NAME IS AKIRA.</td>
<td><strong>THEY COULDN’T SPEAK, AND THEY HAD TROUBLE BREATHING.</strong></td>
<td>FUGU IS SAID TO BE SOMETHING THAT HAS BEEN POPULAR AS GIFTS.</td>
<td>8.3</td>
<td>8.7</td>
<td>1.0</td>
<td>8.3</td>
<td>8.87</td>
<td>0.3</td>
<td>0.339</td>
</tr>
<tr>
<td>4 HI, MY NAME IS AKIRA.</td>
<td><strong>WHENEVER IT IS ATTACKED, THE FISH PUFFS UP ITS BODY TO OVER TWICE ITS NORMAL SIZE.</strong></td>
<td>MOST PEOPLE WHO DIE FROM EATING FUGU THESE DAYS ARE PEOPLE WHO HAVE TRIED THEIR HAND AT PREPARING THE FISH THEMSELVES.</td>
<td>14.0</td>
<td>18.0</td>
<td>1.0</td>
<td>14.0</td>
<td>18.00</td>
<td>0.3</td>
<td>0.343</td>
</tr>
<tr>
<td>5 HI, MY NAME IS AKIRA.</td>
<td><strong>HE WANTS TO KNOW HOW THE DOOR OF THE MACDONALD’S HOUSE WAS BROKEN OPEN.</strong></td>
<td>THE POLICE OFFICER ASKED, “WHY WERE YOUR FINGERPRINTS FOUND ALL OVER THE DOOR?”</td>
<td>8.3</td>
<td>10.3</td>
<td>1.0</td>
<td>8.3</td>
<td>10.33</td>
<td>0.3</td>
<td>0.342</td>
</tr>
</tbody>
</table>

Table 3: Bottom (Worst) 5 combinations of three sentences to maximize correlation

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>Sentence 2</th>
<th>Sentence 3</th>
<th>sentence length</th>
<th>sentence complexity</th>
<th>T-unit length</th>
<th>prosody</th>
<th>automatic score</th>
<th>DNN-GOP</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 HI, MY NAME IS AKIRA.</td>
<td>HE FINDS A BOY STANDING NEARBY.</td>
<td>WHENEVER IT IS ATTACKED, THE FISH PUFFS UP ITS BODY TO OVER TWICE ITS NORMAL SIZE.</td>
<td>9.0</td>
<td>11.3</td>
<td>1.0</td>
<td>9.0</td>
<td>11.33</td>
<td>0.3</td>
<td>0.358</td>
</tr>
<tr>
<td>2 HI, MY NAME IS AKIRA.</td>
<td><strong>WHENEVER IT IS ATTACKED, THE FISH PUFFS UP ITS BODY TO OVER TWICE ITS NORMAL SIZE.</strong></td>
<td>FUGU IS SAID TO BE SOMETHING THAT HAS BEEN POPULAR AS GIFTS.</td>
<td>14.3</td>
<td>19.3</td>
<td>1.0</td>
<td>14.3</td>
<td>19.33</td>
<td>0.3</td>
<td>0.350</td>
</tr>
<tr>
<td>3 HI, MY NAME IS AKIRA.</td>
<td><strong>HE WANTS TO KNOW HOW THE DOOR OF THE MACDONALD’S HOUSE WAS BROKEN OPEN.</strong></td>
<td>THE POLICE OFFICER ASKED, “WHY WERE YOUR FINGERPRINTS FOUND ALL OVER THE DOOR?”</td>
<td>8.3</td>
<td>10.3</td>
<td>1.0</td>
<td>8.3</td>
<td>10.33</td>
<td>0.3</td>
<td>0.342</td>
</tr>
<tr>
<td>4 HI, MY NAME IS AKIRA.</td>
<td><strong>HELLO CAROL. I SAW YOUR HOMEPAGE AND LIKE IT A LOT. YOUR PHOTO WAS REALLY SOMETHING.”</strong></td>
<td>THE POLICE OFFICER ASKED, “WHY WERE YOUR FINGERPRINTS FOUND ALL OVER THE DOOR?”</td>
<td>15.7</td>
<td>20.3</td>
<td>1.7</td>
<td>12.1</td>
<td>15.4</td>
<td>1.0</td>
<td>0.297</td>
</tr>
<tr>
<td>5 HI, MY NAME IS AKIRA.</td>
<td><strong>THE POLICE OFFICER ASKED, &quot;WHY WERE YOUR FINGERPRINTS FOUND ALL OVER THE DOOR?&quot;</strong></td>
<td>GLOVES, CHOCOLATE AND EVEN UNDERWEAR HAVE ALL BEEN POPULAR AS GIFTS.</td>
<td>15.0</td>
<td>20.0</td>
<td>1.7</td>
<td>11.4</td>
<td>15.1</td>
<td>1.3</td>
<td>0.005</td>
</tr>
</tbody>
</table>
7.2 Three attributes of selected sentences

To investigate what kind of characteristics are objectively observed in sentences forming combinations for maximizing correlation, three attributes were added for each of the original 55 sentences: sentence length, sentence complexity and number of liaisons.

Sentence length was calculated by the number of words and syllables in the sentence. Sentence complexity was measured by T-unit length. T-unit (minimal terminable unit) is defined as a main clause plus all subordinate clauses and nonclausal structures attached to or embedded in it [16-17]. T-unit length calculated by the number of words or syllables per T-unit has been reported to indicate sentence complexity and has been widely used in research of applied linguistics to date [18-19]. One T-unit contains only one main clause, which is counted as one. No matter how many subordinate clauses may follow a main clause, none of the subordinate clauses are counted. For example, the number of T-units in sentence A is one, and its T-unit length is four. The number of T-units in sentence B is one and its T-unit length is ten. The number of T-units in sentence C is two, because two main clauses are connected by the conjunction "and." As a result, the number of T-units in sentence C is two, and its T-unit length is five (ten divided by two).

- Sentence A: Her eyes are blue.
- Sentence B: When the movie is over, we are going straight home.
- Sentence C: The police officer blew his whistle and the truck stopped.

7.3 Comparison of top and bottom combinations of three sentences

The number of possible combinations in choosing three sentences out of the 55 original sentences is 26,235. To observe how the three attributes mentioned above are different between the top- and bottom-ranked combinations of three sentences, the top10, 100 and 1000 and the bottom 10, 100 and 1000 combinations were listed up, and the mean scores of their three attributes are shown in Table 4.

8. Discussion

In this study two correlations were compared: correlation A between human ratings and DNN-based GOP scores of the 10 sentences selected by the teacher and correlation B between human ratings of the 10-sentence set by the teacher and DNN-based GOP scores of three-sentence combinations chosen out of the original 55 sentences by greedy search (full search) for maximizing correlation. Since correlation A was calculated based on the same 10-sentence set, this can be regarded as a closed result. On the other hand, correlation B was computed based on two different groups: the teacher-selected 10-sentence set and the three-sentence sets chosen by greedy search (full search). The number of combinations of three sentences is 26,235. In that sense, correlation B can be thought of as a partially open result compared to correlation A.

The correlation values in the open result in Table 2 outperformed the closed result (correlation value, r=0.738, p<0.01). This implies that machine-suggested sentence sets could be more suitable for automatic assessment of L2 shadowing than teacher-selected sentence sets.

To investigate what kind of sentence combinations maximizes/minimizes the correlation, 10, 100 and 1000 combinations were chosen from the top and bottom. To analyze some characteristics of selected sentences, three criteria were set up: sentence length, sentence complexity and number of liaisons. The mean scores of each attribute were calculated across seven groups: top 10, 100 and 1000 combinations, bottom (worst) 1000, 100 and 10 combinations and 10 teacher-selected sentences as in Table 4.

Table 4: Mean scores of attributes of selected sentences between top and bottom (worst) 10, 100 and 1000 combinations

<table>
<thead>
<tr>
<th>sentence length</th>
<th>sentence complexity</th>
<th>T-unit length</th>
<th>prosody</th>
<th>automatic score</th>
</tr>
</thead>
<tbody>
<tr>
<td>word number</td>
<td>syllable number</td>
<td>T-unit number</td>
<td>word number</td>
<td>syllable number</td>
</tr>
<tr>
<td>Top</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>16.2</td>
<td>24.4</td>
<td>1.3</td>
<td>13.9</td>
</tr>
<tr>
<td>STD</td>
<td>1.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>10 Mean</td>
<td>16.2</td>
<td>24.4</td>
<td>1.3</td>
<td>13.9</td>
</tr>
<tr>
<td>STD</td>
<td>1.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>100 Mean</td>
<td>16.2</td>
<td>24.4</td>
<td>1.3</td>
<td>13.9</td>
</tr>
<tr>
<td>STD</td>
<td>1.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>1000 Mean</td>
<td>16.2</td>
<td>24.4</td>
<td>1.3</td>
<td>13.9</td>
</tr>
<tr>
<td>STD</td>
<td>1.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Bottom (Worst)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>16.2</td>
<td>24.4</td>
<td>1.3</td>
<td>13.9</td>
</tr>
<tr>
<td>STD</td>
<td>1.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>10 Mean</td>
<td>16.2</td>
<td>24.4</td>
<td>1.3</td>
<td>13.9</td>
</tr>
<tr>
<td>STD</td>
<td>1.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Discussion
In the top three groups, average scores of sentence length calculated by the number of words were 14.8, 14.5 and 14.0, respectively. Mean scores of sentence length by the number of syllables were 20.8, 21.2 and 20.8, respectively. On the other hand, in the bottom three groups, average scores of sentence length by the number of words were 10.5, 10.2 and 9.8, respectively. These results show that three-sentence combinations consisting of shorter sentences led to lower correlation in the open result. In other words, if stimulus sentences are too short, they are too easy for participants to shadow, and thus they are not suitable for automatic assessment.

To check how well the participants shadowed across the seven groups, the DNN-based GOP scores were calculated. Table 4 shows that the DNN-GOP scores in the three bottom (worst) groups were 0.329, 0.341 and 0.343, respectively, whereas, those in the three top groups were 0.298, 0.295 and 0.293. The higher the DNN-GOP scores were, the better the participants shadowed. Therefore, it could be said that sentences consisting of around 14 words were more suitable as stimulus sentences in these shadowing tasks than those of around 10 words, based on the comparison of sentence length between the top and bottom three combinations in Tables 2, 3 and 4.

The average sentence length of the 10 teacher-selected sentences was 16.2, which is longer than the mean sentence length in the top three groups (around 14) in Table 4. This implies that the teacher-selected sentences were so difficult that participants did not do well in those sentences, and consequently the number of shared sentences between teacher-selected sentences and machine-suggested ones was much smaller than expected.

In the same procedure above, other attributes were calculated and analyzed as follows (Table 4):

The average score of sentence length by the number of syllables in the top three groups was around 20. That in the bottom three groups was around 14. That in the 10 teacher-selected sentences was 24.4.

The mean T-unit score indicating the number of main clauses in the top three groups was around 1.6. That in the bottom (worst) three groups was around 1.1. That in the teacher-selected 10 sentences was 1.3.

T-unit length was calculated by the total number of words or syllables divided by the number of T-units included in a sentence. The average scores of T-unit length by the number of words and those by the number of syllables in the top three groups were around 11 and 16, respectively. Those in the bottom three groups were around 10 and 13, respectively. Those in the 10 teacher-selected sentences were 13.9 and 21.2, respectively.

The average number of liaisons included in a sentence in the top three groups was around 0.8. That in the bottom (worst) three groups was around 0.4. That in the teacher-selected 10 sentences was 1.2.

The average DNN-based GOP scores in the top three groups were around 0.29. Those in the bottom three groups were around 0.34. That in the teacher-selected 10 sentences was 0.287.

Considering these results, sentences included in the bottom groups were easy or too easy for participants to shadow. On the other hand, the 10 teacher-selected sentences were difficult or too difficult for L2 shadowing.

9. Conclusions and future studies

Based on this investigation it could be concluded that stimulus sentence combinations for automatic assessment of L2 shadowing should not be too easy or too difficult. Stimulus sentences should possess moderate complexity in terms of the number of words and syllables and sentence construction complexity measured by T-unit length. These are thought to be required conditions to improve reliability in automatic assessment of L2 shadowing.

Although this conclusion is not so surprising, the results of this investigation might deepen and broaden the insights of teachers and educators into selecting proper materials for assessment. In many cases teachers tend to choose syntactically and semantically difficult sentences as stimuli. Teachers might have the preconception that in L2 shadowing more difficult sentences tend to have more power to discriminate learners’ proficiency, lead to learners’ best performances and reveal what they can do using their utmost linguistic abilities. However, the results of this investigation did not support such a preconceived idea.

In language education stimulus sentences are usually chosen and determined based on veteran teachers’ knowledge and experiences. It is very rare that stimulus sentences are selected systematically by a machine and compared to teacher-selected sentences. The results of this investigation were obtained by objective and systematic analysis, which has rarely been conducted in language education and might shed light on altering teachers’ beliefs in assessment.

L2 shadowing is thought to be a very cognitively demanding task, because it requires learners to do two things simultaneously: listening comprehension and oral reproduction. Therefore, L2 shadowing itself might be difficult enough to differentiate learners’ proficiency, and thus stimuli should not be so difficult.

As for future studies it is very crucial to conduct the same experiment using passages different from those used in this study and examine whether the results obtained from this investigation are dependent on the given sentence sets or not. In other words, if stimulus sentences are selected from different passages to meet the required conditions proposed by the present study, will the assessment result in the same way as in this investigation? After this probing work, conditions obtained from this study can be considered as sufficient conditions as well as necessary ones, and the conclusions will be more generalized.

10. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP16H03084, JP16H03447, and JP26240022. We’d like to thank all students and educators who participated in this experiment, including Prof. Kay Husky (Tokyo International University) for her professional advice on this paper.

11. References


Towards a Tool for Automatic Spelling Error Analysis and Feedback Generation for Freely Written German Texts Produced by Primary School Children

Ronja Laarmann-Quante
Ruhr-University Bochum, Germany
laarmann-quante@linguistics.rub.de

Abstract
This paper proposes a tool for the automatic analysis of spelling errors in freely written German texts. It is based on automatic annotations of spelling errors that comprise various levels, such as linguistic properties of the target word (phonemes, syllables, morphemes) and error-related properties such as error categories which mark whether the misspelling changes the pronunciation of a word or whether the correct spelling can be derived from a related word form. These can be used to create an application that could, for example, help teachers analyze their students’ orthographic skills and give feedback with little manual effort. For the future, it could also be implemented as an automatic tutoring system for children in which case the surface has to be child-oriented and should present error analysis as a kind of game. While the paper presents the capabilities of a first prototype, the concrete implementation for real-world use is open for discussion with experts on orthography instruction.

Index Terms: spelling errors, German, free text, primary school, individual diagnosis, automatic analysis, feedback

1. Introduction
When children learn to write, it is essential that feedback about spelling mistakes is relevant and helpful. A recent study with students of grades 3-5 in German schools has shown that an individual qualitative analysis of the spelling errors in freely written texts from children is a very good basis for providing feedback and fostering the children’s spelling competence [1, 2]. However, even in primary schools, one teacher has to take care of 13-29 children in one class [3], which in reality makes it almost impossible to analyze every child’s misspellings in detail. Furthermore, some teachers and teachers-to-be in Germany have been shown to have deficits in knowledge about the German writing system or in using this knowledge in didactic contexts [4, 5]. It is crucial, though, that feedback about a spelling is not misleading. For instance, in German, a careful pronunciation of a word is often the basis for the correct spelling but not always: the word <Kinder> ‘children’ is pronounced [Kində] in standard German, with the last phone sounding similar to [a]. If a child wrote *<Kinda> instead, it would be wrong to tell the child that it has to pronounce the word more carefully in order to ‘hear’ the correct spelling [4].

In order to support children’s spelling acquisition more individually, automatic analyses of spelling errors with the help of natural language processing methods could reduce the time and cognitive effort that teachers need for a thorough assessment of a child’s text. A close examination of the strengths and weaknesses of a child is then a good basis for further training. This paper proposes a tool which automatically categorizes spelling errors and provides additional information about them which can be used for preparing feedback or individual training material. We present a first prototype which could be adapted to the specific needs of teachers. For the future, if the automatic analysis becomes robust and reliable enough, it is also conceivable to design a child-oriented surface which presents the writing of a free text and a subsequent analysis and correction of spelling errors as a game.

An application with automatic error categorization for diagnostic purposes for German has been developed before [6, 7] but its focus is only on a small number of error categories and it does not provide feedback beyond statistics about possible and committed errors. Our goal is to facilitate a systematic analysis of all the errors in a text and to take different features beyond a one-dimensional error categorization into account. For instance, we want the application to automatically provide feedback about whether the pronunciation of the word changes with a misspelling or how one could otherwise arrive at the correct spelling using a previously developed annotation scheme for spelling errors and its automatic application [8, 9]. The aim of this paper is to show how these rather abstract annotations can be presented in a user-friendly way so that they could be used e.g. in a school setting.

The paper is structured as follows: Section 2 gives a short overview of word spelling in German and summarizes the annotation scheme which forms the basis of the spelling error analysis. Section 3 explains how the annotations are obtained automatically, before Section 4 presents a first prototype of how the annotations can be visualized in a user-friendly application. Section 5 concludes the paper with a summary and outlook.

2. Background: German Orthography and Spelling Error Annotation
The basis of German word spellings are grapheme-phoneme correspondences (GPCs).¹ For instance, the word kalt ‘cold’ can be spelled phonographically, which means that each phoneme in the pronunciation [kalt] corresponds to exactly one grapheme in the correct spelling <kalt>.² The correspondences can be read off from basic (context-independent) GPC-rules such as /k/ → <k> etc. However, there are phenomena which overwrite these basic correspondences, for example consonant doubling (<rennen> [rənən] ‘to run’) or that final devoicing is not reflected in the spelling ([hʊnt] is spelled <Hund> ‘dog’). These can be explained via the word’s syllabic/ prosodic structure (e.g. consonant doubling occurs in the spelling if in the pronunciation a single consonant occurs in a syllable joint

¹The explanations of German word spelling follow [10, 11] in this paper.
²The pronunciation of a word in IPA is presented in square brackets, the spelling in angle brackets and phonemes in slashes.

36
between a stressed lax vowel and an unstressed vowel) or with reference to a related word form: The reference form of the word *Hund* ['hʊnt] is a disyllabic word form such as the plural *Hunde* ['hʊntd] which is spelled *<Hunde>* phonographically. The spelling usually remains constant in all related word forms (principle of morpheme constancy).

Spelling errors can be analyzed on different levels. For German, the classic approach which is used by standardized spelling tests such as the HSP [12] and by other popular spelling error analysis schemes such as AFRA [13] and OLFa [14] for freely written texts, is to categorize errors one-dimensionally by phenomena such as consonant doubling or final devoicing. However, with regard to giving feedback about errors, it is also important to take other features into account as well.3 For instance, some misspellings change the pronunciation of a word (in *<renen>* the first vowel is pronounced long and in *<renen>* short) while others do not (*<Hunt>/*Hund*> are both pronounced with a voiceless obstruent because of final devoicing), some spellings can be derived from related word forms (*<Hund>* from plural *Hunde*> while others cannot. To capture all these differences, we developed a multi-level annotation scheme for misspellings. For the annotation, the target word, i.e. the orthographically correct version of the word that the child wanted to write, has to be known. The annotation scheme then codes the following information:4 For each target word we annotate

- graphemes
- phonemes
- syllables
- morphemes
- whether it structurally belongs to the German core vocabulary or can be seen as a ‘foreign word’ (according to [15])

For each misspelled word, i.e. the original spelling of the child, we annotate whether the child’s spelling

- resulted in another existing German word5 (e.g. *<feld>* ‘field’ for *<fält>* ‘(he/she) falls’): it might be interesting to know whether a child produces forms that he or she may have encountered before (e.g. in a book) vs. forms he or she has never seen before in other texts
- has a plausible syllable structure or violates graphotactic constraints: this can give a hint whether a child has a sense of the structure of German words

Finally, for each error (there can be more than one error in a word) we annotate

- error category
- whether the error influences the pronunciation of the word
- whether the correct spelling can be derived from a related word form

The annotations are stored in a custom-made XML format called LearnerXML [8] and can be visualized in EXMARaLDA [16, 17]. Figure 1 shows a screenshot from EXMARaLDA of all the annotations of the misspelling *<feind>* for *<weint>* ‘(he/she) cries’.6

```xml
<table>
<thead>
<tr>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>tokens.orig</td>
<td>feind</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tokens.target</td>
<td>weint</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[metamarks]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[foreign_target]</td>
<td>false</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[exist_orig]</td>
<td>true</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[characters_orig]</td>
<td>f e i n d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[characters_target]</td>
<td>w e i n t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[phonemes_target]</td>
<td>v al i n t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[graphemes_target]</td>
<td>w e i n t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[syllables_target]</td>
<td>stress</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[syll_orig_plausible]</td>
<td>true</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[morphemes_target]</td>
<td>NN INFL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[error_cat[1]]</td>
<td>PGII:voice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[phon_orig_ok[1]]</td>
<td>false</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[morph_const[1]]</td>
<td>na</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[error_cat[2]]</td>
<td>MO:hyp_final_devoicing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[phon_orig_ok[2]]</td>
<td>true</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[morph_const[2]]</td>
<td>nocus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[error_cat[3]]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Figure 1: Screenshot of EXMARaLDA with the annotations for the misspelling *<feind>* for *<weint>*

One can see that the target word is not a foreign word (foreign_target = false) and that the misspelling resulted in another existing word (exist_orig = true; Feind = ‘enemy’). Moreover, one can see that there are two errors in the word. The first one, *<f>* for *<w>*), has to do with voicing (error_cat[1] = PGII:voice) and changes the pronunciation of the word (phon_orig_ok[1] = false). Morpheme constancy does not explain the correct spelling here (morph_const[1] = na). The second mistake, in contrast, concerns a substitution of *<t>* with *<d>*), which is a hypercorrection of final devoicing (error_cat[2] = MO:hyp_final_devoicing). This does not affect the pronunciation of the word and morpheme constancy does play a role in that the child could have known that the word was spelled with a final *<t>* if it had recognized that this was an inflectional suffix that is always spelled *<t>*.

A more detailed description of the annotation scheme and the XML format can be found in [8]. In a current research project, we analyze a corpus of freely written texts (descriptions of picture stories) produced by primary school children of grades 2-4 that was collected by [18]. We want to investigate the relationship between spelling errors and a word’s properties, including the information coded in the multi-layered annotation scheme presented above. In order to be able to annotate a large number of texts (in total we have over 1800 texts that were manually transcribed and corrected for spelling errors), we implemented an automation of the annotations in the scheme, which will be described in the next section. Thereby, we found that the analysis of orthographic errors can be carried out well with a rule-based approach. This has the advantage that the analyses do not depend on training material and can therefore be applied to any written text. The annotations of a text form the backbone of the application described in Section 4, which presents

--

3AFRA sometimes allows one error to fall into multiple categories, which highlight different aspects, but this is not possible for all errors.

4The complete annotation scheme with all categories can be found under https://www.linguistics.ruhr-uni-bochum.de/litkey/Scientific/CorpusAnalysis/Resources.html.

5Capitalization is not taken into account here.

6Phonemes are represented in SAMPA notation here, see http://www.phon.ucl.ac.uk/home/sampa/german.htm.
an analysis of all the spelling errors in a text in a way that is suitable for e.g. a teacher or other language trainer rather than a corpus linguist.

3. Automatic Spelling Error Analysis

The automatic annotation of a word’s properties according to our annotation scheme described in the previous section is carried out with the help of the web service G2P of the Bavarian Archive of Speech Signals (BAS)\(^7\) [19, 20] and our own processes. These are described in [9] but to convey how the spelling analysis of our proposed application works, the basic procedures will be sketched here again. The G2P web service provides the phonemes, syllables and morphemes for any given word. For the word \textit{fröhlich} [frʊˈliː] ‘happy’, for example, it returns

- phonemes (in SAMPA) \textit{fr2:1IC}
- syllable boundaries (\textit{)} + stress marks (\textit{’}) \textit{fr’2:\;1IC}
- morpheme boundaries \textit{fröh lich}
- morpheme tags \textit{ADJ SFX}

In German, units of more than one letter can correspond to a single phoneme (e.g. \textit{fröhlich}. \textit{<öh} corresponds to \textit{[œː]} in \textit{[œːh]}). To obtain these character units which we call phoneme-corresponding units or PCUs\(^3\), we firstly use string alignment via weighted Levenshtein distance.\(^4\) The resulting 1:1 (or 1:0, 0:1) correspondences between characters and phonemes are then rearranged via rules to get the correct PCUs.

The automatically obtained information about graphemes, phonemes, syllables and morphemes are then used to determine, for a given target word, which of the systematic spelling errors that our scheme defines, \textit{could} be committed in this word. Currently, there are 65 error categories in total (+ 4 ‘other’ categories), of which 62 are implemented. For instance, the category \textit{devoice} (final devoicing), which describes an error such as \textit{maun phriGIS ok} (whether the misspelling changes the pronunciation of the word) and morph\_const (whether the correct spelling can be derived from a related word form) can be read off from the error category, sometimes connected with further restrictions. For example, for \textit{devoice} (see *\textit{maun phriGIS ok} is always true, i.e. by definition, errors of this category never change the word’s pronunciation. For morph\_const, the syllables and morphemes play a role: For this error category, deriving the correct spelling from a related word form is said to be always possible/necessary if the devoiced letter is in final position of the syllable coda of an inflecting morpheme. This captures that words that do not inflect (such as \textit{'und'} and \textit{‘and’}) do not have related word forms and that in words where the devoiced consonant is not in final position of the syllable coda (such as \textit{Obst} ‘fruit’), the consonant is devoiced in \textit{all} related word forms, hence they are of no help. For further information on these features, see [9].

In order to annotate \textit{exist\_orig} (does the misspelling result in another existing German word form?), one has to approximate what ‘existing word forms’ for children are, i.e. which words they are likely to have encountered before. For this purpose, we use an extract from \textit{childLex}, the German Children’s Book Corpus [22]. In order to exclude words which are specific to only particular books or book series, we extracted those words which occurred in at least 10 of the 500 books that were used for the compilation of the corpus. Besides these extracted words themselves, we also included all their related word forms, i.e. those with the same lemma. Furthermore, we did some corpus cleaning in that words in uppercase letters as well as lowercased words which were tagged as a noun and capitalized words which were not tagged as a noun were removed. The final list includes just over 39,000 word forms. For all misspellings which resulted in one of these word forms, \textit{exist\_orig} is tagged as true.

\(^{10}\)The misspelling resulted in another German word meaning ‘(he/she) spoiled’ (stil. for \textit{sho}).

\(^{11}\)The abbreviations \textit{PGIII} and \textit{SL} before the colon mean that the errors refer to the level of grapheme-phoneme correspondence and the syllabic level, respectively. For more information see [8].
The two remaining layers of our annotation scheme, namely foreign_target (is the target word a foreign word?) and syll_orig_plausible (does the syllable in the original spelling adhere to graphotactics?), are currently being implemented.

A big issue when talking about automatic analyses or annotations is of course the reliability. So far, we have studied the agreement between human and automatic annotations for the error category, phon_orig_ok and morph_const (see [9]). Three human annotators (students) and the automatic system annotated 11 texts (866 target tokens) taken from our corpus of freely written children’s texts addressed in Section 2. Table 1 (extracted from the result table presented in [9]) shows the average agreement of human and automatic annotations and among human annotations only.

<table>
<thead>
<tr>
<th>Level</th>
<th>Size</th>
<th>Auto/Human perc.</th>
<th>Human perc.</th>
<th>(\kappa)</th>
<th>Auto/Human perc.</th>
<th>Human perc.</th>
<th>(\kappa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>error cat.</td>
<td>261</td>
<td>78.67%</td>
<td>77.39%</td>
<td>.82</td>
<td>77.74%</td>
<td>77.74%</td>
<td>.82</td>
</tr>
<tr>
<td>phon_orig_ok</td>
<td>227</td>
<td>80.35%</td>
<td>78.03%</td>
<td>.72</td>
<td>78.26%</td>
<td>78.26%</td>
<td>.72</td>
</tr>
<tr>
<td>morph_const</td>
<td>227</td>
<td>79.54%</td>
<td>70.99%</td>
<td>.55</td>
<td>79.46%</td>
<td>70.99%</td>
<td>.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>Size</th>
<th>Auto/Human perc.</th>
<th>Human perc.</th>
<th>(\kappa)</th>
<th>Auto/Human perc.</th>
<th>Human perc.</th>
<th>(\kappa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>error cat.</td>
<td>170</td>
<td>72.16%</td>
<td>72.94%</td>
<td>.79</td>
<td>72.02%</td>
<td>72.02%</td>
<td>.79</td>
</tr>
<tr>
<td>phon_orig_ok</td>
<td>145</td>
<td>80.26%</td>
<td>78.03%</td>
<td>.72</td>
<td>80.26%</td>
<td>78.03%</td>
<td>.72</td>
</tr>
<tr>
<td>morph_const</td>
<td>145</td>
<td>79.46%</td>
<td>70.99%</td>
<td>.55</td>
<td>79.46%</td>
<td>70.99%</td>
<td>.55</td>
</tr>
</tbody>
</table>

Table 1: Average inter-annotator agreement (raw percent and Fleiss’ \(\kappa\)) between human and automatic annotations and among human annotations only.

Since errors of category SN (syntactically determined errors, i.e. capitalization and writing words together or separately) are trivial to detect when the target word is known, the table also presents agreement figures disregarding errors from this category to get a sense of how well the more challenging cases are handled. The features morph_const and phon_orig_ok were only evaluated if the error category was the same for all human annotators (and the system, respectively). As one can see, agreement with the automatic annotations is quite comparable to the agreement among the human annotations. We are currently working on a gold standard for further evaluation results and our automatic procedures are still being improved by considering the sources of disagreements that were discussed in [9]. We are also currently evaluating the information about phonemes, syllables and morphemes obtained from the BAS web service, which is also not 100% correct as one can see in Figure 1, where the stem of the verb weinen is analyzed as a noun. As they form the basis of our error analysis, the correctness of these information in part determines the correctness of our error annotation, and hence the feedback that is generated about a spelling, as well.

Generally, one can see that it is not trivial to interpret spelling errors. As indicated in the introduction, also teachers, whose assessment of spelling errors can have a direct influence on the feedback given to the child and the evaluation of his/her spelling competence, do not always analyze spelling errors correctly. Hence, a “joint effort” of human and automatic analyses should be even more reliable than analyses made by teachers “from scratch” in a school context. With a precise allocation of the error(s) in a misspelled word and various information about the target word and the error, it is possible to create an application which uses these features to create individual diagnoses and feedback. The main task is to find a suitable visualization for the target users and to present the relevant information accordingly. The next section presents a very first prototype of how the rather abstract annotations can be visualized in an application that can be used for example by teachers in order to assess a child’s spellings in a freely written text.

4. Application Prototype

While a manual analysis of spelling errors is very time-consuming, the automatic tool analyzes the spelling errors in the input text within seconds. Figure 2 shows the main page of the application with an example text from our corpus addressed.

---

12Currently, both the original spelling and the target spelling of a word have to be provided as input for the tool. However, we are working on a robust automatic spell checking procedure for children’s writings so that at one point the manual effort of correcting the texts will be reduced to counter-checking the automatic corrections.
in Section 2. Since in our project we are studying orthography errors only, grammatical errors are not corrected and addressed in the texts, and are not part of our annotation scheme. We are planning to deal with grammatical errors in the future.

At the top of the page, one can see some general statistics regarding the number of words and errors in the text. The box on the left hand side displays the input text with each target word printed in gray under the original spelling. Each error is marked in red in the original word. This is achieved with the procedure described in Section 3 that allows a precise allocation of an error. On the right hand side of the page one can see a table of error categories. The column “wrong” shows how many times an error of a particular category was committed and “all” shows how many times this error could have been committed (“base rate” [13, 23]). In that our automatic error analysis first determines which errors could theoretically be committed in a given word before committed errors are interpreted, we automatically obtain the base rate for each category. Errors which can occur in every word (such as capitalization, here low-up or every letter (like unsystematic replacement or deletion of a letter, here repl_v and ins_C) are not given a base rate. What you can see in this table, for instance, is that there are nine occurrences of voc_r, i.e. vocalized <r> (as in [kind]/[Kinder] ‘children’) in the text and three of them were misspelled. As [6] emphasizes, it is important to have a “dual view” on the absolute and relative error rate of a category. Therefore, two different visual markers are used to be able to see the prominence of an error category without effort: The categories are ranked by the absolute number of errors with the category with most errors put on top of the table (errors without base rate are currently displayed under those with a base rate). The relative error rate of a category is indicated with colors: red means the ratio of committed and possible errors is ≥ .5, yellow means < .5 and green means that no error was produced in this category.

Figure 3: Page showing all errors (red) and correct applications (green) of r-vocalization (voc_r) in the learner’s text

While the main page gives a quick overview of the main problem areas in the child’s spellings, you can click on an error category in the table to get more detailed information. Clicking on voc_r yields the page in Figure 3: Here you can see all occurrences of a vocalized <r> in the text with correct spellings in green and mistakes in red. Looking at the text itself gives insights which are hidden if you only look at the abstract statistics. While an error rate of 3/9 for voc_r does not look too alarming, seeing the occurrences in the text quickly reveals that the child only spelled the vocalized <r> correctly in the definite article <der>, which is probably memorized holistically, and made a mistake in all other occurrences, including two nouns (*<Mauer>*/<Mauer> ‘wall’, *<Wasser>*/<Wasser> ‘water’) and one conjunction (*<abba>*/<aber> ‘but’). From seeing this, you can quickly conclude that the phenomenon of r-vocalization has not yet been successfully acquired by the child. In general, it is important to not only analyze errors but also to see where a child dealt with a particular phenomenon correctly (see also [23]). While it is tedious to search a text manually for all occurrences of a phenomenon, the proposed application does exactly this for all of the error categories, which should make it easier for a teacher to evaluate a child’s spelling competence adequately.

Figure 4: Additional generated information concerning the misspelling *<rent>* for *<rennt>* ‘(he/she) runs’

Clicking on a specific mistake in the text opens a box with additional information about certain particularities of a misspelling. This is supposed to help giving relevant feedback about a particular spelling. In principle, all the annotations from our annotation scheme could be used here. For instance, it could be interesting to know which morpheme class was affected by the error: errors in frequent function words may point at more severe deficits in spelling competence. If, in contrast, a foreign word with a non-native structure was misspelled (as could be read-off from the feature foreign_target), these are not very indicative of the child’s spelling abilities per se. So far, the features phon_orig_ph and morph_const have been implemented into the application. Depending on the feature value, a sentence is generated which states whether the error has influenced the pronunciation of the word and whether the spelling can be derived from a related word form, respectively. Figure 4 gives an example for the misspelling *<rent>* for *<rennt>* ‘(he/she) runs’. The application shows here that the error does not influence the pronunciation of the word (so feedback should not refer to the child’s articulation here) but that the correct spelling can be derived from a related word form. For future versions, a desideratum is to also show the related word forms that the spelling can be derived from, in the case of *<rent>* it would for example be *<rennen>* or to pick words from a database which show a similar spelling pattern that can be used for training. What and how to display information here is worth discussing with experts on orthography instruction.

5. Conclusion and Outlook

This paper presented a proposal for the automatic spelling error analysis in freely written German texts. The aim of the paper was to show how previously developed techniques for analyzing spelling errors on various levels could be used to develop a user-friendly application. In the focus of the analyses are misspellings produced by German primary school children.

It could also play a role that in <der> the <er> is pronounced [er], i.e. you can hear [e], while in the misspelled words <cer> is pronounced [e].
Such an application can for example assist teachers in analyzing children’s spelling errors in more detail without a big manual effort and is also capable of giving hints for useful feedback about spellings. The prototype presented is supposed to give an idea of what our proposed application would be able to do. The final implementation, however, is highly dependent on the needs of the target users. So far, the application was presented as an aid for teachers or other language trainers who want to examine the main problem areas of a child’s spellings and be assisted in how to give feedback. Errors are marked in the text itself just like a teacher would do but at the same time one can see where the same phenomenon was handled correctly by the child, which would be tedious to annotate manually. For a smooth integration into the daily routine in schools, one obstacle is that children usually produce handwritten texts while the tool needs digital input. As was also noted by [7], it would be desirable to cooperate with the field of automatic handwriting recognition technology in this respect. On the other hand, with digitization being on the rise in all areas, it is not unlikely that in the future, children will produce a lot of typewritten texts as well. It is also conceivable to implement the automatic spelling error analysis as an application that can directly be used by children. In this case, the whole surface has to look differently of course and one should think about implementing the error analysis as a game.

Anyway, the further development of the tool would gain a lot from the expertise of professionals in orthography instruction. What has to be discussed, for instance, is what kinds of error categories to use (our categorization system is very fine-grained and can easily adapt to more specific needs), what information about an error to display and how to organize the whole application. A goal would then be to have the application tested and evaluated in real-world settings, i.e. by teachers or even children.

Furthermore, it would be interesting to see to what extent L2 learners of German could benefit from such an application as well. To investigate how their orthography errors fit into our annotation scheme is part of our future work.

6. Acknowledgements

This research is part of the project Literacy as the key to social participation: Psycholinguistic perspectives on orthography instruction and literacy acquisition funded by the Volkswagen Foundation as part of the research initiative “Key Issues for Research and Society”.

7. References


Transfer Learning based Non-native Acoustic Modeling for Pronunciation Error Detection

Richeng Duan, Tatsuya Kawahara, Masatake Dantsuji, Hiroaki Nanjo

1School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
2Academic Center for Computing and Media Studies, Kyoto University

Abstract

The scarcity of large-scale non-native corpora and human annotations are two fundamental challenges in the development of computer-assisted pronunciation training (CAPT) systems. We explored several transfer learning based methods to detect the pronunciation errors without using non-native training data. Effects were confirmed in the Mandarin Chinese pronunciation error detection of Japanese speakers. In this paper, we investigate the generality of the methods through application to an English speech data of Japanese speakers. We also evaluate on a non-native phone recognition experiment, which is necessary but challenging in advanced CAPT systems. Experimental results show that transfer learning based acoustic modeling methods can not only be ported to a new target language but also effective in a recognition task.

Index Terms: computer-assisted pronunciation training (CAPT), language independent pronunciation error detection, transfer learning

1. Introduction

In recent years, automatic speech recognition (ASR) has achieved great progress due to the emergence of Deep Neural network (DNN) and big data. However, pronunciation error detection in CAPT systems, which is often based on ASR technology, cannot benefit a lot from the ASR success. One of the most important reasons is lacking large-scale training resources with qualified annotations. As pronunciation of a target language is easily affected by learners’ native language, it is better to train the acoustic model with learners’ speech data. Unfortunately, it is much more challenging in collecting and labeling non-native speech than native speech because of the fewer populations and unnatural pronunciations [1]. To overcome this problem, we have explored several transfer learning based methods which aim at effective learning DNN acoustic models of non-native speakers without using such training data [2].

In this paper, we investigate the generality of transfer learning based methods in terms of different target languages and different evaluation settings. According to the Ethnologue [3], there are around 7000 living languages in the world. It is impossible to collect and annotate every non-native speech corpus considering different language pairs (target language and learners’ mother tongue). Most existing approaches to non-native acoustic modeling in CAPT are language dependent while our proposed approaches do not make any assumption about language pairs. As a result, we are interested in investigating the generality when applying to other language learning corpora. Specifically, a non-native English speech corpus of Japanese students is used in this work (different from non-native Mandarin Chinese in [2]). Experimental results show that transfer learning based modeling methods can generalize well on this new target language learning corpus. The majority of pronunciation error detection works are based on read speech, which means students have to repeat the learning material. Providing more choices for learners is desirable for advanced language learning [4]-[6]. However, allowing more freedom means the system needs to recognize the learners’ speech firstly, which is a challenging task in ASR because of a broad range of pronunciation variations in learners’ speech. Here, we also investigate the effectiveness of transfer learning based methods when evaluated in the non-native speech recognition.

The rest of this paper is organized as follows: In Section 2, pronunciation error detection based on DNN articulation models is described. Section 3 presents transfer learning based methods to enhance the learning of the DNN models. The performance of these modeling methods is firstly evaluated on native articulatory attributes recognition task in Section 4. Section 5 and 6 confirm the effectiveness in pronunciation error detection experiment and speech recognition of non-native speakers. Conclusions are in the final section.

2. Pronunciation error detection based on DNN articulation model

2.1. Pronunciation error detection

Pronunciation error detection on segmental level has been a core component in CAPT system. Most of the prior works focused on detecting phone substitution errors. Some researchers targeted a few specific problematic phones and explore the distinctive features and classifiers [7]-[9]. Others conducted the detection based on ASR technology, either incorporating the possible errors into the pronunciation lexicon or directly adding them into the decoding grammar [10]-[16]. The ASR-based method is more general than the specially designed ones since it can detect all the phones in a unified framework. A typical feedback of phone error detection approach is: “You made an r-l substitution error.” when a user pronounces the word “red” as “led”.

Instead of providing phone substitution feedbacks, giving the feedbacks directly related with articulation is more attractive [17]-[19]. Facing the same pronunciation error, learners could be instructed with “Try to retract your tongue and make the tip between the alveolar ridge and the hard palate”. This approach has been demonstrated helpful in many areas, such as speech comprehension improvement [20], speech therapy [21] and pronunciation perceptual training [22].
2.2. Context-dependent Articulation Modeling with DNN

Articulation means the movement of the tongue, lips, and other organs to make speech sounds. Generally, place of articulation and manner of articulation are used to describe the attributes of consonant sounds, while vowels are described with three-dimensional features: horizontal dimension (tongue backness), vertical dimension (tongue height), and lip shape (roundedness). We investigate articulatory models to recognize these articulatory attributes of foreign language learners.

Considering the co-articulation effect, context-dependent tri-attribute modeling is employed. Similar to context-dependent tri-phones used in ASR, labels for tri-manners and tri-places are generated by taking into account the labels of neighboring attributes. Since the mapping relation between articulatory attributes and phones is many-to-many, we prepare four kinds of transcriptions (manner, place-roundedness, place-backness and place-height) to represent all articulatory attributes. Articulatory attribute transcription is derived from the phone transcription. We exploit native data of target language (English in this work) to train the articulatory models (see Figure 1). These models can be directly used to detect pronunciation errors of language learners as a baseline.

![Figure 1: Context-dependent modeling of articulatory attributes.](image)

3. Enhancing articulatory attribute modeling with transfer learning

The idea of transfer learning, which should trace back to 20 years ago, has been successfully employed in broad research fields [23]-[26]. We employ transfer learning on the articulation modeling of non-native speech. Inter-language transfer learning, related-task transfer learning, and combination of these two methods are explored.

3.1. Related-task transfer learning on articulatory attribute modeling

Multi-task learning is an approach of transfer learning that learns a task together with other related tasks at the same time. In this study, phone classification task is served as the secondary task, which aims at helping the primary task learn better feature representation of attributes with the phonetic information.

3.2. Inter-language transfer learning on articulatory attribute modeling

In inter-language transfer learning method, two large native speech corpora of learners’ native language (Japanese) and a target language (English) are used to model the inter-language phenomenon since many articulatory attributes are shared between the two languages and we can easily get a large-scale corpus. Shared hidden layers in multi-lingual DNN (ML-DNN) allow for learning non-native articulatory features without using such data set.

3.3. Combining related-task and inter-language transfer learning for articulatory attribute modeling

We further investigate the combination of the related-task and inter-language transfer learning (see Figure 2). The related-task transfer learning learns the commonality through co-supervision of different tasks. The inter-language transfer learning aims at learning a better feature representation of non-native speech. As a result, their combination can have a synergetic effect.

![Figure 2: Enhancing the articulatory models through related-task and inter-language transfer learning.](image)

4. Native Attribute Recognition Experiment

4.1. Database

Two native speech corpora are used in this experiment. The native English corpus is Wall Street Journal (WSJ) databases [27], which is used to train the target articulatory models and validate different modeling methods. The Japanese corpus named JNAS [28] is also a commonly used database for Japanese large-vocabulary continuous speech recognition research. Sixty-four hours speech data from each corpus were selected after filtering noisy utterances. We conduct the evaluation on both Nov ‘92 and Nov ‘93 data sets of WSJ.

4.2. System Configuration

All different methods use the following DNN configuration: the acoustic feature consists of 40-dimensional filter bank outputs plus their first and second temporal derivatives. The input to the network is made by splicing 11 frames, 5 frames on each side of the current frame. The neural network has 7 hidden layers with 2048 nodes per layer. DNN training consists of unsupervised pre-training and supervised fine-tuning.
4.3. Experimental Results

The experimental results of different articulatory attributes are shown in Figure 3 to Figure 6. From these 4 figures, we observe the effects of all three transfer learning based methods. Compared with the conventional DNN, all the methods achieve lower recognition error rates. ML-DNN could benefit from more training data than conventional DNN though it comes from another language. MT-DNN is more effective than ML-DNN for native speech because it explicitly takes advantage of more direct information (phonetic labels). We highlight the effect of their combination (ML+MT DNN) as they are complementary to each other and can further reduce the error rate.

5. Non-native Pronunciation Error Detection

5.1. Experiment Setup

The evaluation data for pronunciation error detection is a corpus of English words spoken by Japanese students [29]. There are 7 speakers (2 male, 5 female) and each speaker uttered a same set of 850 English words. The database contains phonemic hand-labels, which were transcribed faithfully. We employ finite state decoding network for pronunciation error detection, which includes the canonical pronunciation and possible pronunciation errors.

5.2. Pronunciation Error Types

In this experiment, three pronunciation error types are focused which involve 5 specific vowels:

- Lip shape error: vowels with spread lips have problems of rounded sound.
- Tongue position error (horizontal): inappropriate tongue position with a little front or back.
- Tongue position error (vertical): inappropriate tongue position with a little high or low.

5.3. Evaluation Metrics

Two common used metrics of Detection Accuracy (DA) and F-score are used to evaluate the detection performance of different methods:

\[
DA = \frac{N_{TE} + N_{TC}}{N} \\
F-score = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{Precision} = \frac{N_{TE}}{N_{ID}} \\
\text{Recall} = \frac{N_{TE}}{N_{E}}
\]

\(N_{TE}\) is the number of pronunciation errors correctly detected by the system. \(N_{TC}\) is the number of correct pronunciation detected as correct one by the system. \(N\) is the total number of test samples. \(N_{ID}\) is the number of all detected pronunciation errors. \(N_{E}\) is the total number of pronunciation errors in the test set.

5.4. Experimental Results

Figure 7 compares the overall detection performance of different methods: conventional DNN, MT-DNN, ML-DNN, the combined ML+MT DNN. We can see that both MT-DNN and ML-DNN perform better than the conventional DNN. While MT-DNN is consistently better than ML-DNN in the previous native attribute classification experiment, ML-DNN is generally more effective for modeling non-native speech. This is because MT-DNN is trained with English data only while we add Japanese characteristics by using both English and Japanese data sets. Detailed detection results of individual error types are shown from Figure 8 to Figure 10. Among these three errors, the system detects the lip shape error best, while the tongue position error types are less accurate. This

Figure 3: Manner attribute classification.

Figure 4: Place-roundedness attribute recognition.

Figure 5: Place-backness attribute recognition.

Figure 6: Place-height attribute recognition.
tendency is similar to what we observed in the native speech attribute recognition experiment. However, the absolute performance of tongue backness error detection is rather low compared with the high accuracy of native place-backness attributes (Figure 5). This is partly due to the significant phonological differences between Japanese and English, especially the vowel system. In terms of vowel inventory, there are only five vowels in Japanese language while sixteen vowels (including the schwa sound) are in English. The considerably more vowels in English not only brings a big challenge for Japanese students learning English vowels but also for the annotators when labeling the non-native speech.

We also add the result of Mandarin Chinese pronunciation error detection [2] in Figure 11. We can see that transfer learning based methods can generalize well in different target languages. Compared with Mandarin Chinese pronunciation error detection, however, detection performance of English pronunciation error is much lower. This is because there are only vowel errors considered in this preliminary study, and the acoustic difference among English vowels is very subtler as discussed above.

Figure 7: Overall system performance for non-native English pronunciation error detection.

Figure 8: DA & F-score for lip shape error.

Figure 9: DA & F-score for tongue backness error.

Figure 10: DA & F-score for tongue height error.

Figure 11: Overall system performance for non-native Mandarin Chinese pronunciation error detection.

Figure 12: Phone error rate (PER) of non-native English speech recognition.

6. Non-native Speech Recognition

Above-mentioned pronunciation error detection is text-dependent, which means the system provides a set of predefined scripts for the students to read. In this case, the system knows what the target pronunciations should be, and the scripts can be used for constructing a finite state decoding network for pronunciation error detection. However, as the students improve, especially for those advanced learners, it would be better to let them speak freely and create their own sentences rather than reading a given text. The text-independent system should provide two functions. One is recognizing the learner’s speech, the other is detecting the pronunciation errors based on recognized text.

We conduct a free phone recognition experiment, in which the system recognizes the non-native speech without using any constraint. This setting is the most general but hardest condition. Figure 12 demonstrates the phone error rate (PER) of different methods. We can see that transfer learning based
methods perform better than conventional DNN even when they are ported to speech recognition of non-native speech. We also see that there is a large room to improve the performance as the absolute recognition error rate is high. A promising solution is that adding a lexicon with reasonable size to control the search space.

7. Conclusions
In this paper, we investigate the generality of transfer learning based modeling methods by applying to different target languages and different tasks. We have performed pronunciation error detection experiments on two different language learning corpora (Mandarin Chinese and English). Experimental results show the effectiveness of transfer learning based modeling methods on both corpora. We then investigate the usability in speech recognition of non-native speech. The experimental result shows that transfer learning based methods are still effective even with the free phone recognition.

In theory, the proposed approach can be applied to any language pairs as long as there is a native standard corpus. It opens new possibilities in language-independent pronunciation error detection. In future, we will apply these methods to more language learning corpora. Text-independent pronunciation error detection will be our next direction.

8. Acknowledgements
The author would like to acknowledge the financial support from Chinese Scholarship Council (CSC).

9. References
[10] Y. Tsubota, T. Kawahara, and M. Dantsuji, “Recognition and verification of English by Japanese students for computer-
Effects of Educational Context on Learners’ Ratings of a Synthetic Voice

Neasa Ní Chiaráin, Ailbhe Ní Chasaide

The Phonetics and Speech Lab., School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland

neasa.nichiarain@tcd.ie, anichsid@tcd.ie

Abstract

Studies suggest that the context in which synthetic voices are evaluated has a significant bearing on the evaluators’ assessments of the voices. The present study is based on learners’ evaluations of an Irish synthetic voice which is used in a Computer-Assisted Language Learning (CALL) context.

Previous studies have allowed a number of extraneous variables, such as different voices being used or pitch or speech rate manipulations, to enter the study. The present study involves a single synthetic Irish HTS voice being evaluated in three different contexts, one of which is highly interactive, a second less so and a third which does not require the evaluator to perform any specific task. Results show that the voice was rated more favourably when presented in a highly interactive context. This finding has significant implications for the way we interpret the results of evaluations of synthetic voices.

Index Terms: text-to-speech synthesis, evaluation, computer-assisted language learning, Irish (Gaelic)

1. Introduction

The question of how to evaluate synthetic voices has been a subject of discussion for some time now [1]. Criteria such as intelligibility, naturalness and attractiveness are frequently applied though it is not always clear that those who are assessing the voices have a common understanding of what exactly is intended by each of the terms. Handle [2] suggested that one’s judgment of the voices was dependent on the role which those voices had in the context in which they were being judged. Ni Chiaráin [3] suggested that one’s evaluation of the voice was dependent on the degree to which one engaged with the situation in which the voice was being used. The latter study involved three distinct environments which were used as CALL platforms, or serious games, for the teaching of the Irish language to sixteen year-old students. One involved navigating a virtual world in order to complete a task, a second had a rich visual environment with a low level of interactivity, while the third involved a dialogue partner, or chatbot, which had relatively simple graphics but was highly interactive. The synthetic voices in question were HTS voices which were developed as part of the ABAIR initiative at the Phonetics and Speech Lab., CLCS, TCD (www.abair.ie).

It emerged that the platform which involved the highest level of interactivity was scored the highest by learners (N=251). Although the same voices were used for each of the platforms, it was necessary to carry out pitch and speed manipulations on them so that different characters in the various platforms could be distinguished from each other.

The present study examines the relationship between learners’ evaluations of a synthetic Irish voice and the degree of interactivity and engagement which the learners have with the CALL platform in which the voice is being used. It uses a single voice which has not been manipulated thereby reducing the chance of extraneous variables such as pitch or speech rate changes influencing the outcome. The single voice is evaluated in terms of intelligibility, quality and attractiveness in three diverse CALL contexts:

1. a highly interactive context where the learner interacts directly with a chatbot
2. a task-based activity where learners are asked to transcribe sentences as they hear them
3. a non-task-based activity where learners merely listen and follow the text on the screen

Participants are asked to evaluate the voice after each context. The hypothesis is that ratings for the interactive context (1) will be the highest, followed by the task-based context (2), and that the least highly rated will be the non-task-based activity (3), where the activity is such that learners are least likely to be engaged.

This study has significant implications for the way we interpret evaluations of synthetic speech and may account for inconsistent evaluations of the same voice.

2. Motivation and Rationale for this Study

There is a threefold motivation for the present study. Firstly, it wishes to extend the debate on the evaluation of synthetic speech with particular reference to the connectivity between evaluators’ perception of the synthetic voices and the context in which the voices are being experienced.

The second motivation for conducting the study relates to the usefulness of synthetic speech as an aid to language teaching/learning in a CALL context. To date, synthetic speech has been used in a very restricted way for language teaching. Continuous improvements in the quality of the voices could be expected to add greatly to their usefulness as language teaching aids. In the context of interactive CALL platforms, where natural language processing/artificial intelligence is used to create chatbot-type platforms where the computer interacts in a non-linear fashion with the input of human interlocutors, then the synthetic speech is the only type of speech which can be used. The computer can generate an infinite number of responses which could clearly not be pre-recorded. The present study is part of a larger study that aims to evaluate the usefulness of synthetic speech for such a context.

Finally, the study is part of the overall ABAIR initiative, which has developed synthetic voices for the Irish dialects. The Irish language is designated as one of the endangered...
languages by UNESCO [4]. There are relatively few native speakers of Irish available as speaker models for learners and the availability of interactive software programs is similarly restricted. The ABAIR initiative aims to develop interactive Irish language teaching/learning platforms into which the synthetic speech is integrated. This study important to provide an evaluation of the suitability of the synthetic voices for such platforms at this stage in their development.

3. Literature Review

To date, text-to-speech (TTS) synthesis has not been used very frequently as an integral part of Computer-Assisted Language Learning (CALL) systems [5]–[7]. In earlier years this was largely because of the relatively poor quality of the synthetic voices and their inability to closely imitate the human voice [8]. More recent advances in the development of TTS have now reached the point where most systems are almost 100% intelligible [1]. Indeed, in one case, language learners found it difficult to distinguish between natural and synthetic speech [1], [9]. Despite the advances and the vastly improved quality of many of today’s synthetic voices the observations of Keller and Zellner-Keller [10] that they are insufficiently expressive and lack any form of human emotion, are still valid to-day.

The greatest weakness being reported in TTS being used for CALL purposes is that at the suprasegmental level of production where it lacks appropriate prosody with the result that listeners can experience difficulty in focusing on the output leading to lower levels of comprehension than would be the case for a natural voice [11].

To date, synthetic speech in CALL platforms have tended to be used for the teaching of individual skills such as listening comprehension, reading or pronunciation skills [6], [12], [13]. Modern approaches to language teaching favour the learner being involved in a task in a communicatively interactive way, using the target language in order to achieve some pre-defined goal. Some language learning ‘games’ are emerging which allow the learner to interact with a virtual interactive dialogue partner, or chatbot, in a non-linear fashion so that so that the output responds to the input of the learner in an individualized fashion. This type of platform is frequently referred to as a ‘serious game’. Until recently there have been few instances of the integration of TTS into serious CALL games [14]–[17]. The use of TTS synthetic voices becomes essential in CALL platforms which are genuinely interactive in a non-linear fashion. In this situation pre-recorded voices cannot be used as there may be a near infinite number of possible responses. Since we are at the very initial stages of integrating synthetic voices into interactive CALL platforms with dialogue games we must now revisit the criteria by which TTS is evaluated in order to establish its fitness-for-purpose in this evolving context [18].

King [1] suggests that subjective listening tests are the only sure way to evaluate synthetic speech. Diagnostic or comparative laboratory tests may not equate to the perceptions of the human brain. Furthermore, when the objective is a learning context, it is vital that the adequacy of the voice be assessed by learners, where additional factors such as cognitive load are realistically mimicked. When one introduces subjectivity to any evaluation then context and evaluators’ characteristics come into play. The effects of context in visual perception have been well documented and there are numerous examples of how the perception of objects changes when its context changes without any physical changes in the object itself [19]. This same phenomenon, though less widely reported, occurs in the case of auditory perception. Situational context has been described in terms of the psychological, social, physical, physiological and linguistic constraints which come to bear on the individual’s perception of auditory cues [20] [21] [22] [23] [24]. The phenomenon known as the ‘McGurk Effect’, for example, illustrates the multimodal nature of speech perception since it takes auditory and visual cues in processing an auditory stimulus [25]. Similarly, Munhall et al. noted the relationship between physical head movements and linguistic meaning amongst Japanese speakers [26].

Context and perception are inextricably linked and consequently evaluations of phenomena such as synthetic speech remain somewhat elusive [1]. Evaluators’ judgments vary with context, as in the case in the perception of natural language, and the literature draws attention to a variety of factors found to influence perceptions of synthetic speech. It has also been suggested that a learner’s evaluation may be quite different to that of a native speaker. Kang et al. [7] found that learners’ general listening comprehension abilities influenced their ratings of synthetic voices. Stern et al. [27] suggested that synthetic voices are judged more positively when they are presented as emanating from a computer rather than being human voices. Cryer and Home suggested that listeners’ predisposition to synthetic speech based on their previous experiences with it may influence their attitude or prejudices towards it [28]. One may reasonably claim that the context in which synthetic speech is being judged has an influence on the listeners’ evaluations of it. Ni Chiaráin suggested that in the case of a CALL platform, one’s evaluation of synthetic speech may be highly influenced by one’s perceptions of other facets of the platform such as graphics, plot, attractiveness, playfulness, etc. [3].

4. Research Design & Experimental Setup

The present study examines the degree to which learners’ perceptions of one Irish TTS synthetic voice (see Section 4.2) vary within three diverse CALL contexts (see Section 4.3).

4.1. Experimental setup

Three schools were selected for participation in this study. These are Irish-medium schools and were selected to ensure participants would have a high level of Irish. This was done to ensure the scores for intelligibility weren’t affected by lack of knowledge of the language, although it is important to mention that the students may not necessarily be familiar with the dialect spoken by the HTS system.

It was decided to use a 6-point Likert scale to elicit the subjective intelligibility, quality and attractiveness ratings from the participants. This was intended to avoid the tendency of respondents to opt for the mid-point in a 5-point scale. Selections to the level of one decimal point were made possible. It was felt that allowing ratings to a level of one decimal point would allow respondents to give a more fine-grained response rather than relying on whole numbers. The scale was presented in the form of a sliding bar where the default was set at the midpoint of 2.5 and participants were free to drag the slider to the left towards 0 or to the right towards 5. Meaningful statements labels the axes (e.g. Intelligibility: 0 “impossible to understand” to 5 “very easy to understand”). This configuration allows for the calculation of
a mean score for scale items which can then be used to produce the parametric mean and standard deviation.

It was decided to host the evaluations online so that they could be delivered to students in their own school environment. Each student worked individually and was equipped with a set of headphones to listen to the speech output.

On the evaluation day, students were given a broad general introduction to the area of speech and language technology by the researcher as well as a short introduction to the Irish ABAIR initiative. This entailed explaining that the project aimed to develop multi-dialect synthetic voices for Irish. It was not specified to the students how many voices were involved in the initiative, rather that all three of the main dialects of Irish were represented. They were told that they were evaluating the synthetic speech and that their opinions were valued because of their advanced level of Irish. They were given a half an hour to work individually at their computers and were presented with three diverse CALL contexts in random order. Each context used an embedded synthetic voice. A pre-evaluation questionnaire was presented to students to elicit some basic background information, such as gender and age. Empty text boxes inviting feedback were included on each page of the evaluation.

4.2. One HTS voice

A synthetic voice was built with HTS [29] using a corpus recorded from a middle-aged male speaker of the Connacht dialect of Irish. This voice was built based on c. 3,000 utterances at a 32 kHz sample rate. An example of the synthetic voice can be heard by visiting www.abair.ie, selecting ‘Conamara HTS’ and synthesizing some Irish text.

4.3. Three CALL contexts

The three diverse CALL contexts were presented to students in random order to limit as much as possible fatigue impacting on the scoring of any individual context.

4.3.1. A highly interactive context

Students were given the opportunity to have a playful, non-directed interaction with a chatbot call Taidhgín which was developed as a prototype spoken system by Ní Chiaráin [3]. Taidhgín is a chatbot in the form of a monkey who appears as though he is engaging in conversation by typing prompts or responses into a laptop (see Figure 1). Learners make their input by typing into a text box. Irish does not yet have a speech recognition system and hence the only input possible for now is in text form. The chatbot is built on the basis of pattern matching and responds to input from a database of preprogrammed content, which is appropriate to the language level of a 16–17 year old learner in the Irish school system. For more details on Taidhgín see Ní Chiaráin and Ni Chasaide [18]. Students were not given a specific time limit for interacting with the chatbot but most spent 10-15 minutes in the engagement.

Figure 1: Context 1–Taidhgín, the interactive dialogue partner.

4.3.2. A task-based activity

In this activity the evaluators were asked to transcribe 20 sentences as they heard them. The sentences were composed especially for this evaluation – the language used was at an appropriate level of difficulty for the target group, using vocabulary which belongs in the domain of schooling and is commonplace for them. Once synthesized, there were no modifications made to the output. All sentences were unseen and students were asked to orthographically transcribe them. Each sentence was decontextualized and presented in random order with no common semantic connections between them. Students were free to listen repeatedly, if necessary. The average sentence length was 13.45 words. The time spent was comparable to the interactive context (c. 10-15 minutes).

Figure 2: Context 2–presentation of the twenty utterances for individual transcription & rating.

4.3.3. A non-task-based activity

Students were asked to listen to a continuous piece of prose enunciated by the synthetic voice. Unlike the sentences in the previous context, these formed a coherent whole and evaluators could also read the script on screen. It was comprised of 421 words divided into 3 separate sections and lasted 2 minutes 23 seconds. As in the previous two contexts, the content and difficulty level were considered appropriate and there was no restriction on the number of repeats they were allowed to make. Typically, this task took about 5 minutes.
5. Results

Participants (N=52) were from three Irish-medium schools in an urban setting. The group was predominantly female (N=38) and all were 16-17 year old pupils in their fourth and fifth years of post-primary education. While they have had their education through the medium of Irish and may be described as having a high level of language ability, nevertheless they are based in neighbourhoods where English is the predominant language and many may not be familiar with the Connacht dialect used for this evaluation.

Subjective ratings on a 6-point Likert scale were given by each participant. The arithmetic mean was calculated and ratings for the intelligibility, quality and attractiveness of each of the three CALL contexts are shown in the table below.

5.1. Intelligibility

Table 1 shows clearly a high level rating for intelligibility in all three contexts. The highest intelligibility rating is given to the voice when it was used for the highly interactive context. In the non-interactive contexts, the task-based activity yielded a slightly higher intelligibility rating than did the non-task based one. These findings support the initial hypothesis that the level of interaction and engagement with the CALL activity would influence judgments concerning the synthetic voice.

One factor should be mention which might conceivably have influenced the present results. The initial conversational turns with the chatbot entailed very short greeting sentences. It is possible that this served as a warm up so that learners relaxed and felt they got to know the character and the game and that this contributed towards a sense of the synthesis being more intelligible. While it is unlikely that this would account for the appreciable and consistent difference that emerged, this is a factor that future work should control for.

5.1.1. Transcription Results and Intelligibility

Further indicators of an overall high level of intelligibility of this synthetic voice (regardless of context) is provided by the transcriptions made by students in the task-based context. The vast majority of utterances were successfully transcribed. The few errors which did arise fell into two categories: (1) errors associated with students’ own fossilized language level where they used a grammatically incorrect version of a phrase even though they had been presented with the grammatically correct utterance (see example in Table 2), and (2) Mondegreens, where the students erroneously transcribed phrases bearing a close phonetic resemblance to the prompt (see example in Table 3).

Table 2: grammatical error in transcriptions.

<table>
<thead>
<tr>
<th>Error</th>
<th>Utterance</th>
<th>Translation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>An bhfuil tú ciinte nach bhfaca tú duine ar bith ar an mbealach abhaile?</td>
<td>Are you sure you didn’t see anyone on the way home?</td>
<td>An bhfuil tú ciinte nach bhfaca tú duine ar bith ar an mbealach abhaile? (Did you see anyone on the way home?)</td>
<td></td>
</tr>
<tr>
<td>An bhfuil tú ciinte nach əchonaic ə dúine ar bith ar an mbealach abhaile?</td>
<td>Are you sure you didn’t see anyone on the way home?</td>
<td>An bhfuil tú ciinte nach əchonaic ə dúine ar bith ar an mbealach abhaile? (Did you see anyone on the way home?)</td>
<td></td>
</tr>
</tbody>
</table>

Note: ‘feic’ (‘to see’) is an irregular verb. Irish has no translation of yes/no, one answers with a positive/negative repetition of the verb. In this case the learner is using the wrong form of the verb, a common error made by learners.

Table 3: example of Mondegreen in transcriptions.

<table>
<thead>
<tr>
<th>Error</th>
<th>Utterance</th>
<th>IPA</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>An bhfuil tú ciinte nach bhfaca tú duine ar bith ar an mbealach abhaile?</td>
<td>An bhfuil tú ciinte nach bhfaca tú duine ar bith ar an mbealach abhaile?</td>
<td>An bhfuil tú ciinte nach bhfaca tú duine ar bith ar an mbealach abhaile?</td>
<td></td>
</tr>
<tr>
<td>An bhfuil tú ciinte nach əchonaic ə dúine ar bith ar an mbealach abhaile?</td>
<td>An bhfuil tú ciinte nach əchonaic ə dúine ar bith ar an mbealach abhaile?</td>
<td>An bhfuil tú ciinte nach əchonaic ə dúine ar bith ar an mbealach abhaile?</td>
<td></td>
</tr>
</tbody>
</table>

Note: ‘feic’ (‘to see’) is an irregular verb. Irish has no translation of yes/no, one answers with a positive/negative repetition of the verb. In this case the learner is using the wrong form of the verb, a common error made by learners.

5.2. Quality

The ratings for the quality of the synthetic voice were fairly similar for all three contexts, but nevertheless the ranking of the ratings were in line with the hypothesis. Overall, these ratings are rather high and are reassuring for potential CALL users. Note that the ratings for intelligibility and quality are identical in the non-interactive contexts. For the interactive context, there is a considerable difference in the ratings for intelligibility and quality.

5.3. Attractiveness

While the ratings for attractiveness are lower than for either intelligibility or quality, the scores (close to the midpoint of
2.5) point to a fairly neutral judgment, as being neither particularly attractive/unattractive.

6. Discussion

The hypothesis is supported by these findings: one HTS voice, several CALL contexts, evoked different reactions. This finding highlights the need to reconsider how and why we evaluate synthetic speech and provides further evidence that context influences our perception of auditory cues.

Of the three measures used in this study, intelligibility is undoubtedly the most critical. It is probably also the most objective since subjective evaluations can be supported by more objective indicators, such as the results of transcription tasks, as was done here. It is interesting to note that while some students got a perfect or near-perfect performance score in the transcription task, they nevertheless evaluated the intelligibility and quality of the speech as falling short of perfect. This may be interpreted as indicating that while the speech is intelligible, nevertheless it may take greater mental effort to understand the synthetic voices than would be the case for natural voices. This greater effort gives rise to what is termed “cognitive loading” [30] and can be expected to depress subjective evaluation scores. Ni Chiaráin & Ni Chasáide used a formula to produce an intelligibility and clarity index which averaged the subjective and objective scores and argued that this produced a far more realistic representation of the intelligibility of synthetic speech than either measure on its own [31].

Intelligibility is the single most important factor when one is using synthetic speech for interactive language learning games in that it allows the learner to interact with the chatbot in a coherent manner. While intelligibility cannot be sacrificed, a relatively lower rating for quality may not be as critical as it has been shown that learners are more tolerant of the voice when eminates from a machine [28].

The judgments on attractiveness are the lowest of the three measures. One potentially important factor here is that the voice was of a middle-aged male, of rural background and the evaluators were predominantly female adolescents from an urban background. Clearly, the voice they were rating is unlikely to conform to their image of ‘cool’ or ‘attractive’. (We speculate that very different and more positive responses might be obtained from older evaluators from a rural background). Attractiveness is highly subjective and is likely to vary in a way that reflects the evaluation group. Importantly, in the context of game-characters it is only one of many potential attributes one would wish for. For example, an authoritative, formal, friendly or playful voice may be more appropriate depending on the role which is being assigned to the game character to make it more credible. Given that listeners make subjective judgments about personality based on the quality of the voice [32], [33], the future for CALL may be to match the synthetic voice(s) to the character-role(s). As we move towards more sophisticated interactive dialogue-based CALL scenarios, our evaluations may need to be broadened to explore the perceived age and ‘personality’ of the character conjured by the synthetic voices at our disposal. Looking further into the future, we would also aspire to voices where the quality can credibly be modulated in ways that are appropriate to the context.

On the question of evaluation metrics, it is worth considering whether the widely-used ‘intelligibility’, ‘quality’ and ‘attractiveness’ are the right measures of synthesis quality. Other metrics can also be very important depending on the circumstances. In the Irish context, for example, ‘nativeness’, i.e. the native-like quality of the voice is a priority for many, given that the interactive CALL applications aim to remedy the lack of availability of native-speaker models of the language. But beyond the realm of the endangered language, nativeness/authenticity can also be important in assistive technologies, where the synthetic voice becomes the user’s voice.

7. Conclusions

The present results support the initial hypothesis that the context in which the voice is used will affect how it is perceived by the user. It seems clear that the highly interactive playful chatbot was more engaging and that this in itself enhanced listeners’ attitudes towards the synthetic voice.

Results also indicate that the synthetic Irish voice that has been developed as part of the ABAIR initiative is fit-for-purpose for interactive CALL activities and has a very high level of intelligibility among 16-17 year old learners of Irish. In the absence of a wide variety of synthetic voices, as in the case of minority languages, one may still proceed with a voice that is highly intelligible so that learners can engage with the target language in an interactive, task-based manner [34].

The voice deployed here is but one of three current HTS ABAIR voices, spanning the three major dialects and offering choice in terms of gender and age. As further voices are developed, a priority is to ensure that they measure up to the contexts in which they are to be used and for this early consideration of appropriate evaluation criteria will be critical.

8. Acknowledgements

Funding from the Irish Government’s Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs (ABAIR Initiative) and An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (Cabrair project) is gratefully acknowledged. Bualochas mór leis na múinteoirí agus agus Gaelscoileachtaí a ghlac páirt sa thaidhde seo.

9. References


Linguistic Factors Affecting Evaluation of L2 Korean Speech Proficiency

Seung Hee Yang¹, Minhwa Chung¹, ²

¹Interdisciplinary Program in Cognitive Sciences, Seoul National University, Republic of Korea
²Department of Linguistics, Seoul National University, Republic of Korea

Abstract

Much research attention has been directed to identify how native speakers perceive non-native speakers’ oral proficiency. To investigate the generalizability of previous findings, this study examined segmental, phonological, accentual, and temporal correlates of native speakers’ evaluation of L2 Korean proficiency produced by learners with various levels and nationalities. Our experiment results show that proficiency ratings by native speakers significantly correlate not only with rate of speech, but also with the segmental accuracies. The influence of segmental errors has the highest correlation with the proficiency of L2 Korean speech. We further verified the validity of this finding across different L1 backgrounds. Although phonological accuracy was expected to be highly correlated with the proficiency score, it was the least influential measure. Another new finding in this study is that the role of pitch and accent has been underestimated so far in the non-native Korean speech perception studies. This work will serve as the groundwork for the development of automatic assessment module in Korean CAPT system.

Index Terms: CAPT (Computer-Assisted Pronunciation Teaching), Second Language Learning, Oral Proficiency Assessment, Non-native Korean Speech

1. Introduction

In second language (L2) acquisition, a growing number of researchers have emphasized the importance of assessing L2 speech proficiency based on judgments of comprehensibility, accentual, and intelligibility [1][2][3][4][5][6][7]. They have studied what kinds of linguistic properties, such as phonetic accuracy, fluency, and grammar errors, are relatively crucial for native speakers’ assessment under various task conditions.

This is a significant topic in language learning, since the results from these studies direct how L2 learners can achieve successful oral communication in the language. Indeed, the factors that may interfere with communication, and the degree to which they determine perceptual significance need to be identified for second language instructors, curriculum designers, and language learning software developers, since these standards will be manifested in the language teaching methodology. For instance, accentuatedness may not always spill over to lack of intelligibility. In a similar way, mispronunciation at the phonetic level may not necessarily interfere with communication. Some types of error influence the overall proficiency score more than others, and it is pedagogically important to obtain a better understanding of the correlation among the factors affecting listeners’ perceptions.

This study has been conducted within the framework of CAPT (Computer-Assisted Pronunciation Teaching) system development for Korean. A CAPT system automatically assesses learners’ speech and provides corrective feedback. In order to implement an automatic assessment module, it is necessary to examine how the task is carried out by human raters, and what criteria of judgment are used. For example, whether or not phonological and phonetic errors should be scored with equal weights in Korean, the context in which they cause miscommunication, and the degree of listeners’ perceptual sensitivity need to be investigated.

It is noteworthy that previous literature on speech assessment considered L2-dependent factors in dealing with judgments of L2 English, German, Spanish, Japanese, Dutch and Chinese [1][2][3][4][5][6][7]. For example, in case of L2 English, due to the empirical evidence in favor of prosody as a factor of intelligibility, it was included as a variable in the experiment [2]. Studies on L2 Chinese investigated tonal realization patterns [6], whereas pitch accent was studied for L2 Japanese [4]. It remains open to question whether and to what degree such findings can be generalized to other linguistic contexts. In this study, L2-dependent factors for Korean were included as variables in order to identify the linguistic properties of Korean speech that are perceived as crucial for native evaluators.

Section 2 summarizes previous findings on non-native speech assessment and Section 3 proposes an improved experiment design. Section 4 describes the results of our experiment, followed by discussion and conclusion in Sections 5 and 6.

2. Related Studies

This section summarizes related studies on non-native speech assessment. The studies on L2 Korean will be examined in more detail in Section 2.2., before proposing an improved experiment for the current research in Section 3.

2.1. Non-native Speech Assessment

L2 education researchers have emphasized the importance of analyzing L2 speech with comprehensibility, intelligibility, and accentuatedness. According to Derwing and Munro’s seminal work on accentuatedness and native speakers’ comprehensibility [1][2], utterances that are perceived as heavily accented can be highly comprehensible. The finding showed that the degree to which learners approximate the native speaker norm does not necessarily measure how easily L2 utterances are understood. More empirical studies have examined phonological, temporal, lexical or grammatical correlates of L2 German [3], Japanese [4], Dutch [5], Chinese [6] and Spanish [7] comprehensibility.
2.2. Non-native Speech Assessment for L2 Korean

Table 1 shows the evaluation criteria employed by previous experiments that deal with L2 Korean. They assessed whether or not meaningful correlation can be observed between a fixed number of factors and proficiency scores [8][9][10][11][12]. The proficiency criterion refers to a holistic measure according to the rater’s impression of proficiency in the utterance. Holistic measure of proficiency is distinguished from analytic measures since the raters rely on comprehensive impression across the entire utterance rather than paying attention to particular linguistic properties, such as fluency, phonology, or phonetics [13].

The filled circles in Table 1 indicate the factors that are highly correlated with the overall proficiency score according to each experimental results. There is no filled circle in the columns corresponding to [14] and [15] because these studies did not measure correlations between variables, but were interested in the change within a variable across time. Results in [14] concluded that fluency score improves for 6 months and starts degrading, while [15] found that all learners show different improvement patterns over time.

All studies that measured fluency as the evaluation criteria agree that it is a useful measure which highly correlates with native listeners’ perception of speech proficiency [8][9][10]. The correlation was shown to be stronger than segmental accuracy [8][11]. However, segmental accuracy, including all substitutions, deletions, and insertions, is still important according to [9], while number of juncture insertion is an important consideration in [10].

The causes of these inconsistencies can be partly explained by the nature, design, and purpose of different experimental set-ups, which is summarized in Table 2. For example, L1 specificities may have introduced some disagreements, and it may be the case that segmental accuracy is more crucial for speakers with Japanese background, while suprasegmental features are more important for speakers with Chinese background. Moreover, whether the evaluation sample was read speech or spontaneous speech would play a big role in the assessment, as it introduces other factors such as orthographical influence and the learners’ knowledge of the words. For evaluating read speech prompts, they used proficiency score as a criterion [8][9][10], while the studies using spontaneous speech included comprehensibility or complexity measures [11][12][14][15]. The disagreements raise further research questions and call for an improved experiment that can clarify which areas of language (fluency, juncture, segmental, phonological accuracy) influence native speakers’ judgments of L2 Korean.

Moreover, some variables have been left out or have been insufficiently considered, which can be a limitation of the experiments. For instance, most of them did not measure the effect of pitch and stress errors, which have been influential factors in L2 evaluation [1][2][4]. Also, the coverage of phonological rules is not comprehensive. One of the characteristics of Korean is its usage of phonological rules, i.e., pronunciation at the surface level is changed by certain conditions in the underlying representation. Several studies have reported that learners of Korean are pronouncing the segments according to their underlying representation, and phonological rules are not realized [16][17][18][19]. Therefore, the extent to which phonological accuracy affects the assessment of L2 Korean speech needs to be thoroughly investigated. However, only a portion of the phonological phenomena has been covered in [8], and it is necessary to design an experiment that is comprehensive in scope. The next section will discuss in more detail what the missing phenomena were, and how we propose to improve the experiment.

Table 1: Evaluation criteria used in previous studies assessing non-native Korean speech (= used as a variable, ●= used as a variable and found to be an important feature).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch</td>
<td>●</td>
<td>●</td>
<td></td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>Juncture</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td>○</td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>Segmental accuracy</td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>Phonological accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>○</td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>Hesitation</td>
<td></td>
<td></td>
<td>○</td>
<td></td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>Comprehensibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Comparison of experimental set-ups in previous studies assessing non-native Korean speech (NS = Native Speaker).

<table>
<thead>
<tr>
<th>Study</th>
<th>Task</th>
<th>No. of Speakers (L1)</th>
<th>No. of Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>[8]</td>
<td>Read 10 sentences</td>
<td>33 (Chinese)</td>
<td>3 (NS)</td>
</tr>
<tr>
<td>[9]</td>
<td>Read 210 sentences</td>
<td>24 (Japanese)</td>
<td>4 (NS)</td>
</tr>
<tr>
<td>[10]</td>
<td>Read 25 sentences</td>
<td>24 (Chinese)</td>
<td>10 (NS)</td>
</tr>
<tr>
<td>[11]</td>
<td>Interview after watching Mr. Bean movie</td>
<td>130 (Chinese, French, Japanese, English)</td>
<td>6 (NS)</td>
</tr>
<tr>
<td>[12]</td>
<td>Open question answering, picture story telling</td>
<td>37 (Thai)</td>
<td>5 (NS)</td>
</tr>
<tr>
<td>[14]</td>
<td>Interview after watching a movie</td>
<td>20 (Chinese)</td>
<td>3 (NS)</td>
</tr>
<tr>
<td>[15]</td>
<td>Talk about a given topic</td>
<td>3 (English, Chinese, Russian)</td>
<td>1 (NS)</td>
</tr>
</tbody>
</table>

3. Experiment Design

This section describes our methodology and what improvements have been made compared to the previous studies. It elaborates on how we composed the reading prompts, variables, evaluation method, speakers, and evaluators.

3.1. Reading prompts

50 speakers were given 100 sentences to read. The sentences used everyday vocabulary from L2 Korean text books, such as “How many times have you been to Korea?” and “I usually eat dinner when I go home.” Since spontaneous speech is closer to the real communication phenomenon, using read speech corpus can have limitations. However, canonical pronunciation is predefined in read speech, which can be an
advantage for discovering error patterns, and also for conducting a research with the beginner level speakers, whose canonical form of the utterance are often impossible to identify.

Moreover, using read speech prompt enables a comprehensive analysis of phonological accuracy. For this study, we used 50 sentences containing carefully selected phonological rules. As shown in Table 3, there are 264 instances of phonological rules occurring in the prompt. Five common phonological rules for both cross and within-morphemes are included. For example, tensification rule in the word “worry” [gok̚z̚a] occurs within morpheme, whereas the aspiration rule in the word “would like” [gok̚es’uminda] occurs across morpheme. The frequencies are balanced to cover diverse linguistic phenomena. Regarding sentence types, the 50 sentences consist of 42 statements and 8 questions.

Table 3: Distribution of phonological rules occurring in the 50 sentences used in this experiment.

<table>
<thead>
<tr>
<th>Phonological Rule Type</th>
<th>Cross-Morpheme (freq.)</th>
<th>Within-Morpheme (freq.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lenition</td>
<td>82</td>
<td>5</td>
</tr>
<tr>
<td>Tensification</td>
<td>33</td>
<td>13</td>
</tr>
<tr>
<td>Nasalization</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>Aspiration</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td>Palatalization</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>158</td>
<td>106</td>
</tr>
</tbody>
</table>

3.2. Selection of variables

The purpose of the current investigation is to examine the generalizability of previous findings in [8][9][10][11][12] and resolve their disagreements. We also examine whether and to what degree the correct realization of phonological rules affect L2 Korean perception. In order to answer these research questions, the following five variables have been defined: segmental accuracy, phonological accuracy, pitch and accent, fluency, and holistic impression of proficiency. Upon listening to each sample, the raters used 1-5 Likert scale (5: perfect, 4: good, 3: acceptable, 2: poor, and 1: very poor) to evaluate each variable.

Figure 1 illustrates the difference between phonological and segmental accuracies. The top row is the pronunciation according to the underlying form of the characters, before the phonological rules are applied, whereas the bottom row is the canonical pronunciation after correct application of the rules.

In this example, the phonological phenomenon include aspiration, tensification, and nasalization rules, in which case the correct—realized substitutions of k>g, s>s, m>⸘ at these positions will be counted as phonological errors. All the positions where the rules occur and their types are marked in advance, so that the raters know which errors to listen to. All other substitutions are considered as segmental errors.

1) Proficiency: As employed in previous studies [8][9][10], proficiency was rated by evaluators’ impressionistic and holistic judgments of the overall utterance in the scale of 1 to 5, without paying attention to specific linguistic features. For example, even if a part of an utterance digresses from the canonical, they can assign high scores if it is perceived as acceptable.

2) Fluency: The evaluators rated fluency based on rate of speech, juncture, pause, and filled pause. For example, novice learners tend to speak slowly and pronounce each syllable separately, which is not observed in native speech. This would discount the fluency score.

3) Phonological accuracy: All syllables where phonological rules occur and their types are marked in advance for the five different phonological phenomenon, which are listed in Table 3. In this way, the raters know which errors to listen to. They counted the number of errors and gave scores based on the count.

4) Segmental accuracy: The raters phonetically transcribed all segments and rated segmental accuracy according to the number of mismatch between the canonical and realized pronunciations.

5) Pitch and accent: The evaluators judged the appropriateness of pitch and accent realized at lexical and sentential levels. For example, if a question is perceived as a statement due to inappropriate pitch realization, the utterance will receive a low score.

Note that the raters had a prior knowledge of the read prompts and therefore, comprehensibility was not included as a variable. That is, the degree of effort required by raters to understand an utterance could not be independently measured in this study by the nature of read speech task. This is also consistent with the previous studies that evaluated read speech by holistic impression of proficiency, instead of comprehensibility.

Moreover, only one pronunciation per word is defined as the canonical form in this experiment. There are certainly variants and regional variations that are also recognized as acceptable pronunciations, which means that some may not be counted as an error. However, we did not allow multiple correct answers in this experiment because there is no consensus on what counts as an ‘acceptable variance,’ and would cause confusion in the scoring process. In case when multiple pronunciation rules can be applied to the same phonemic context, we take the most common pronunciation as the canonical pronunciation. In addition, predefined standard Korean pronunciation exists according to the National Institute of Korean Language, and is the form of Korean that is accepted as a national norm. Therefore, considering that the purpose of current research is a pedagogical application, it seems desirable to keep the correct reference as the gold standard. Such variations that are observed in native speech will be considered with lower priority in later feedback stage.

Figure 1: Pronunciations before and after the application of phonological rules occurring in the phrase [gok̚es’uminda] (would like). The highlighted boxes show where phonological errors may occur, while mispronunciations of segments outside the boxes are counted as segmental errors.

3.3. Speakers

50 speakers of L1 Mandarin Chinese, Japanese, Cambodian, Vietnamese, and Filipino residing in Korea are included. Gender, age, and learner levels are balanced from beginner to advanced. The speaker age ranges from 18 to 60, with the average of 24. Their average length of residence in Korea is 1.5 years, ranging from 2 months to 6 years.
3.4. Evaluators

Each utterance was scored by four native Korean graduate students in Seoul National University with knowledge in Korean phonetics and phonology. Since phonological accuracy was included in the evaluation criteria, it was necessary to recruit raters with detailed knowledge in Korean phonological rules in this evaluation task.

The evaluators practiced scoring with the established guidelines to ensure inter-rater consistency. Before the four raters could officially start scoring, we made sure that the correlation was consistent on the first 50 utterances for training purposes. To ensure consistent scoring quality, we held biweekly training and discussion sessions for monitoring inter-rater consistency throughout the scoring and annotation period, which took about five months.

3.5. Summary of methodological improvements

Compared to the previous experiments, improvements have been made in the coverage and depth of variables and speaker diversities. First, we have increased the number and the scope of variables. Most previous researches mentioned in Section 2.2 studied correlations between two or three variables [9][10][11][12], whereas our study looks at them more comprehensively. We expect this will enable a direct comparison among the variables, and thereby resolve the existing disagreements. Second, the coverage of phonological rules has been increased by adding aspiration and palatalization error types, in addition to lenition, nasalization, and tennification. We believe this is important because phonological rule is a pedagogically meaningful area where the learners may need explicit instruction. Third, we attempt to reduce the disagreements arising from L1 effect by including participants of various backgrounds, with L1 Mandarin Chinese, Japanese, Cambodian, Vietnamese, and Filipino, which is more diverse than other previous experiments with L2 Korean.

4. Experimental Results

The four raters demonstrated general agreement ($\alpha = 0.88$) on the proficiency rating task over 2,500 utterances, suggesting that they share similar intuitive notion of what it meant by holistic impression of proficiency in L2 Korean speech. Inter-rater reliability was calculated by using Cronbach's alpha. The coefficients reported in the previous studies confirm that the results are reliable ($\alpha = 0.82$ [4], 0.88 [9], 0.89 [10], 0.74 [12]).

The mean and standard deviation of proficiency scores are 2.94 and 0.98, respectively, and their distribution is summarized in Table 4. In the following analyses, all raters' scores were averaged to derive a single score for the perceived proficiency of each utterance.

### Table 4: Proficiency score distribution for 2,500 utterances, each rated by four native speakers.

<table>
<thead>
<tr>
<th>Score</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Utterances</td>
<td>1,273</td>
<td>2,906</td>
<td>3,388</td>
<td>1,971</td>
<td>462</td>
<td>10,000</td>
</tr>
</tbody>
</table>

4.1. Linguistic correlates of speech proficiency

We conducted a set of correlation analyses to examine how proficiency rating scores were related to the four linguistic variables defined in the previous section. The results of the linguistic influences on their proficiency evaluation are summarized in Figure 2. All variables are strongly correlated with proficiency scores. Among the variables, proficiency was most strongly correlated with segmental accuracy ($r = 0.81$) and fluency ($r = 0.80$), and relatively weakly correlated with pitch and accent ($r = 0.76$) and phonological accuracy ($r = 0.74$). All correlations are statistically significant ($p<0.0001$).

In other words, speech with higher proficiency ratings was comprised of fewer segmental errors, and was fluently spoken with an appropriate rate of speech. This suggests that the raters similarly relied on segmental and fluency information during their proficiency judgments.

![Mean Proficiency Score per Utterance](image)

**Figure 2:** Linguistic correlation with proficiency scores according to Pearson measure. Correlation is the highest in the order of segmental accuracy ($r = 0.81$), fluency ($r = 0.80$), pitch and accent ($r = 0.76$), and phonological accuracy ($r = 0.74$).

Similar to the correlation analyses in previous studies, we confirm that speech rate is a significant predictor of speech proficiency. However, it is a new finding in this study that segmental accuracy is an even better predictor of proficiency than speech rate. Although the difference between them is small, the finding is significant because most previous studies with L2 Korean did not take segmental accuracy to be an important measure.

Another important finding in this experiment is that pitch and accent is correlated with proficiency ($r = 0.76$). Most studies with L2 Korean did not include pitch and accent as a variable, which can be partially explained by the fact that Korean is a syllable-timed language and it is easy to assume that the difference in pitch and accent would not be perceptually significant. However, the experimental result in the current study is contrary to the expectation. In fact, pitch and accent is shown to have even higher correlation than phonological accuracy, which was initially predicted to be a challenging area according to the related works in L2 Korean.

Therefore, it is surprising that pitch and accent plays a significant role in L2 Korean perception, to an equivalent or higher degree than phonological accuracy.

A possible explanation for this new observation lies in L1 interference by Chinese speakers. The influence of tonal pitch can result in pitch lowering at the end of accentual phrases, which may be perceived as accented and unnatural for the native listeners [8]. In fact, the average score for pitch and accent is lower for L1 Chinese, 3.04, than L1 Japanese, 3.18 with the standard deviations 0.99 and 1.07, each respectively. An independent group t-test was performed to compare the averages between the two groups. The t-statistic was significant at the 0.05 level ($p=0.0071$). This stresses the importance of considering L1 dependent factors in experiment design. It seems that such L1-dependent factors have been underemphasized so far, and pitch and accent have more sensitivity in Korean speech perception than was predicted.
4.2. L1 dependent segmental variation patterns

The experiment results show that segmental accuracy affects the proficiency score more strongly than other factors. We further verified this finding across L1 backgrounds. Figures 3, 4, 5, and 6 show that there are diversities in segmental production patterns across L1 backgrounds. Southeast Asian languages here include Filipino, Cambodian, and Vietnamese, which belong to Austronesian language group. The patterns are organized into five groups (coda deletion and insertion, detensification, aspiration, and vowel epenthesis), which are the salient mispronunciation patterns in non-native Korean speech [20][21]. Chinese and Japanese L1 speakers show higher coda deletion errors than Southeast Asian learners (Figure 3). Detensification and coda insertion error is characteristic of Chinese learners (Figure 4), while vowel epenthesis is characteristic of Japanese speakers (Figure 6). For Southeast Asian speakers, aspiration error is more frequent than other groups (Figure 5). The analysis demonstrates that all L1 groups have salient segmental error patterns that contribute to lower proficiency scores, and confirms the new finding of this research across learners’ L1 backgrounds.

5. Discussion

In Section 2, the research questions were formulated as follows:
1. Which linguistic features predict the holistic speech proficiency of L2 Korean speech?
2. To what extent does the correct realization of phonological rules affect L2 Korean speech proficiency?

In response to the first research question, the results indicate that the effects of segmental accuracy, fluency, pitch and accent, and phonological accuracy are all positively correlated with L2 speakers’ oral proficiency scores obtained from native listener judgements, in the order of their importance. Therefore, segmental accuracy is significantly predictive of L2 Korean proficiency judgment. This resolves the disagreement regarding the relative importance of segmental and suprasegmental features in the previous researches. That is, there was a disagreement whether fluency, segmental accuracy, or both are strong predictors of L2 Korean speech proficiency. We showed that both are highly correlated with proficiency, and segmental accuracy is in fact, more strongly correlated.

As for the second research question, we found that phonological accuracy is least correlated with the proficiency scores. This is interesting because it is contrary to the previous studies’ prediction that phonological accuracy is an L2-specific property in Korean that learners would find difficult. A possible explanation is that not all mispronunciations at the phonological level cause difference in meaning and therefore, does not influence raters’ impression of proficiency. For example, pronouncing the word [ʨʰukʰa] “to congratulate” without applying the phonological rule to aspirate and reading it as the underlying text form, [ʨʰukʰa], does not change the word into a different meaning. The ways in which phonological accuracy affect raters’ perceptual impression can be a subject of future research.

There is a room for improvement in the current study regarding the independence of individual variables, especially between the segmental and pitch accuracies. For example, syllables with tense consonants in Korean are realized with higher pitch than the lax counterparts [22], leading to interaction effect between segmental and pitch variables. In such cases, it can be difficult to identify if the error originates from phonetic or prosodic category. In the future studies, it seems desirable to make more efforts to minimize such effects.
In the future works, we plan to extend this study to spontaneous speech. The findings of read speech may not necessarily generalize into spontaneous speech. We believe that using read speech recording has an advantage in that the reference pronunciation is known, and therefore, error patterns can be efficiently discovered. However, spontaneous speech is closer to natural communication settings, and we plan to extend this study to spontaneous speech in the future works.

6. Conclusion

Using the speech produced by fifty L2 learners of Korean, the current study investigated the linguistic correlates of oral proficiency. Certain errors are believed to entail more perceptual value than others. According to the results in the correlational analysis, our findings for L2 Korean were generally consistent with the previous literature, in that fluency score is a good measure of oral proficiency, including speech rate, juncture, and other temporal features. Furthermore, positive correlation was found between accuracies in Korean phonological rules and oral proficiency scores, which also echoed previous literature that highlighted the importance of assessing L2-specific variables.

The new finding in this study is that segmental accuracy demonstrates the highest correlation with the proficiency scores. Moreover, native listeners are more sensitive to pitch and accent than it was predicted, and deserve more attention in L2 Korean studies. In contrast, phonological accuracy is relatively less important. Building on the previous studies concerning L2 acquisition, the results presented here postulate a verified hypothesis on the oral assessment for L2 Korean. In future works, these findings will not only be extended to spontaneous speech, but also will serve as the ground for the development of automatic assessment module in Korean CAPT system.

7. Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, the Republic of Korea (No. 2015R1D1A2A01061378).

8. References

Automatic Characterisation of the Pronunciation of Non-native English Speakers using Phone Distance Features

K. Kyriakopoulos, M.J.F. Gales, K.M. Knill

ALTA Institute / Engineering Department
Cambridge University, UK

{kk492,mjfg,kate.knill}@eng.cam.ac.uk

Abstract

The distances between and relative movements of phones in acoustic space in language learners have been shown to be indicative of the speaker’s proficiency, in a way that is compact and independent of bias-inducing voice qualities. Typically these features are based on known transcriptions, “read aloud” style tasks. This paper examines the information that can be extracted about speakers from phone distance features (PDFs) when the transcription is unknown. Here, phone distances are obtained by measuring the relative entropy between a distribution trained on the speaker’s manner of pronunciation of each of the phones of the English language and distributions trained on each of the other phones. These features are extracted from untranscribed audio and so rely on automatic speech recognition (ASR) output. The ASR can have high word error rates, as spontaneous, non-native speech is being recognised. Two forms of speaker characterisation are examined using these features: first, the use of PDFs to predict the speaker’s proficiency and second, their use in classifying the mother tongue (L1) of the speaker. For both tasks, recorded answers to sections of the BULATS English Speaking test were used. Using only PDFs for predicting the grade within a Gaussian Process based grader showed performance comparable to using a range of standard fluency style features. This indicates the robustness of PDFs to errors in ASR output. Additionally, the same PDF features can detect with high accuracy the L1 of the speakers from among 21 L1s using a deep neural network based classifier. Experiments on South American Spanish show that it is further possible to discriminate between the speakers’ countries of origin.

1. Introduction

The process by which a language learner improves their pronunciation can be thought of as a path through acoustic space from their initial incorrect pronunciation, affected by their native language (L1) and dialect, towards a pronunciation more closely resembling native speech. It is therefore useful, in the context of Computer Aided Pronunciation Training (CAPT), to be able to automatically characterise the path that the learner is following and evaluate their position along it.

This paper investigates phone distance features (PDFs) for characterising the pronunciation of a non-native speaker of English, from recordings of un-transcribed spontaneous speech. It examines to what extent these features carry information about the speaker’s starting point (L1 and country of origin) and their progress along the learning path (as measured by human-assigned proficiency scores).

Approaches for automatic assessment of pronunciation in the literature often include comparison to native speaker models [1, 2, 3], which can introduce considerable bias with regards to accent and voice quality. The most common features used are prosodic [4, 5, 6] and ASR confidence measures (both at the word and phone level) [7, 2, 1]. Recent investigations have shown promising results based on phone distance based measures [8, 9], on which this paper is based. Most existing systems rely on “read aloud” style tasks with known transcriptions. There are a few systems which can grade spontaneous speech, e.g. [10, 11, 12], however, their scope is usually limited. This paper uses phone distance features, an extension on the vowel formant distances, with the aim of being able to predict the L1 of the speaker from among 21 other L1s, from recordings of un-transcribed spontaneous speech.

2. Phone Distance Features

Pronunciation is a key predictor of speaker proficiency, and is expected to become more native, reducing strain to the listener caused by L1 effects, as the learner progresses up the CEFR lev-
A large component of pronunciation is the manner in which the phones of the language are rendered. Extracting features to represent pronunciation of phones, however, presents a number of difficulties, particularly when dealing with spontaneous speech. First, acoustic models of the phones are not a robust predictor of proficiency, due to the large variation across speakers with different accents, voice qualities and L1s but of otherwise similar level. The forms of native pronunciation being emulated may also vary from speaker to speaker, owing to the large variation in English native speech, creating problems with using native speaker comparisons. The spontaneous nature of the speech further complicates obtaining comparable native speaker models and strengthens the need for general non-native reference approaches.

To overcome these issues, this paper employs an approach based on the distances between phones. Rather than characterising each phone by the distribution of acoustic features in its articulations, it is defined relative to the pronunciation of each of the other phones, with the full set of phone-pair distances describing the speaker’s overall accent. Distances between acoustic models should be more robust to speaker variability than the models themselves. In [9] phonetic pronunciation features consisting of a set of phone-pair distances were proposed for vowels and applied to read speech. Here, the features consist of a set of phone-pair distances covering all 47 phones in English and are applied to both read and spontaneous speech. This yields 1081 distances in total.

Phone distance features should thus robustly represent the pronunciation of a speaker in samples of spontaneous, untranscribed audio, in a way that is compact and independent of the speaker’s irrelevant voice qualities.

Figure 2: Illustration of the phone distance concept

The speaker’s recorded utterances are passed through an ASR and time aligned to the most probable phone sequence given the recognised word sequence. A set of statistical models is then trained to represent the manner of pronunciation of each of the phones in the English language. For each possible phone pair, the distance between the phone models is measured by the symmetric Kullback-Leibler (K-L) divergence [18]. If the statistical models for phones $\phi_i$ and $\phi_j$ are $p(\phi_i)$ and $p(\phi_j)$, respectively, the K-L divergence between the two phones is defined as

$$D_{KL}(p_i||p_j) = \int p(\phi_i) \log \left( \frac{p(\phi_i)}{p(\phi_j)} \right) d\phi_i. \quad (1)$$

Since the K-L divergence is not symmetric and the distance measure should be invariant of the order in which the distributions are taken, one type of the symmetric K-L divergence (also known as Jensen–Shannon divergence [19]) is used, which can be written as

$$D_{JS}(p_i||p_j) = \frac{1}{2} \left[ D_{KL}(p_i||p_j) + D_{KL}(p_j||p_i) \right]. \quad (2)$$

Each phone is modeled by a single multivariate Gaussian with a mean, $\mu$, and diagonal covariance matrix, $\Sigma$. The input vector consists of PLP features, extracted from the speaker’s audio. For each speaker, a model set is trained on all the speech from that speaker. Full recognition is run to acquire 1-best hypotheses from which time aligned phone sequences are generated. Single Gaussian models for each phone are then trained given these alignments. The K-L divergence of $D_{JS}(p_i||p_j)$ is calculated as

$$D_{KL}(p_i||p_j) = \frac{1}{2} \left[ tr \left( \Sigma_i^{-1} \Sigma_j \right) + (\mu_i - \mu_j)^T \Sigma_i^{-1} (\mu_i - \mu_j) - d + \ln \left( \frac{|\det \Sigma_i|}{|\det \Sigma_j|} \right) \right], \quad (3)$$

where $tr(\cdot)$ and $det(\cdot)$ are the operators for the trace and determinant of the matrix, respectively.

If there is insufficient data to train the multivariate Gaussian of a particular phone, the PDFs corresponding to all phone-pairs containing that phone are set to -1. The resultant vector thus also contains information about which phones the speaker avoided pronouncing. This information may itself help predict speaker proficiency and L1. This approach was found to yield higher score prediction accuracy than replacing missing PDFs with the mean value based on other speakers in the training set.

This paper investigates the hypothesis that features extracted in this way are strongly representative of the speaker’s accent, by evaluating how well they predict the speaker’s proficiency, native language and country of origin.

### 3. Data

The experiments reported in this paper are based on candidate responses to the spoken component of the Business Language Testing Service (BULATS), provided by Cambridge English Language Assessment. The BULATS speaking test has five sections, all related to business scenarios [20]. Section A consists of short responses to prompted questions. Candidates read 8 sentences aloud in Section B. Sections C-E consist of spontaneous responses of several sentences in length to a series of spoken and visual prompts. Candidates are scored on a scale from 0 to 30, based on their overall proficiency, mapping to standard CEFR levels as shown in Table 1.

<table>
<thead>
<tr>
<th>BULATS score range</th>
<th>Level description</th>
<th>CEFR level</th>
</tr>
</thead>
<tbody>
<tr>
<td>29-30</td>
<td>Upper advanced</td>
<td>C2</td>
</tr>
<tr>
<td>25-29</td>
<td>Advanced</td>
<td>C1</td>
</tr>
<tr>
<td>20-25</td>
<td>Upper intermediate</td>
<td>B2</td>
</tr>
<tr>
<td>15-20</td>
<td>Intermediate</td>
<td>B1</td>
</tr>
<tr>
<td>10-15</td>
<td>Elementary</td>
<td>A2</td>
</tr>
<tr>
<td>5-10</td>
<td>Beginner</td>
<td>A1</td>
</tr>
<tr>
<td>0-2</td>
<td>Fail/Incomprehensible</td>
<td>pre-A1</td>
</tr>
</tbody>
</table>

Table 1: Equivalence between BULATS scores and CEFR levels (adapted from [21])
For the purposes of the experiments in this paper, the data is segmented into four non-overlapping data sets: BLX0, which is used to train the ASR, TRN, which is used to train the regressors and classifiers, EVL1 which is used to evaluate the classifiers, and EVL2, which is used to evaluate the regressors, classifiers and ASR.

It is important that none of the speakers in the training or evaluation sets (TRN, EVL1 or EVL2) are present in the ASR training set, so that ASR induced error in the PDFs is uniform across the data used to train and evaluate the other systems.

For each speaker in each of the four sets there is available the audio, a human assigned proficiency score and meta-data describing the candidate’s L1 and country of origin. BLX0 and EVL2 additionally are accompanied by crowd-sourced transcriptions, which are used in training and evaluating the ASR. Finally, each speaker in EVL2 has also been scored by a highly qualified expert grader, who are known to have extremely high inter-annotator agreement (upwards of 0.95). Evaluation of the regressor (which is done by Pearson Correlation Coefficient - PCC - of actual to predicted grades) is therefore only performed on EVL2.

When evaluating country of origin on speakers whose L1 is known to be Spanish, subsets of TRN and EVL1 including only Spanish speakers (called TRN_S and EVL1_S) are used.

Tables 2 and 3 show the breakdown of L1 and country of origin in the latter three sets (TRN, EVL1 and EVL2) used in the course of this investigation.

<table>
<thead>
<tr>
<th>L1</th>
<th>TRN</th>
<th>EVL1</th>
<th>EVL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>4502</td>
<td>2156</td>
<td>-</td>
</tr>
<tr>
<td>Tamil</td>
<td>1468</td>
<td>790</td>
<td>-</td>
</tr>
<tr>
<td>Gujarati</td>
<td>1015</td>
<td>230</td>
<td>94</td>
</tr>
<tr>
<td>Hindi</td>
<td>563</td>
<td>294</td>
<td>-</td>
</tr>
<tr>
<td>Telugu</td>
<td>462</td>
<td>250</td>
<td>-</td>
</tr>
<tr>
<td>Malayalam</td>
<td>395</td>
<td>184</td>
<td>-</td>
</tr>
<tr>
<td>Bengali</td>
<td>333</td>
<td>152</td>
<td>-</td>
</tr>
<tr>
<td>Russian</td>
<td>303</td>
<td>170</td>
<td>-</td>
</tr>
<tr>
<td>French</td>
<td>291</td>
<td>115</td>
<td>36</td>
</tr>
<tr>
<td>Polish</td>
<td>258</td>
<td>69</td>
<td>39</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>245</td>
<td>67</td>
<td>37</td>
</tr>
<tr>
<td>Kannada</td>
<td>226</td>
<td>131</td>
<td>-</td>
</tr>
<tr>
<td>Arabic</td>
<td>202</td>
<td>51</td>
<td>39</td>
</tr>
<tr>
<td>Portuguese</td>
<td>176</td>
<td>78</td>
<td>-</td>
</tr>
<tr>
<td>Dutch</td>
<td>173</td>
<td>47</td>
<td>32</td>
</tr>
<tr>
<td>Thai</td>
<td>144</td>
<td>43</td>
<td>36</td>
</tr>
<tr>
<td>Japanese</td>
<td>135</td>
<td>68</td>
<td>-</td>
</tr>
<tr>
<td>Marathi</td>
<td>106</td>
<td>67</td>
<td>-</td>
</tr>
<tr>
<td>Italian</td>
<td>107</td>
<td>41</td>
<td>-</td>
</tr>
<tr>
<td>Korean</td>
<td>90</td>
<td>53</td>
<td>-</td>
</tr>
<tr>
<td>Oriya</td>
<td>65</td>
<td>26</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Breakdown of number of speakers in each data set by native language (L1)

<table>
<thead>
<tr>
<th>Country</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRN_S</td>
<td></td>
</tr>
<tr>
<td>EVL1_S</td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td>798</td>
</tr>
<tr>
<td>Mexico</td>
<td>3208</td>
</tr>
<tr>
<td>Spain</td>
<td>359</td>
</tr>
</tbody>
</table>

Table 3: Breakdown of speakers in Spanish data set by country of origin

the accuracy of standard commercial “off-the-shelf” ASR systems is too low for non-native learner English. Instead specific ASR systems are trained.

A stacked-hybrid DNN-HMM acoustic model is used for ASR. It is trained on a 108 hour (1075 speaker) Gujarati L1 BULATS data set with merged crowd-sourced transcriptions [22] (BLX0, mentioned above), using the HTK toolkit [23, 24].

4.2. Regression and Classification DNNs

Both regression and classification tasks are performed using deep neural networks (DNNs), constructed using Torch. The networks each have 6 hidden layers (8 layers in total), with 1200 hidden units per hidden layer. Dropout of 50% of the units is implemented at each layer. Weight decay and stop validation are employed to further improve generalisation. The regression networks are trained for minimum MSE, while the classification networks are trained for minimum cross-entropy. The networks are trained in batches of 1000 speakers at a time, for 200 iterations. The regression networks are evaluated (on the evaluation sets described in Section 3, using the Pearson correlation coefficient (PCC) and mean square error (MSE) between predicted and actual results. The classifiers are evaluated by the percentage of speakers correctly classified.

4.3. Baseline Features

The fluency and prosodic features described in [5], plus the number and fraction of disfluencies, fraction of speech in the recording duration and vowel frequency, are used as baseline features for both the grader and the classifiers. Each system is implemented using just baseline features, just phone distance features and the two combined.

5. Results and Discussion

The results of four experiments using the systems described in the previous sections are presented and discussed in the following subsections. First, the ASR trained on BLX0 is evaluated on EVL2 to obtain indicative ASR word error rates, by profi-
ciency and language. Using phone distance features, extracted as described in Section 2, using the ASR trained on BLX0, a score predictor is trained on TRN and evaluated on EVL2, an L1 classifier is trained on TRN and evaluated on EVL1 and a country of origin classifier is trained on TRN and evaluated on EVL1.S.

5.1. ASR Performance

This ASR trained on BLX0 as described in Section 3.1 is evaluated using the transcriptions of EVL2. It has an overall word error rate (WER) of 47.5% and a phone error rate (PER) of 33.9%. The relatively high PER suggests a considerable amount of noise will be present in the phone distance features. Note, however, that the inherent inaccuracies and noisiness of crowd-sourced transcriptions may lead error rates to be exaggerated.

The WER for the mixed-L1 data is further broken down by L1 and CEFR level in Table 4. As can clearly be seen, recognition error decreases with increasing proficiency of the speaker, a result that holds across all L1s. This is to be expected as higher proficiency speakers are likely to speak more intelligibly and their speech is therefore easier for the ASR to recognise.

5.2. Score Prediction

The baseline and phone distance features are now used to build an automatic grader, which attempts to predict human-assigned scores, trained on the ordinary human grader assigned scores in TRN and evaluated on the expert-assigned scores in EVL2.

As can be seen in Table 5 below, PDFs outperform baseline features in both MSE and PCC when used on their own and yield considerable improvements when used in combination with them. This is particularly promising when considering that these results are in the presence of considerable ASR error and that the scores being predicted are general proficiency scores and not pronunciation-specific.

5.3. Candidate L1 Classification

Having established that PDFs are strong predictors of proficiency, a DNN classifier is now built to determine whether they can also be used to predict candidates’ native languages.

Table 6 below shows the performance of the same features when used to classify the speakers’ native language (L1) from among 21 candidates. The baseline features already perform significantly better than random chance, but the PDF-trained DNNs significantly outperform them, suggesting the phone distances are indeed indicative of speaker L1. Combining PDFs and baseline features degrades the accuracy slightly, suggesting most information about L1 contained in the baseline features is also captured by the PDFs.

As can be seen in Table 7, L1 classification performance is highest for those languages with the most data in the training and testing sets (e.g. Gujarati, Spanish and Tamil) and lowest for those with the least data (e.g. Marathi, Italian, Korean and Oriya).

The identities of the L1s that speakers with each L1 are most frequently misclassified as confirm expectations regarding similarities between languages. Romance languages (Italian, Portuguese and French) are most frequently confused with Spanish, Indo-Aryan languages (Gujarati, Marathi, Bengali and Oriya) are most commonly misclassified as Hindi, and Dravidian languages (Telugu and Malayalam) most commonly misclassified as Tamil. This is confirmed by observing the confusion matrices in Tables 8, 9 and 10 below, which demonstrate the majority of incorrect classifications within each group of languages are as other L1s in the same group.

Breaking down L1 classification accuracy by CEFR level (Table 11), the classifier is the least accurate (with both baseline and PDF features) for poor speakers (A1) and its performance then increases with proficiency. This is attributed to the decrease in WER with increasing proficiency. As WER falls, features are more representative of the speaker’s actual pronunciation and less noisy, resulting in more powerful classification.

---

**Table 4:** Word error rates (WER) of ASR on indicative Mixed-L1 data set (EVL2) broken down by L1 and CEFR level

<table>
<thead>
<tr>
<th>L1</th>
<th>Spanish</th>
<th>Arabic</th>
<th>Dutch</th>
<th>French</th>
<th>Thai</th>
<th>Viet.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>69.8</td>
<td>69.7</td>
<td>78.7</td>
<td>55.8</td>
<td>65.4</td>
<td>65.4</td>
</tr>
<tr>
<td>A2</td>
<td>58.7</td>
<td>67.4</td>
<td>45.7</td>
<td>48.0</td>
<td>56.0</td>
<td>55.9</td>
</tr>
<tr>
<td>B1</td>
<td>48.6</td>
<td>47.2</td>
<td>41.3</td>
<td>45.0</td>
<td>50.7</td>
<td>53.5</td>
</tr>
<tr>
<td>B2</td>
<td>47.1</td>
<td>47.3</td>
<td>40.3</td>
<td>45.0</td>
<td>48.1</td>
<td>56.6</td>
</tr>
<tr>
<td>C</td>
<td>48.8</td>
<td>48.6</td>
<td>43.1</td>
<td>36.7</td>
<td>41.3</td>
<td>43.6</td>
</tr>
<tr>
<td>All</td>
<td>50.9</td>
<td>52.0</td>
<td>42.5</td>
<td>43.6</td>
<td>50.2</td>
<td>33.0</td>
</tr>
</tbody>
</table>

**Table 5:** Performance (PCC and MSE) of DNN grader described in §4.2, trained on TRN and evaluated on EVL2

<table>
<thead>
<tr>
<th></th>
<th>PCC</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>0.737</td>
<td>26.4</td>
</tr>
<tr>
<td>PDF</td>
<td>0.751</td>
<td>23.6</td>
</tr>
<tr>
<td>Base+PDF</td>
<td>0.832</td>
<td>15.8</td>
</tr>
</tbody>
</table>

**Table 6:** Accuracy (percentage of the speakers correctly classified) for DNN L1 classifier described in §4.2, trained on TRN and evaluated on EVL1 and EVL2

<table>
<thead>
<tr>
<th>L1</th>
<th>% Correct</th>
<th>% # Speakers Most confused</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>97.7</td>
<td>4502</td>
</tr>
<tr>
<td>Tamil</td>
<td>76.7</td>
<td>1468</td>
</tr>
<tr>
<td>Gujarati</td>
<td>74.5</td>
<td>1015</td>
</tr>
<tr>
<td>Hindi</td>
<td>62.3</td>
<td>563</td>
</tr>
<tr>
<td>Marathi</td>
<td>0.0</td>
<td>106</td>
</tr>
<tr>
<td>Italian</td>
<td>2.4</td>
<td>107</td>
</tr>
<tr>
<td>Korean</td>
<td>3.7</td>
<td>90</td>
</tr>
<tr>
<td>Oriya</td>
<td>0.0</td>
<td>65</td>
</tr>
</tbody>
</table>

**Table 7:** Breakdown by L1 of accuracies for L1 classifier evaluated on EVL1, using baseline + PDF, for L1s with the most and leasts number of speakers in the training set (TRN), along with L1 that its speakers are most frequently misclassified as.
Performance rises faster with proficiency for baseline features than for PDFs, suggesting the latter are more robust to ASR error. Performance levels out and slightly dips as proficiency enters the C levels, which can be attributed to the speakers’ pronunciation becoming more similar to native speech and therefore less indicative of their L1.

Table 8: Percentage of speakers of each Romance L1 classified as other Romance languages

<table>
<thead>
<tr>
<th>L1</th>
<th>Spanish</th>
<th>French</th>
<th>Portugese</th>
<th>Italian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>97.7</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>French</td>
<td>16.5</td>
<td>43.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Portugese</td>
<td>21.8</td>
<td>24.4</td>
<td>29.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Italian</td>
<td>36.6</td>
<td>26.8</td>
<td>0.0</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 9: Percentage of speakers of each Indo-Aryan L1 classified as other Indo-Aryan languages

<table>
<thead>
<tr>
<th>L1</th>
<th>Gujarati</th>
<th>Hindi</th>
<th>Bengali</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td>74.3</td>
<td>10.9</td>
<td>1.7</td>
</tr>
<tr>
<td>Hindi</td>
<td>11.6</td>
<td>62.9</td>
<td>0</td>
</tr>
<tr>
<td>Bengali</td>
<td>10.5</td>
<td>55.9</td>
<td>20.3</td>
</tr>
<tr>
<td>Marathi</td>
<td>3.0</td>
<td>74.6</td>
<td>0</td>
</tr>
<tr>
<td>Oriya</td>
<td>3.8</td>
<td>73.1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 10: Percentage of speakers of each Dravidian L1 classified as other Dravidian languages

<table>
<thead>
<tr>
<th>L1</th>
<th>Tamil</th>
<th>Telugu</th>
<th>Malayalam</th>
<th>Kannada</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamil</td>
<td>76.7</td>
<td>8.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Telugu</td>
<td>37.6</td>
<td>27.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Malayalam</td>
<td>49.5</td>
<td>23.4</td>
<td>12.4</td>
<td>0</td>
</tr>
<tr>
<td>Kannada</td>
<td>24.4</td>
<td>19.1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 11: Detection rate for speaker L1 classifier, evaluated on EVL1, broken down by CEFR level

<table>
<thead>
<tr>
<th>CEFR Level</th>
<th>%Baseline</th>
<th>%PDF</th>
<th>%Baseline+PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>53.1</td>
<td>69.0</td>
<td>66.5</td>
</tr>
<tr>
<td>A1</td>
<td>41.3</td>
<td>60.0</td>
<td>56.5</td>
</tr>
<tr>
<td>A2</td>
<td>47.0</td>
<td>60.1</td>
<td>58.5</td>
</tr>
<tr>
<td>B1</td>
<td>55.2</td>
<td>70.0</td>
<td>67.8</td>
</tr>
<tr>
<td>B2</td>
<td>53.5</td>
<td>70.5</td>
<td>67.7</td>
</tr>
<tr>
<td>C1</td>
<td>56.3</td>
<td>71.8</td>
<td>67.8</td>
</tr>
<tr>
<td>C2</td>
<td>50.0</td>
<td>57.5</td>
<td>77.5</td>
</tr>
</tbody>
</table>

5.4. Candidate Country of Origin Classification

A similar methodology to the previous section is now employed to attempt to classify the speaker’s country of origin, to see whether the predictive powers of PDFs can narrow foreign accents down further than the level of L1.

As seen in Tables 12 and 13, the classifier performs very well when identifying the country of origin of speakers with a known L1. This suggests that PDFs (as well as, to some extent, the baseline features) are able to capture phonological differences in the way speakers of the same L1 with different regional dialects pronounce the phones of English. As with the L1 classifier, PDFs considerably outperform the baseline features, while adding the baseline features to the PDF only slightly increases performance.

As with L1 classification, performance increases with CEFR level (Table 14), again clearly attributable to decreasing word error rates.

6. Conclusions

Phone distance features were presented as a means of representing the relative manner in which a learner renders the phones of the English language, based only on recordings of spontaneous speech. They were shown to be a strong predictor of the speaker’s proficiency (as assigned by human graders), their native language and, at least for speakers of Spanish, their country of origin, performing significantly better than baseline features at all three prediction tasks. Although they depend on accurate ASR for best results and their predictive power decreases with increasing WER, they were found to be more robust to ASR errors than the baseline features.

7. Acknowledgements

This research was funded under the ALTA Institute, Cambridge University. Thanks to Cambridge English, Cambridge University, for supporting this research and providing access to the BULATS data.

8. References

### Table 12: Accuracy (percentage of the speakers correctly classified) for DNN Country of Origin classifier described in §4.2, trained on the Spanish speakers in TRN (aka. TRN§) and evaluated on the Spanish speakers in EVL1 (aka EVL1§).

<table>
<thead>
<tr>
<th></th>
<th>Base 77.3</th>
<th>PDF 85.5</th>
<th>Base+PDF 87.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>51.1</td>
<td>34.4</td>
<td>51.1</td>
</tr>
<tr>
<td>A2</td>
<td>55.0</td>
<td>75.7</td>
<td>78.1</td>
</tr>
<tr>
<td>B1</td>
<td>79.7</td>
<td>88.7</td>
<td>90.3</td>
</tr>
<tr>
<td>B2</td>
<td>81.8</td>
<td>89.8</td>
<td>89.6</td>
</tr>
<tr>
<td>C1</td>
<td>77.3</td>
<td>86.3</td>
<td>85.8</td>
</tr>
<tr>
<td>C2</td>
<td>86.7</td>
<td>86.7</td>
<td>93.3</td>
</tr>
</tbody>
</table>

### Table 13: Breakdown by country of detection rates for country classifier, using baseline + PDF features, trained and evaluated on Spanish-only data (TRN§/EVL1§).

<table>
<thead>
<tr>
<th>Country</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>71.5</td>
</tr>
<tr>
<td>Colombia</td>
<td>45.5</td>
</tr>
<tr>
<td>Mexico</td>
<td>97.5</td>
</tr>
</tbody>
</table>

### Table 14: Detection rate for speaker country classifier, evaluated on Spanish data, broken down by CEFR level.

<table>
<thead>
<tr>
<th>CEFR</th>
<th>%Baseline</th>
<th>%PDF</th>
<th>%Baseline+PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>77.3</td>
<td>85.5</td>
<td>87.0</td>
</tr>
<tr>
<td>A1</td>
<td>51.1</td>
<td>34.4</td>
<td>51.1</td>
</tr>
<tr>
<td>A2</td>
<td>55.0</td>
<td>75.7</td>
<td>78.1</td>
</tr>
<tr>
<td>B1</td>
<td>79.7</td>
<td>88.7</td>
<td>90.3</td>
</tr>
<tr>
<td>B2</td>
<td>81.8</td>
<td>89.8</td>
<td>89.6</td>
</tr>
<tr>
<td>C1</td>
<td>77.3</td>
<td>86.3</td>
<td>85.8</td>
</tr>
<tr>
<td>C2</td>
<td>86.7</td>
<td>86.7</td>
<td>93.3</td>
</tr>
</tbody>
</table>


Mispronunciation Diagnosis of L2 English at Articulatory Level Using Articulatory Goodness-Of-Pronunciation Features

Hyuksu Ryu¹, Minhwa Chung¹

¹Department of Linguistics, Seoul National University, Seoul, Republic of Korea
oster01@snu.ac.kr, mchung@snu.ac.kr

Abstract
This paper proposes a method to provide an articulatory diagnosis of English produced by Korean learners using articulatory Goodness-Of-Pronunciation (aGOP) features, which are based on the distinctive feature theory in phonology. Previous studies on mispronunciation diagnosis have mainly dealt with pronunciation errors at phone-level. They inform learners of which phone is recognized as a diagnosis, when the corresponding segment is realized as a mispronunciation. However, to provide learners more effective corrective feedback, diagnosis had better be performed at articulatory-level, such as place and manner of articulation, rather than at phone-level. This study aims to provide automatic articulatory diagnosis using articulation-based confidence scores. At first, the speech of learners is forced-aligned and recognized to compute the GOP and aGOPs. When the forced-aligned segment is a consonant, articulatory diagnosis is conducted in three articulatory categories: voicing, place of articulation, and manner of articulation. Otherwise, diagnosis is performed in terms of rounding, height, and backness corresponding to articulatory characteristics of vowels. Experimental results show that F1 scores for voicing, place, and manner corresponding to consonants are 0.828, 0.754, and 0.781, respectively, whereas F1 score for rounding, height, and backness corresponding to vowels are 0.843, 0.782, and 0.824, respectively. These results indicate that the proposed method yields effective articulatory diagnosis.

Index Terms: articulatory Goodness-Of-Pronunciation, mispronunciation diagnosis, CAPT, English produced by Korean learners

1. Introduction

Corrective feedbacks explaining where learners are making pronunciation errors and how to correct them are essential for Computer-Assisted Language Learning (CALL) and Computer-Assisted Pronunciation Training (CAPT) systems [1]. That is, mispronunciation detection and diagnosis of mispronunciation using speech technology are necessary for conducting effective CALL/CAPT.

There have been several studies to detect pronunciation errors of learners [2][3][4][5]. The study of [2] suggested an extended recognition network (ERN), which expands pronunciation dictionaries of learners by predicting frequent erroneous pronunciation sequences. When the erroneous pronunciation sequences are recognized, it is considered that learners made pronunciation errors. However, ERN approach has difficulties in identifying mispronunciation patterns that learners frequently show in terms of each L1-L2 pair. Also, it is difficult to guarantee that ERN covers most of the possible mispronunciations [6].

Another approach for detecting pronunciation errors is using confidence scores such as Goodness-Of-Pronunciation (GOP) [3][4]. Confidence score-based approach has virtues that it has L1/L2 independence and it is easy to compute [7]. However, it is difficult to provide corrective feedback, since learners do not know how to interpret with confidence scores alone and improve their pronunciation with the scores. Diagnosis for the detected pronunciation errors was not provided in these researches.

Several previous studies [6][8][9] conducted diagnosis for mispronunciation as well as detection of pronunciation errors. Li et al. [6] suggested multi-distribution DNN (MD-DNN) by using acoustic features, graphemes, and canonical pronunciation as inputs of DNN to predict actual pronunciation of learners. When the predicted pronunciation is different from the canonical pronunciation, it is considered as a mispronunciation. Wang and Lee [8] proposed hierarchical multi-layer perceptrons (MLPs). First MLP is binary and classifies each frame as correct or incorrect. Then, second MLP classifies each frame identified as incorrect by the first MLP into one of the Error Patterns (EP) as diagnosis. Xie et al. [9] extracted landmark features for nasal codas spoken by learners of Chinese, and detected pronunciation errors by applying SVM. In these studies, diagnosis is performed in a hierarchical way as shown in Figure 1.

At first, the pronunciation error detector that they proposed distinguishes between mispronunciation and correct pronunciation. In addition to binary mispronunciation detector, diagnosis is carried out for instances which are correctly detected as mispronunciations (True Rejection in Figure 1). The diagnosis performance is reported by diagnosis error rate (DER), which is defined as the percentage of incorrectly recognized phones among correctly identified as a mispronunciation.

These hierarchical approaches for diagnosis have a limitation that they provide diagnosis at phone level only. For example, let’s assume a learner pronounced a word “give” /gɪv/ as /gɪb/. When the CAPT system detects pronunciation error at coda position and recognizes the phone as /b/, the system reports a diagnosis of /v/→/b/. However, for more effective corrective feedback or learners, it had better provide diagnosis information at articulatory level such as voicing, place, and...
manner of articulation, not just at phone level. In the example above, it would be more effective if the system tells a learner that there is a diagnosis of fricative→stop at articulatory level of the manner of articulation, rather than diagnosis of /s/→/h/ at phone level. In addition, the diagnosis procedure presented in the existing studies is performed in two steps; detection and recognition. Since mispronunciation detection errors and recognition errors are piled up, diagnosis accuracy could be degraded. For instance, diagnosis will be incorrect when mispronunciation detection shows false acceptance or false reject, as well as when the segment produced by learners is incorrectly recognized, even if the system detects mispronunciation correctly. When considering limitations mentioned above, it will be helpful when mispronunciation diagnosis is performed at articulatory level.

There are several studies utilizing articulatory features for pronunciation assessment and mispronunciation detection. The study of Ryu and Chung [10] proposes articulatory Goodness-Of-Pronunciations (aGOPs) as novel features for pronunciation assessment in English spoken by Korean learners. Furthermore, Li et al. [5] extended GOP into speech attributes to detect mispronunciation of onset consonants in learners’ Chinese by decision trees. By using speech attributes, they have shown the possibility to provide corrective feedback to improve pronunciation. However, there is a limitation that they did not present the actual experiments about diagnosis or corrective feedback. The goal of this paper is to propose a method to provide an articulatory diagnosis in English produced by Korean learners using articulatory Goodness-Of-Pronunciation (aGOP) based on the distinctive feature theory. The remainder of this study is organized as follows. In Section 2, we describe theoretical backgrounds and computational models regarding articulatory features. Section 3 presents the details of methods for diagnosis modeling at articulatory level. In Section 4, we provide quantitative analysis of salient mispronunciation patterns in English by Korean learners. The results of experiments for articulatory diagnosis are presented in Section 5, which is followed by conclusion in Section 6.

2. Articulatory features

Phonemes are differentiated by phonological features [11]. For example, /p/ and /b/ are distinguished by the phonological feature of voicing; /b/ has the feature of voicing ([+voice]), while /p/ does not ([−voice]). In other words, the voicing feature is a phonological criterion which makes the two phonemes ‘distinctively’ different. The minimum unit that discriminates phonemes in a language is called distinctive features. Therefore, phonemes are represented by a set of distinctive features, which is called natural class [11]. Chomsky and Halle [12] explained that the distinctive features have only two values of presence or absence such as [+voice] or [−voice], and that it is possible to distinguish various phonemes by using a natural class.

The study of [10] suggested articulatory-based GOPs for pronunciation assessment. The aGOP features are used to compare articulatory characteristics between natives and learners regarding the articulatory attributes. The aGOP features are computed based on phone GOP [4] as the normalized posterior probability for each articulatory attribute as shown in (1).

\[ a\text{GOP}^k(p) \equiv \log \frac{P(\text{aGOP}_q^k)}{\max_q P(\text{aGOP}_q^k)} \]

where \( k \) and \( q^k \) denote the sort of articulatory attribute presented in [12] and the canonical value of the \( k^{th} \) articulatory attribute at the position of the forced-aligned target phone \( p \), respectively. \( N(p) \) and \( P(\text{aGOP}_q^k) \) mean the number of frames composing target phone \( p \) and the probability of observing \( q^k \) given \( q^k \), respectively. The higher the value of GOP and aGOP, the higher the likelihood that learners utter the canonical phone.

Articulatory features are classified into three categories such as manner, place, and laryngeal [11]. Articulatory features used in this study consist of 24 attributes such as sonorant and continuant (manner), labial and round (place), and voice (laryngeal). The details of articulatory features and the corresponding phones are presented in [10].

The study of [10] used aGOP features as novel predictors for automatic pronunciation assessment of English produced by Korean learners. In addition to well-known features such as Rate of Speech (ROS) in previous studies, [10] showed that the performance was improved by including aGOP features and applying the statistical method such as best subset selection.

The study of [5] also proposed similar articulatory features for mispronunciation detection of Mandarin learners. However, they limited the focus of the study on onset consonants, while this paper deals with every phone including vowels. Also, they performed the articulatory modeling based on categories [13]. For example, the articulatory model of [13] classifies the category of ‘place’ into multiple attributes, such as bilabial, alveolar, and dental. This kind of modeling has a limitation that it shows low performance when the category has multiple attributes, such as place, as discussed in [13]. On the contrary, aGOP features suggested in [10] and this study are computed based on each attribute, not category. For instance, our model classifies the articulatory attribute of ‘alveolar’ into a binary value of presence or absence. Furthermore, we specify articulatory attributes in more details by using the phonological theory. Therefore, we can compute more various information and use them for mispronunciation diagnosis.

3. Method

3.1. Corpus and annotation

We use the ETRI (Electronics and Telecommunications Research Institute) English speech corpus produced by Korean speakers. The corpus consists of 21,110 sentences (21 hours) spoken by 151 learners.

Ten Korean annotators participate in phone-level transcription by the procedure in [14]. They are native Koreans who can speak English as L2 and have expertise in phonetics or phonology. Transcribers present 88.13% of the phone-level agreement, which means the annotation is reliable [15]. The result of transcriptions shows 6.32% of overall mispronunciation rate.

3.2. Acoustic model

We use an acoustic model using WSJ corpus of 37,000 sentences spoken by North American native English speakers [16]. The CD-DNN-HMM acoustic model [17] is trained using 39-dim. MFCC+Δ+ΔΔ features. In addition to phone acoustic model, we separately train the articulatory attribute models to calculate aGOPs. The parameter configuration and architecture of DNN used in this study is provided by the default configurations of the Kaldi toolkit [18].
3.3. Diagnosis modeling

Mispronunciation diagnosis at the articulatory level suggested in this study is performed as shown in Figure 2. We use GOP [4] and aGOPs [10] as predictors for articulatory diagnosis. Based on forced-alignment, we examine whether the corresponding segment is a consonant or a vowel. When the segment is a consonant, mispronunciation diagnosis is performed in voicing, place of articulation, and manner of articulation, since consonants are phonetically classified in terms of these three dimensions at articulatory level [19]. On the contrary, if the segment is a vowel, mispronunciation diagnosis is carried out at an articulatory level in terms of rounding, height, and backness, corresponding to articulatory characteristics of vowels. Although phonetic transcription of mispronunciation can be converted into articulatory transcription by rules and phone-level mispronunciation can be detected using GOP, rule-based articulatory conversion could be inaccurate, when phone-level mispronunciation detection shows false results. Thus, we perform separate diagnosis modeling at articulatory level using aGOP features as well as GOP.

For consonants, mispronunciation diagnosis at articulatory level is carried out separately in terms of voicing, place, and manner as demonstrated in Figure 3. As stated above, GOP and 24 aGOPs for the associated articulatory attributes are used as predictors for diagnosis modeling. The binary value of correctness or incorrectness is used for the response variable of each diagnosis model by comparing the canonical pronunciation and the actual realization in terms of articulatory levels. For example, let’s assume an observation of /s/ → /T/, which means the canonical pronunciation is voiced alveolar fricative. In this case, the value of ‘correctness’ is assigned to the articulatory level of voicing and manner of articulation, since these articulatory levels of the canonical pronunciation and the actual realization are matched as voiceless and fricative, respectively. On the other hand, in terms of place of articulation, the canonical and the actual are dental and alveolar respectively, so the value of incorrectness is given to the articulatory level of place.

Feed-Forward Neural Network (FFNN) is applied to each diagnosis modeling of mispronunciation at articulatory level. All the diagnosis models are implemented using TensorFlow [20]. The FFNNs are configured with fully connected hidden layers and a softmax output layer. We tune the number of hidden layers between 3 and 7 and the number of nodes in the range [128, 256, 512, 1024]. Each layer is followed by an Exponential Linear Unit (ELU) [21], which is known to show better performance compared to Rectified Linear Unit (ReLU). We also apply batch normalization [22] and dropout with rate of 0.5. Weights are initialized in the form of He initialization [23]. We use a learning rate of 0.005 with the Adam optimizer [24] which uses cross-entropy for the loss function. The FFNNs are trained for 10,000 epochs with a batch size of 100 and early-stopping is performed based on the accuracy of the validation set.

For vowels, mispronunciation diagnosis at articulatory level is separately performed in terms of rounding, height, and backness as shown in Figure 4. For instance, when there is an observation of /a/ → /o/, the canonical pronunciation /a/ is an unrounded low back vowel, while the actual realization /o/ is a rounded mid back vowel. Therefore, ‘incorrectness’ is assigned to rounding and height where the value of the canonical (unrounded low) does not match with that of the actual realization (rounded mid). On the contrary, the articulatory level of backness has ‘correctness’, because both the canonical and the actual are back vowel. The detailed configurations of diagnosis modeling are identical to that of consonants as mentioned above.

4. Quantitative analysis of salient mispronunciations

Prior to experiments for mispronunciation diagnosis at the articulatory level, we quantitatively analyze the tendency of mispronunciation patterns in English produced by Korean learners using the corpus mentioned in 3.1. By the result of corpus anal-
ysis, there are entirely 602,810 phones. Among them, 38,100 phones are marked as incorrect, which shows 6.32% of variation rate. We choose phones with the variation rate above the overall variation rate (6.32%) and the entire frequency over 500 instances as salient phones. Details of salient phones and the corresponding variation frequency and variation rates are presented in Table 1. The sum of variation frequencies of nine salient phones is 26,553 instances, occupying around 70% of entire variation frequency (38,100 instances).

Table 1: Frequency and rate of salient phones

<table>
<thead>
<tr>
<th>Category</th>
<th>Phone</th>
<th>Entire Freq.</th>
<th>Variation Freq.</th>
<th>Variation rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consonant</td>
<td>/z/</td>
<td>12,603</td>
<td>3,425</td>
<td>27.18%</td>
</tr>
<tr>
<td></td>
<td>/θ/</td>
<td>14,967</td>
<td>3,488</td>
<td>23.30%</td>
</tr>
<tr>
<td></td>
<td>/θ/</td>
<td>3,392</td>
<td>613</td>
<td>18.07%</td>
</tr>
<tr>
<td></td>
<td>/l/</td>
<td>9,492</td>
<td>1,434</td>
<td>15.11%</td>
</tr>
<tr>
<td></td>
<td>/l/</td>
<td>23,814</td>
<td>2,999</td>
<td>12.59%</td>
</tr>
<tr>
<td></td>
<td>/l/</td>
<td>49,804</td>
<td>4,206</td>
<td>8.45%</td>
</tr>
<tr>
<td>Vowel</td>
<td>/a/</td>
<td>11,327</td>
<td>2,381</td>
<td>21.02%</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>10,204</td>
<td>1,690</td>
<td>16.56%</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>44,490</td>
<td>6,317</td>
<td>14.20%</td>
</tr>
</tbody>
</table>

To determine the most noticeable variations which appear only in the learners’ speech, we choose variations which are more frequent than in native speech [25] among salient phones in Table 1. The most noticeable variations are determined by adopting the analysis of [25], and we consider the most noticeable variations as salient mispronunciation patterns. Salient variations for /d/ and /t/ are deletion in consonant clusters (ex. ‘just’ /dθæst/ → /θæst/) and flapping (ex. ‘body’ /bæθdi/ → /bæθdI/). Because such variations frequently appear in natives’ speech [25], they are not included in the list of salient mispronunciation patterns. Details of salient mispronunciation in consonants and vowels are shown in Table 2 and Table 3, respectively. The parenthesis in the Canonical column (Canon.) denotes the corresponding frequency of the canonical phones where variations occur.

Table 2: Salient mispronunciation patterns in consonants in terms of articulatory level

<table>
<thead>
<tr>
<th>Level</th>
<th>Canon. (Act.)</th>
<th>Example</th>
<th>Freq.</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel</td>
<td>/z/ (3,425)</td>
<td>/s/</td>
<td>does</td>
<td>2,935</td>
</tr>
<tr>
<td></td>
<td>/θ/ (1,434)</td>
<td>/t/</td>
<td>love</td>
<td>305</td>
</tr>
<tr>
<td></td>
<td>/θ/ (3,488)</td>
<td>/θ/</td>
<td>this</td>
<td>3,235</td>
</tr>
<tr>
<td></td>
<td>/s/ (613)</td>
<td>/s/</td>
<td>thing</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td>/t/ (613)</td>
<td>/t/</td>
<td>thank</td>
<td>331</td>
</tr>
<tr>
<td>Place</td>
<td>/θ/ (3,488)</td>
<td>/θ/</td>
<td>this</td>
<td>3,235</td>
</tr>
<tr>
<td></td>
<td>/s/ (613)</td>
<td>/s/</td>
<td>thing</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td>/θ/ (1,434)</td>
<td>/b/</td>
<td>give</td>
<td>766</td>
</tr>
</tbody>
</table>

In consonants, salient mispronunciation patterns are classified into voicing, place, and manner at articulatory level. As presented in Table 2, in terms of voicing, learners show devoicing patterns such as /z/ → /s/ and /θ/ → /θ/. Devoicing on /z/ mainly occurs at word final, while devoicing on /θ/ is mostly caused by regressive assimilation. In terms of place of articulation, salient mispronunciation pattern is shown as a variation from dental to alveolar. Finally, in terms of manner of articulation, Korean learners have difficulties of producing English fricatives. These mispronunciation patterns are the cases where the learners failed to produce fricatives which do not exist in native language and substituted them with their corresponding stops.

Table 3: Salient mispronunciation patterns in vowels in terms of articulatory level

<table>
<thead>
<tr>
<th>Level</th>
<th>Canon. (Act.)</th>
<th>Example</th>
<th>Freq.</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round</td>
<td>/a/ (2,381)</td>
<td>/oU/</td>
<td>project</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td>/a/ (2,381)</td>
<td>/oU/</td>
<td>/prədʒekt/ → /prədʒekt/</td>
<td>12.39%</td>
</tr>
<tr>
<td></td>
<td>/s/ (6,317)</td>
<td>/l/</td>
<td>law</td>
<td>735</td>
</tr>
<tr>
<td></td>
<td>/s/ (6,317)</td>
<td>/a/</td>
<td>another</td>
<td>1,106</td>
</tr>
<tr>
<td></td>
<td>/s/ (6,317)</td>
<td>/æ/</td>
<td>/æn/</td>
<td>1,030</td>
</tr>
<tr>
<td></td>
<td>/t/ (613)</td>
<td>/t/</td>
<td>Helen</td>
<td>654</td>
</tr>
</tbody>
</table>

In vowels, salient mispronunciation patterns are classified into rounding, height, and backness at articulatory level. As presented in Table 3, in terms of rounding, learners tend to realize unrounded /a/ as rounded vowel /oU/. In terms of height, there are two variations, one is raising from low to mid vowels, and the other is lowering from mid to low vowels. In backness, all salient patterns occur as fronting from back to front vowels. Substitution of /s/ for /oU/ occurs because /s/ does not exist in native language and learners tend to replace it with the most similar phoneme /oU/ which exists in L1. The remaining mispronunciation patterns in Table 3 are explained as orthographic interference, in which inappropriate inference from the spelling influences the learners’ pronunciation [25]. For instance, the learners substituted /oU/ for /æ/, as in /prədʒekt/ for the word ‘project’ /prədʒekt/, where the variation is influenced from the grapheme ‘o’ for /æ/.

5. Experiments

5.1. Experimental setup

Based on the quantitative analysis of salient mispronunciation patterns in Section 4, we perform an articulatory diagnosis experiment on the seven salient phones. As can be seen in Section 4, the number of the correctly pronounced phone is much larger than the number of mispronounced, which could make trained models biased. To prevent the bias problem, we adopt other phones’ correctly pronounced observations as mispro-
nounced samples of the target phone as much as the difference between the number of correct instances and the number of incorrect instances to make a balance [5].

We split balanced observations of each phone into the training and the test set with a ratio of 8:2. Furthermore, 20% of the training set is assigned as a validation set to determine hyperparameters of FFNN, such as the number of layers and the number of nodes. We do not include artificially augmented instances as mentioned above in the test set, maintaining the balance between the number of correct and incorrect instances. The details of the training, validation, and test sets are presented in Table 4.

Table 4: Details of training, validation, and test sets in terms of salient phones

<table>
<thead>
<tr>
<th>Cat. Level</th>
<th>Phone</th>
<th>Training (Validation)</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>consonant</td>
<td>/z/</td>
<td>14,685 (2,937)</td>
<td>3,671</td>
<td>18,356</td>
</tr>
<tr>
<td></td>
<td>/s/</td>
<td>18,367 (3,673)</td>
<td>4,591</td>
<td>22,958</td>
</tr>
<tr>
<td></td>
<td>/s/</td>
<td>4,447 (889)</td>
<td>1,111</td>
<td>5,558</td>
</tr>
<tr>
<td></td>
<td>/v/</td>
<td>12,893 (2,578)</td>
<td>3,223</td>
<td>16,116</td>
</tr>
<tr>
<td>vowel</td>
<td>/a/</td>
<td>14,314 (2,862)</td>
<td>3,578</td>
<td>17,892</td>
</tr>
<tr>
<td></td>
<td>/o/</td>
<td>13,624 (2,724)</td>
<td>3,405</td>
<td>17,029</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>61,077 (12,215)</td>
<td>15,269</td>
<td>76,346</td>
</tr>
</tbody>
</table>

5.2. Experimental results

Table 5 shows the performance of mispronunciation diagnosis for salient consonants in terms of articulatory levels for the test set. By the results, mispronunciation diagnosis at articulatory level using articulatory features shows more than 70% accuracy and 0.75 F1 scores for all articulatory levels in average. The results denote that the proposed method using articulatory features is effective for articulatory diagnosis for consonants.

Although the proposed method shows high accuracy in average at articulatory levels, the articulatory level of place for /O/ and /B/ presents slightly lower performance than average. These phones are the only dental sounds at the articulatory level of place. Since they are inter-dental fricative, they all have relatively small amount of amplitude, which make them difficult for them to distinguish mispronunciation. These factors appear to affect the performance.

Table 5: Diagnosis performance for salient consonants in terms of articulatory levels

<table>
<thead>
<tr>
<th>Phone</th>
<th>Level</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>/z/</td>
<td>voice</td>
<td>70.14%</td>
<td>0.683</td>
<td>0.890</td>
<td>0.773</td>
</tr>
<tr>
<td></td>
<td>place</td>
<td>85.57%</td>
<td>0.857</td>
<td>0.877</td>
<td>0.867</td>
</tr>
<tr>
<td></td>
<td>manner</td>
<td>79.38%</td>
<td>0.821</td>
<td>0.825</td>
<td>0.823</td>
</tr>
<tr>
<td>/s/</td>
<td>voice</td>
<td>83.60%</td>
<td>0.837</td>
<td>0.898</td>
<td>0.866</td>
</tr>
<tr>
<td></td>
<td>place</td>
<td>60.50%</td>
<td>0.623</td>
<td>0.670</td>
<td>0.646</td>
</tr>
<tr>
<td></td>
<td>manner</td>
<td>62.13%</td>
<td>0.632</td>
<td>0.852</td>
<td>0.726</td>
</tr>
<tr>
<td>/s/</td>
<td>voice</td>
<td>79.68%</td>
<td>0.814</td>
<td>0.857</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td>place</td>
<td>65.83%</td>
<td>0.672</td>
<td>0.697</td>
<td>0.684</td>
</tr>
<tr>
<td></td>
<td>manner</td>
<td>71.76%</td>
<td>0.761</td>
<td>0.830</td>
<td>0.794</td>
</tr>
<tr>
<td>/v/</td>
<td>voice</td>
<td>80.18%</td>
<td>0.821</td>
<td>0.859</td>
<td>0.840</td>
</tr>
<tr>
<td></td>
<td>place</td>
<td>75.43%</td>
<td>0.795</td>
<td>0.842</td>
<td>0.818</td>
</tr>
<tr>
<td></td>
<td>manner</td>
<td>71.40%</td>
<td>0.751</td>
<td>0.815</td>
<td>0.782</td>
</tr>
<tr>
<td>average</td>
<td>voice</td>
<td>78.40%</td>
<td>0.789</td>
<td>0.876</td>
<td>0.828</td>
</tr>
<tr>
<td></td>
<td>place</td>
<td>71.83%</td>
<td>0.737</td>
<td>0.772</td>
<td>0.754</td>
</tr>
<tr>
<td></td>
<td>manner</td>
<td>71.17%</td>
<td>0.741</td>
<td>0.831</td>
<td>0.781</td>
</tr>
</tbody>
</table>

Table 6 presents a performance of mispronunciation diagnosis for salient vowels in terms of articulatory levels for the test set. As well as consonants, vowels also show high average articulatory diagnosis performance, except in height. Other levels have more than 70% accuracy, but height shows 65% accuracy in average. In the case of vowels, the training sets contain variations from salient phones to diphthongs, such as /aI/→/aI/. Articulatory characteristics of diphthongs vary during pronunciation of a vowel. For example, the tongue in /aI/ moves from low to high in terms of height. Also, /aI/ shows changes of articulation in height (mid→high) and backness (back→front). Thus, such drastic changes in the articulation of a vowel could affect low performance.

Table 6: Diagnosis performance for salient vowels in terms of articulatory levels

<table>
<thead>
<tr>
<th>Phone</th>
<th>Level</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>rounding</td>
<td>83.79%</td>
<td>0.839</td>
<td>0.888</td>
<td>0.863</td>
</tr>
<tr>
<td></td>
<td>height</td>
<td>59.60%</td>
<td>0.658</td>
<td>0.853</td>
<td>0.743</td>
</tr>
<tr>
<td></td>
<td>backness</td>
<td>80.94%</td>
<td>0.811</td>
<td>0.896</td>
<td>0.851</td>
</tr>
<tr>
<td>/o/</td>
<td>rounding</td>
<td>70.43%</td>
<td>0.731</td>
<td>0.855</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>height</td>
<td>70.93%</td>
<td>0.759</td>
<td>0.898</td>
<td>0.823</td>
</tr>
<tr>
<td></td>
<td>backness</td>
<td>75.98%</td>
<td>0.753</td>
<td>0.887</td>
<td>0.815</td>
</tr>
<tr>
<td>/u/</td>
<td>rounding</td>
<td>88.25%</td>
<td>0.883</td>
<td>0.865</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td>height</td>
<td>65.16%</td>
<td>0.693</td>
<td>0.887</td>
<td>0.778</td>
</tr>
<tr>
<td></td>
<td>backness</td>
<td>65.59%</td>
<td>0.761</td>
<td>0.859</td>
<td>0.807</td>
</tr>
<tr>
<td>average</td>
<td>rounding</td>
<td>80.82%</td>
<td>0.818</td>
<td>0.869</td>
<td>0.843</td>
</tr>
<tr>
<td></td>
<td>height</td>
<td>65.23%</td>
<td>0.703</td>
<td>0.879</td>
<td>0.782</td>
</tr>
<tr>
<td></td>
<td>backness</td>
<td>74.17%</td>
<td>0.775</td>
<td>0.881</td>
<td>0.824</td>
</tr>
</tbody>
</table>

6. Conclusion

In this paper, we proposed a method to provide an articulatory diagnosis of English produced by Korean learners using articulatory Goodness-Of-Pronunciation (aGOP) features, which are based on the distinctive feature theory in phonology. So far, previous studies regarding mispronunciation diagnosis have limitations that they have carried out diagnosis at phone level. To provide effective corrective feedback, mispronunciation diagnosis had better be performed at the articulatory level, rather than at phone level. We applied different models of articulatory diagnosis depending on the consonants and vowels considering articulatory characteristics. Mispronunciation diagnosis for consonants was conducted in three articulatory levels: voicing, place of articulation, and manner of articulation. On the other hand, articulatory diagnosis for vowels was performed in terms of rounding, height, and backness. Furthermore, we quantitatively analyzed salient mispronunciation patterns in English produced by Korean learners using corpus-based analysis. By the results, the proposed method for articulatory diagnosis presented more than 70% accuracy and 0.75 of F1 scores in average for all articulatory levels except height in vowels. It is noteworthy that these results indicate the proposed method yields effective mispronunciation diagnosis at articulatory level.

However, there is a limitation that the proposed method only decides that the pronunciation is correct or not at the articulatory level and does not provide corrective feedback on how to correct the pronunciation at articulatory level. Thus, in future work, we will extend the experiment to provide corrective feedback at articulatory level as well as mispronunciation diagnosis.
7. Acknowledgements

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-1600-00464) supervised by the IITP (Institute for Information & communications Technology Promotion).

8. References


Overview of the 2017 Spoken CALL Shared Task
Claudia Baur1, Cathy Chua4, Johanna Gerlach1
Manny Rayner1, Martin Russell2, Helmer Strik3, Xizi Wei2
1 FTI/TIM, University of Geneva, Switzerland
2 Department of Electronic, Electrical and Systems Engineering, University of Birmingham
3 Centre for Language Studies (CLS), Radboud University Nijmegen
4 Independent researcher
Claudia.Baur@unige.ch, Johanna.Gerlach@unige.ch, Emmanuel.Rayner@unige.ch, XXW395@student.bham.ac.uk, m.j.russell@bham.ac.uk, w.strik@let.ru.nl

Abstract
We present an overview of the shared task for spoken CALL. Groups competed on a prompt-response task using English-language data collected, through an online CALL game, from Swiss German teens in their second and third years of learning English. Each item consists of a written German prompt and an audio file containing a spoken response. The task is to accept linguistically correct responses and reject linguistically incorrect ones, with “linguistically correct” being defined by a gold standard derived from human annotations; scoring was performed using a metric defined as the ratio of the relative rejection rates on incorrect and correct responses. The task received twenty entries from nine different groups. We present the task itself, the results, a tentative analysis of what makes items challenging, a comparison between different metrics, and suggestions for a continuation.

Index Terms: CALL, shared tasks, speech recognition, metrics

1. Introduction
There are now many episodes in the history of human language technology showing that the introduction of a shared task1 has had a positive effect on a particular area. A prominent series of examples are the various tasks based on the Wall Street Journal corpus, including speech recognition [1], parsing [2] and several types of semantic analysis [3]. Perhaps even more importantly, work on machine learning during the 21st century has to a considerable extent been driven by the handwritten digit recognition task [4]. Other well-known examples of shared tasks include ATIS in the early 90s [5], which had a strong effect on interactive spoken language systems; the Named Entity Recognition task [6], which similarly influenced work on information extraction; and the Recognizing Textual Entailment task [7], which has influenced work on question answering.

In all these cases, introduction of the shared task created a new community with frequent productive interactions between many groups, and substantially advanced a whole subfield inside the space of a few years. The sociology of the process has become familiar to many researchers. A shared task forces each group to look closely at what other groups are doing, and in particular to study methods which are achieving high scores in the competitions. It encourages development of a common vocabulary of concepts. Above all, it introduces widely accepted evaluation procedures and metrics that permit objective comparisons, both between systems developed by different groups and between different versions of single systems. It is easier to achieve progress when people agree on what “progress” consists of, and how it can be measured.

As the series of ‘Speech and Language Technology in Education’ (SLaTE) workshops2 attests, speech recognition for CALL has become an established field. At the 2015 workshop in Leipzig, the authors of the present paper suggested that the time might have arrived to define a shared task for this area. The response was positive enough that we presented a paper at the LREC conference in May 2016 [8] with a concrete definition of a possible task, and released training data and resources on July 13 of the same year. Test data was released on March 13, 2017, with a deadline of one week to submit results; each competing group was allowed to submit up to three entries. We were pleased to receive a total of twenty submissions from nine different groups.

In the rest of this paper, we present an overview of the shared task. §2 describes the task, data, metric and resources. §3 presents the results. In the following two sections we present some analysis: §4 considers what the data tells us about the issues that appear to make test items easier or harder, and §5 discusses the adequacy of the scoring metrics used. In conclusion, §6 makes suggestions about what to do next and §7 briefly discusses ethical issues.

2. Task, data, resources and metric
One of the most common types of spoken CALL exercise is prompt-response: the system gives the student a prompt, the student responds, and the system either accepts or rejects the response, possibly giving some extra feedback. The prompt can be of various forms, including L2 text (“read the following sentence”), L1 text (“translate the following sentence into the L2”), multimedia (“name this object”), or some kind of combination. Prompt-response exercises are for example used heavily in the popular Duolingo application.3

We proposed a spoken prompt-response task using data collected from an English course developed for German-speaking Swiss teenagers doing their first to third year of English [9, 10]. The course runs on CALL-SLT [11], a spoken CALL platform which has been under development at Geneva University since 2009.4 It is based on a textbook commonly used in German-speaking Switzerland [12] and consists of eight lessons: (1) at

1Another common term is “competitive-collaborative task”.
2http://hstrik.ruhosting.nl/slate/
3https://www.duolingo.com/
4http://callslt.unige.ch/demos-and-resources/
the train station, (2) getting to know someone, (3) at the tube sta-
tion, (4) at the hotel, (5) shopping for clothes, (6) at the restaur-
ant, (7) at the tourist information office, and (8) asking/giving
directions. Each lesson offers an interactive dialogue permitting
many variations, which allows the students to practise their oral
conversational skills. The emphasis is on a communicative ap-
proach to second language acquisition, putting more weight on
achieving a successful interaction than on minor grammatical
or pronunciation flaws in the utterances.

Each prompt in the course is a combination of a multimedia
file in the L2 (English) and a written text instruction in the L1
(German). To give a typical example, the system plays a short
animated clip with an English native speaker asking the question,
“How many nights would you like to stay at our hotel?” and
simultaneously displays the German text, “Frag: Zimmer
für 3 Nächte” (Ask: room for 3 nights). The text indicates how
the student is supposed to answer in the L2. In this case, an
acceptable response would be something like “I want a room
for three nights”. “Do you have a room for three nights?” or
“I would like to stay for three nights”. The intention is that a
reasonably wide variety of grammatically and linguistically cor-
rect utterances are accepted, as long as they correspond to the
meaning of the German prompt, so the student is able to prac-
tise spontaneous generative language skills. A response can be
rejected for a variety of reasons, including incorrect use of vo-
cabulary, grammatical incorrectness, incorrect use of the user
interface, bad pronunciation, bad recognition due to insufficient
recording quality, etc.

Once the student has responded, by speaking into the head-
set or onboard mic, the system performs speech recognition
and then matches the recognised utterance against the prompt’s
specification of what should be counted as a correct answer. If
there is a match, the system gives positive feedback by display-
ing a green frame around the text prompt, and moves on to the
next dialogue state. If the utterance is rejected, a red frame
(negative feedback) is shown and the student is asked to repeat
or reformulate their response. The screenshot in figure 1 illus-
trates the process.

To make the data set more interesting and challenging, short
utterances such as “hello”, “bye”, “yes”, “no” and “thanks”,
which occur very frequently in the corpus and are almost always
well pronounced by the subjects, were removed.

The data was selected so that the sets of speakers in the
test and training portions were disjoint. The test set contained
5,222 utterances and the test corpus 996 utterances. The ut-
terances in the training and test data sets were selected based on
the following criteria with decreasing level of importance: 1)
student’s total number of interactions (grouped into three cate-
gories: <50, 50–200, >200 interactions), 2) pre-placement test
score (out of 41), 3) gender, 4) age (ranging between 12 to 15
years). This methodology allows us to have a representative and
well-balanced selection of interactions in both the training and
test corpora. The two data sets contained utterances from mo-
tivated and less motivated students, from stronger and weaker
students, from both male and female students and from students
with different ages. Table 1 presents data selection details for
both the training and test data sets.

<table>
<thead>
<tr>
<th>Training data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterances</td>
<td>5222</td>
</tr>
<tr>
<td>Speakers</td>
<td>66</td>
</tr>
<tr>
<td>&gt; 200 interactions</td>
<td>19</td>
</tr>
<tr>
<td>50–200 interactions</td>
<td>35</td>
</tr>
<tr>
<td>&lt; 50 interactions</td>
<td>12</td>
</tr>
<tr>
<td>Average pre-placement test score</td>
<td>19.18</td>
</tr>
<tr>
<td>Female</td>
<td>34</td>
</tr>
<tr>
<td>Male</td>
<td>32</td>
</tr>
<tr>
<td>Average age</td>
<td>13.9</td>
</tr>
</tbody>
</table>

Table 1: Data selection criteria for training and test data sets

The core resource for the task was an English speech corpus
collected with the CALL-SLT dialogue game. In total, the
corpus contains 38,771 spontaneous speech acts in the form
of students’ interactions with the dialogue system. The data
was collected in 15 school classes at 7 different schools in Ger-
manophone Switzerland during a series of experiments in 2014
and early 2015 [10]. For the task, we made available an an-
notated subset of the corpus. The training corpus contained
5,222 utterances and the test corpus 996 utterances. The ut-
terances in the training and test data sets were selected based on
the following criteria with decreasing level of importance: 1)
student’s total number of interactions (grouped into three cate-
gories: <50, 50–200, >200 interactions), 2) pre-placement test
score (out of 41), 3) gender, 4) age (ranging between 12 to 15
years). This methodology allows us to have a representative and
well-balanced selection of interactions in both the training and
test corpora. The two data sets contained utterances from mo-
tivated and less motivated students, from stronger and weaker
students, from both male and female students and from students
with different ages. Table 1 presents data selection details for
both the training and test data sets.

To make the data set more interesting and challenging, short
utterances such as “hello”, “bye”, “yes”, “no” and “thanks”,
which occur very frequently in the corpus and are almost always
well pronounced by the subjects, were removed.

The data was selected so that the sets of speakers in the
test and training portions were disjoint. The test set contained
data from 25 speakers, 12 female and 13 male. Table 2 presents
details.

2.2. Annotated data and annotation procedure

The audio data for the training and test sets was released on
the shared task site\textsuperscript{5} together with accompanying metadata
in the form of CSV-formatted spreadsheets, with one line for each

\textsuperscript{5}http://regulus.unige.ch/spokencallsharedtask/
audio file. For the training data, the spreadsheet columns represented the identifier, the prompt, the name of the audio file, the transcription, and judgements (correct/incorrect) for “language” and “meaning”. A judgement of “correct” in the “language” column meant that the audio file was a fully correct response to the prompt. “Incorrect” in the “language” column and “correct” in the “meaning” column meant that it was linguistically incorrect, but semantically correct (cf. Table 3). The metadata for the test set was similar, but the version released to participants omitted the transcription and judgement columns. These were kept secret until after the result submission deadline.

For the training data, we thought we would be able to carry out the annotation process efficiently by splitting it into two phases, respectively for transcription and text judging. In the first phase, audio files were transcribed by native German/Swiss German speakers fluent in English; in the second, each prompt/written response pair was judged for linguistic and semantic correctness by three native speakers of English. We hoped that by adopting this procedure and only retaining items where the judges were unanimous, we would obtain highly reliable judgements; however, careful analysis of a sample of the data convinced us that we had been too optimistic. The annotations contained more noise than we had expected, and perhaps as many as 3–4% of the judgements were incorrect.

For the test data, where this level of noise was clearly unacceptable, we rechecked the judgements after receiving the submissions. Having the submissions available made it possible for us to focus attention on the examples for which the largest number of submissions disagreed with the judgements, implying that the judgements in question were the ones most likely to be incorrect. (Details are presented in the next section). Ordering the examples in this way, by the number of submissions which gave the example the wrong label, we carefully went through the top half of the set. This meant that we checked each example where at least four submissions disagreed with the current judgement. Two of the authors, one an English native speaker and one a German native speaker fluent in English, carefully listened to the examples together and discussed each one until they had reached clear agreement. The principles used to adjudicate borderline examples (in all cases according to the principle “other things being equal”) were as follows:

- Examples with background noise, crosstalk or interrup-

tions were counted as linguistically and semantically correct.
- Disfluencies and repetitions were counted as linguistically and semantically correct.
- Mispronunciations were counted as linguistically and semantically correct, unless the mispronunciation either a) resulted in a different English word meaningful in the given context or b) was incomprehensible.
- Foreign words (usually German/Swiss German but occasionally French) were counted as linguistically incorrect, but semantically correct if they were close enough to the corresponding English word.

### Table 2: Distribution of test data by speaker. “m/f” = male/female, #Utts = number of recorded utterances for speaker in test set.

<table>
<thead>
<tr>
<th>ID</th>
<th>m/f</th>
<th>#Utts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>m</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>m</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>m</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>m</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>f</td>
<td>29</td>
</tr>
<tr>
<td>8</td>
<td>m</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>m</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>m</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>f</td>
<td>32</td>
</tr>
<tr>
<td>12</td>
<td>f</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>f</td>
<td>201</td>
</tr>
</tbody>
</table>

2.3. Resources

In order to make it easier for groups to attempt the proposed task, we provided some other resources: a baseline Kaldi recogniser for the domain, a baseline domain grammar, sets of speech data preprocessed through two domain recognisers, and a baseline Python script that combined the other resources to perform the task. All these materials were made available for free download from the shared task site. We present brief descriptions of these resources below.

2.3.1. Baseline Kaldi recogniser

For people who wanted to experiment with recognition methods, we included acoustic models, language models and scripts for Kaldi [13], a state-of-the-art open-source recogniser platform. Three baseline deep neural network (DNN)-hidden Markov model (HMM) hybrid systems were trained on the training part of WSJCAM0 [14] (WSJCAM0_TR) and 90% of the training corpus of the Spoken CALL Shared Task (90ST), Base_DNN1, Base_DNN2 and Base_DNN3 were trained on 1. WSJCAM0_TR, 2. WSJCAM0_TR plus 90ST, and 3. 90ST, respectively. A Gaussian mixture model (GMM)-HMM system is initially trained to provide the triphone tree and the state level time alignment. The hidden layers of the DNN are unsupervised pre-trained as a stack of restricted Boltzmann machines (RBMs). The output layer is initialized randomly. The network has 4 hidden layers with 1024 neurons in each layer and a softmax output layer with 1516 units. A basic bigram language model, trained on the Shared Task training data, was also included. Recognition experiments on the remaining 10% of the Shared Task data (Table 4), show that Base_DNN1 and Base_DNN2 with WSJCAM0 performed worse than Base_DNN3. This may be because the majority of the training set was from WSJCAM0, and the adult speech from WSJCAM0 is not like the speech from the task. To make the system more like the shared task testing data, we further trained the network from the Base_DNN2 system with 90ST to obtain DNN4. With 13.92% word error rate, DNN4 outperforms the other three baseline systems. After ten cross-validation experiments, this method of adaption was chosen to train our final baseline system for the shared task.

2.3.2. “Pre-recognised” versions of data

For the benefit of groups who only wished to explore the language processing aspects of the task, we processed test and training data through both the baseline Kaldi recogniser and a Nuance Toolkit recogniser with a language model created from the baseline response grammar, and supplied versions of the task metadata which included the recognition results produced.
I would like a room for six nights
I want a room for six nights
I want a room for five nights
It’s raining outside

Table 3: Examples showing use of the “language” and “meaning” judgements in the annotated data. The prompt, “Frag: Zimmer für 6 Nächte” means “Request: room for 6 nights”.

<table>
<thead>
<tr>
<th>system</th>
<th>training data</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base_DNN1</td>
<td>wsjcam0</td>
<td>72.12</td>
</tr>
<tr>
<td>Base_DNN2</td>
<td>wsjcam0+90ST</td>
<td>19.62</td>
</tr>
<tr>
<td>Base_DNN3</td>
<td>90ST</td>
<td>16.61</td>
</tr>
<tr>
<td>DNN4</td>
<td>90ST</td>
<td>13.92</td>
</tr>
</tbody>
</table>

Table 4: Results(% WER) of DNN baseline systems

2.3.3. Baseline response grammar

We also provided a version of the existing CALL-SLT response grammar, which contained 565 prompts with a total of 43,862 possible responses. The grammar was supplied in a minimal XML format, where each item consisted of the original German text prompt, an English translation of the prompt, and a list of possible responses. A typical record from the grammar is shown in Figure 2. We stated clearly that the response grammar was not intended to be exhaustive. The task is open-ended; ideally, the system should accept any grammatically correct, adequately pronounced response which corresponds to the prompt, and the grammar only gives plausible examples of such responses. Since the grammar was automatically derived from the one used to perform the actual data collection, we knew that it gave useful coverage, but it could evidently be improved.

2.3.4. Baseline system

Finally, we supplied a simple Python script which instantiated a minimal example of a system capable of performing the shared task. The script reads a “pre-recognised” set of metadata (§2.3.2) and the baseline response grammar (§2.3.3) and produces a CSV-formatted spreadsheet of accept/reject decisions, labelling a response as “accept” if and only if the recognition result is in the response grammar.

2.4. Metrics

For reasons explored in our 2016 paper, we are not convinced that standard metrics such as the F-measure are appropriate for scoring tasks like the one we proposed here. In particular, one serious objection is that these metrics can give a higher score to a system which always accepts than they do to a system which behaves in a normal way, accepting correct responses more frequently than incorrect ones. Intuitively, a system which always accepts is useless; if a system is useful, it is precisely because it responds differently to correct and incorrect responses. We consequently decided to base our metric directly on the idea of measuring the degree of difference between the system’s behaviour in the case of the two types of response. A second intuition was that false accepts are not all equally serious. A false accept of a response which is linguistically incorrect, but has the right meaning, should be less serious than a false accept of a response which is not even semantically correct. We call these two kinds of false accepts respectively “plain false accepts” and “gross false accepts”.

Slightly adapting the treatment of [15], we define metrics as follows. We assume that we are given a set of annotated prompt/response interactions, where in each case the annotations show whether the response was correct or incorrect, both linguistically and semantically, and whether it was accepted or rejected. We write CA for the number of of correct accepts, CR for the number of correct rejects, FA for the number of plain false accepts, FA2 for the number of gross false accepts and FR for the number of false rejects. We set

\[
FA = FA_1 + k FA_2
\]

for some constant k, weighting gross false accepts k times more heavily than plain false accepts, and

\[
Z = CA + CR + FA + FR
\]

Then we write \(C_A = \frac{C}{F + A}, C_R = \frac{C}{F + R}, F_A = \frac{F}{A}, F_R = \frac{F}{R}\) and define metrics in terms of the four quantities \(C_A, C_R, F_A, F_R\), which total to unity. Looking first at traditional metrics, we consider precision \((P = \frac{C_A}{C_A + F_A})\), recall \((R = \frac{C_R}{C_R + F_R})\), F-measure \(F = \frac{2PR}{P + R}\) and scoring accuracy \((SA = C_A + C_R)\). Scoring accuracy \(SA\) is related to classification error \(E\) by the equation \(SA = 1 - E\), and maximising \(SA\) is equivalent to minimising \(E\).

Generally, all of the above metrics are based on the idea of minimising some kind of error. In contrast, \(D\), the metric based on differential response which we used for the task, is defined as the ratio of the relative correct reject rate (the reject rate on incorrect responses) to the relative false reject rate (the reject rate on correct responses). We put \(RC_R = \frac{C_R}{C_R + F_A}\) and \(RF_R = \frac{F_R}{F_R + C_A}\), then define

\[
D = \frac{RC_R}{RF_R} = \frac{C_R/(C_R + F_A)}{F_R/(F_R + C_A)} = \frac{C_R(F_R + C_A)}{F_R(C_R + F_A)}
\]

We announced when proposing the task that results would be ranked in terms of \(D\) with \(k = 3\), but we also present scores for the other metrics.

3. Results

We received twenty entries from nine different groups. Of these, ten entries (four groups) used the pre-recognised baseline Kaldi data; four entries (three groups) used the pre-recognised baseline Nuance data; and six entries (three groups) used their own recognisers. One group submitted entries both for their own recogniser and for Nuance data. Table 5 present submissions and scores using D and several other metrics, also comparing with three baseline systems built using the resources described in §2.3.
4. What makes items difficult?

In order to study the degree of difficulty of the test items, we combined the results of all 20 submitted entries, with those of the “Nuance” and “Kaldi” baseline systems, making a total of 22 entries. A full table is available on the “Test data” tab of https://regulus.unige.ch/spokencallsharedtask/. Table 6 presents a summary and some representative examples. The greater part of the test set, as can be seen, was quite straightforward; 581 of the 996 items end up in the “Easy” group, and were scored correctly by at least 17 of the 22 entries. (232 were scored correctly by all entries). Another 265 items, the “Intermediate” group, were scored correctly by more than half of the entries. The remaining 150 items were clearly challenging, with 24 being scored correctly by only five or fewer entries.

In order to gain some understanding of a few obvious factors which might make items easier or harder, we performed an annotation of the test data. Three of the authors (a native Swiss German speaker fluent in English, a native English speaker with some German, and a native English speaker with no German) listened to each audio file separately using an online tool and categorized them on the following five scales:

- **Crosstalk** Could you hear anyone other than the student talking? (yes/no)
- **Non-speech noise** Could you hear any non-speech noises, for example background noise, breath noise, etc? (no/weak/strong; “weak” was defined as “clearly softer than the speech” and “strong” as “comparable in loudness with the speech”).
- **Stuttering/repetition** Did the student stutter, repeat themselves, or in some other way clearly change their mind about what they were going to say? (yes/no)
- **Incomprehensible** Was any word spoken by the student incomprehensible to you? (yes/no)
- **Faint** Was the volume of the student’s speech clearly much fainter than usual? (yes/no)

Figure 2: XML reference grammar example.

The categories were chosen as labeling conditions which a) occurred reasonably often in the data, much of which was recorded in noisy environments, b) could reasonably be expected to make the accept/reject decision difficult and c) were easy to judge. Unfortunately, we were not able to include any measure of pronunciation quality; previous experience had convinced us that we would not be able to judge this usefully in the time available. We found very poor inter-annotator agreement on “Non-speech noise”, where about half the data was annotated as “Weak” by each judge. We consequently collapsed “No” and “Weak” together for this scale, making all five scales binary. The judgements from the three annotators were combined using majority voting; Table 7 shows agreement between annotators. We also added a sixth category, out-of-vocabulary (OOV), which was computed automatically; an example in the test data was counted as OOV if at least one word in the transcription failed to occur in either the training data or the baseline response grammar (cf. §2.3.3). Finally, we computed the number of word errors (sum of insertions, deletions and substitutions) for each item, using the recognisers for which we had available data. These were the baseline Kaldi and Nuance recognisers, and the recogniser for the JJJ entry, whose author submitted them to us to be made publicly available on the shared task site. Table 8 shows the distribution of the resulting metrics over the different bands from Table 6.

We draw the following tentative conclusion from this data. We see that for all 6 measures, the percentages are lower for the easy cases (0–5 labelling errors). The relative differences are smaller for Faint, and larger for OOV and the other four properties of the utterances (CT, NSN, Stut, Inc). Thus it seems that esp. the latter 5 measures have a substantial effect on item difficulty. At the same time, the fact that only 38 percent of the “Difficult” items are marked for any of the categories suggests that other factors might be relevant. Obvious candidates are of course pronunciation related measures, at segmental and/or prosodic level. Second, looking at the word error rate columns, we see that WER for all three recognisers is much lower for the “Easy” group; this supports the commonsense hypothesis that recognition quality is an important factor in determining whether an item is easy or difficult. Conversely, the fact that the WER is only slightly higher in the “Difficult” group than it is in the “Medium” group also suggests that other factors might be involved.

These intuitive impressions are strengthened by carrying out a basic ANOVA analysis. The effect of word error on labelling error is very significant ($p < 10^{-15}$) for the entries using recognisers where WER data is available, but WER still accounts for less than a quarter of the variance in the labelling error for any recogniser. ANOVA also shows strongly significant effect on labelling error from the categories OOV, Crosstalk and Stuttering/Repetition ($p < 10^{-8}, p < 10^{-4}$ and $p < 10^{-3}$ respectively), but these categories account for only a few percent of the variance in the labelling error. This supports the intuitive conclusion that these three categories, while significant, are not central to the task.

<table>
<thead>
<tr>
<th>Id</th>
<th>Prompt</th>
<th>Transcription</th>
<th>language</th>
<th>meaning</th>
<th>#Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>3709</td>
<td>Sag: Ich habe 2 jüngere Brüder</td>
<td>i have two young brothers</td>
<td>incorrect</td>
<td>incorrect</td>
<td>22</td>
</tr>
<tr>
<td>3944</td>
<td>Frage: Hosen</td>
<td>i want pants</td>
<td>correct</td>
<td>correct</td>
<td>20</td>
</tr>
<tr>
<td>4271</td>
<td>Frage: ein T-Shirt</td>
<td>i will like a t-shirt</td>
<td>incorrect</td>
<td>correct</td>
<td>19</td>
</tr>
<tr>
<td>4519</td>
<td>Frage: ein Ticket zum Green Park</td>
<td>can i have a ticket to the green park</td>
<td>incorrect</td>
<td>correct</td>
<td>15</td>
</tr>
<tr>
<td>4080</td>
<td>Frage: mein Steak rare</td>
<td>i want a rare steak</td>
<td>correct</td>
<td>correct</td>
<td>13</td>
</tr>
<tr>
<td>4683</td>
<td>Frage: Gibt es einen Lift?</td>
<td>is there an ascenseur</td>
<td>incorrect</td>
<td>incorrect</td>
<td>11</td>
</tr>
<tr>
<td>4540</td>
<td>Sag: Kann ich mit Kreditkarte bezahlen</td>
<td>i like to pay with credit card</td>
<td>incorrect</td>
<td>correct</td>
<td>9</td>
</tr>
<tr>
<td>4155</td>
<td>Frage: Wo kann ich ein Shampoo kaufen?</td>
<td>where i can buy shampoo</td>
<td>incorrect</td>
<td>correct</td>
<td>6</td>
</tr>
<tr>
<td>3877</td>
<td>Frage: Erbsen</td>
<td>i would like some peas</td>
<td>correct</td>
<td>correct</td>
<td>6</td>
</tr>
<tr>
<td>4679</td>
<td>Frage: Doppelzimmer</td>
<td>can i have a double room</td>
<td>incorrect</td>
<td>correct</td>
<td>4</td>
</tr>
<tr>
<td>4419</td>
<td>Frage: ein Ticket für Mamma Mia</td>
<td>i want one ticket for mamma mia</td>
<td>correct</td>
<td>correct</td>
<td>2</td>
</tr>
<tr>
<td>4085</td>
<td>Frage: Ich möchte die Dessertkarte</td>
<td>i would like the dessert card</td>
<td>incorrect</td>
<td>incorrect</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6: Examples of prompt/response pairs with different levels of difficulty, estimated by the number of entries assigning the wrong label. “Id” = identifier of utterance in Shared Task test set (clickable link to audio file), “Prompt” = German language prompt, “Transcription” = transcription of student response, “language” = judgement of responses’s correctness in terms of both language and meaning, “meaning” = judgement of response’s correctness only in terms of meaning, “#Bad” = number of submissions out of 22 making incorrect decision.

Table 7: Results of annotation on test set by three judges: proportion of data agreeing and Light’s $κ$ scores.

5. Metrics

We calculated the correlation between metrics using Kendall’s $τ$, a common statistic for measuring similarity of ordinal sequences. Two metrics have a Kendall’s $τ$ of 1 on a set if they put the elements of the set in the same order. −1 if they put them in reverse order. Table 9 summarises the results.
First of all, Precision and Recall correlate negatively with each other; which isn’t surprising, as there is in general a trade-off between these two metrics. Below we will focus on the three remaining metrics, the “main metrics”: D, SA, and F.

We find that the D-metric used for the shared task correlates well with the scoring accuracy (hence also with the simple error rate, since the scoring accuracy is 1 minus the error rate), but rather less well with the F-measure. Interestingly, of the three main metrics, for SA the highest correlations with the other two are observed, while the correlation of D and F are lower.

Looking at Table 5, it is comforting to see that of all submitted twenty entries, the same entry, KKK, was best according to all three of the main metrics. We are however struck by the fact that the baseline “perfect-recognition” system slightly outperformed KKK on F-score, but not on D-score. We are not certain what interpretation to put on these results, and welcome discussion.


<table>
<thead>
<tr>
<th></th>
<th>D</th>
<th>SA</th>
<th>F</th>
<th>Rec</th>
<th>Pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>*</td>
<td>0.731</td>
<td>0.557</td>
<td>0.560</td>
<td>-0.036</td>
</tr>
<tr>
<td>SA</td>
<td>0.731</td>
<td>*</td>
<td>0.763</td>
<td>0.217</td>
<td>0.091</td>
</tr>
<tr>
<td>F</td>
<td>0.557</td>
<td>0.763</td>
<td>*</td>
<td>-0.004</td>
<td>0.328</td>
</tr>
<tr>
<td>Rec</td>
<td>0.360</td>
<td>0.217</td>
<td>-0.004</td>
<td>*</td>
<td>-0.676</td>
</tr>
<tr>
<td>Pr</td>
<td>-0.036</td>
<td>0.091</td>
<td>0.328</td>
<td>-0.676</td>
<td>*</td>
</tr>
</tbody>
</table>

### Table 6: Searching for missing factors

The quality of pronunciation (including suprasegmental/prosodic aspects) is intuitively a very important factor when determining the extent to which an item is challenging for a CALL system to handle. The results from §3 are compatible with this hypothesis; the factors we examined only account for a small proportion of the variance in the labelling error, so some large influence must be missing. It would however be very good to have some kind of direct evidence that the missing factors are related to pronunciation, and attempt to quantify their contribution. Groups with experience in phonetic transcription may find this an interesting topic for further investigation.

### 6.3. A continuation of the shared task

Finally, it is natural to think about possible continuations of the shared task itself. Here we consider four options: a) nothing, this was a one-off; b) the same task again, with new test data; c) use the same data, but define a more challenging task; d) a completely new task. Our suggestion is that we start discussing this at the SLaTE 2017 workshop, and e.g. make democratic decisions about it.

### 6.3.1. No continuation

Given the substantial number of entries for the task, it seems logical to follow it up. Of course, this entirely depends on continued interest. We will try to make an inventory of parties interested in a continuation, starting by asking people present at the SLaTE 2017 workshop.
be the following:

- Interspeech 2018 (Hyderabad, India)
- LREC 2018 (Miyazaki, Japan)
- Building Educational Applications 2018 (location to be announced)

Another option might be a SLaTE workshop in 2019, which could be a satellite of Interspeech 2019 in Graz.

6.3.3. More challenging version of current task

Ideally, it would be desirable to redefine the task in some way so that pronunciation quality could be taken into account. The practical problem is that this would appear to require phonetic annotation of both the training and the test data, i.e. a minimum of 6,000 utterances.

The Geneva group, who have so far taken responsibility for data annotation, do not have resources to do this, so another partner would need to get involved.

6.3.4. New task

The current task is perhaps a little too easy (in particular, its vocabulary is only about 450 words), and there is an argument for moving to a more challenging one. This would require relevant data, including the required annotations. If you have ideas and/or data for such a new task, please contact us.

7. Ethical considerations

The ethics of shared tasks has recently been the subject of some attention. We briefly address the issues raised by Escartin and her colleagues [16], considering conflicts of interest, anonymity, gaming of the system and the balance between competitiveness and collaboration.

To start with the more obvious items, there was a potential conflict of interest in that one of the competing groups (Birmingham) was also involved in organising the task. To avoid any possibility of giving the Birmingham group an unfair advantage, we processed all the data at the Geneva site, only making it available to Birmingham according to the normal schedule. With regard to anonymity, all the results have been presented under anonymised IDs. Each group has been given the key to their own results, so they are free to choose whether to stay anonymous or reveal their identity. We were concerned at all times that pronunciation quality could be taken into account.

With regard to anonymity, all the results have been presented under anonymised IDs. Each group has been given the key to their own results, so they are free to choose whether to stay anonymous or reveal their identity. We were concerned at all times that pronunciation quality could be taken into account.

8. Acknowledgements

Work on the shared task at the University of Geneva was partially funded by the Swiss National Science Foundation (SNSF) under project 105219_153278. We would like to thank Karen Fort for helpful comments on ethical issues.

9. References


Spoken CALL Shared Task system description

Andrew Caines
Automated Language Teaching & Assessment Institute, Department of Theoretical & Applied Linguistics, University of Cambridge, U.K.
apc38@cam.ac.uk

Abstract

We describe the systems we applied to the Spoken CALL Shared Task text-processing track in which an ‘accept’ or ‘reject’ label had to be applied to transcribed responses to a given ‘prompt’ stimulus. The data come from a language tutoring system for Swiss-German learners of English. In our system we attempt to capture the grammaticality and semantic dimensions of language assessment: dimensions which are explicitly labelled in the training corpus. With the training data we achieved accuracies and differential scores well above the baseline. However, our system performed less well on the test data and we discuss ways to improve performance.

Index Terms: language assessment, spoken CALL, shared task

1. Introduction

We present a system description for our submission to the Spoken CALL Shared Task, a special session at SLaTe 2017. The rationale for proposing the shared task, described in [1], was to bring together groups working on computer-assisted language learning (CALL) systems for speech, to demonstrate that the field has reached a level of maturity, and to provide a relevant and useful way of benchmarking progress in spoken CALL.

We describe our approach to the text-processing task (the one with speech transcriptions provided), explain methodological decisions, and assess system weaknesses which need to be addressed. We made two entries to the competition: one using generic features only, the other drawing on a task-specific grammar as well, named Caines_GENERIC and Caines_SPECIFIC respectively.

Results of the competition indicate that our entries ranked low, having performed reasonably well in terms of precision and recall, but being heavily penalised in the key evaluation measure for ‘gross false accepts’ – wrongly accepting utterances in which meaning was judged to be incorrect. We discuss ways to deal with this problem, but nevertheless propose this is a useful system design, since all features were extracted using well-established methods and open-source tools.

2. Related Work

This research relates to work on methods of linguistic feature extraction for assessment and error detection in learner language, on which there has been a multitude of work (e.g. [2, 3, 4]). Much of the literature attends to written inputs – whether that be learner essays or transcripts of speech – and the set of 5222 files in the training data are the subset for which all three judges were in agreement, from a total of 6000 files judged.

Prompts are all of the form, Frag:... (‘say’) or Sag:... (‘say’), and translations were provided in the referenceGrammar.xml file, with allowances for justifiable variation in response. In Table 1, for example, the three prompts mean respectively: ‘ask for red boots’, ‘ask how much it costs’, ‘ask for (bill | check)’. CALL-SLT users then had to do as instructed, carrying out the speech act in English in a fully constructed way.

Language and meaning judgements were obtained by the shared task organisers from three native speakers of English, and the set of 5222 files in the training data are the subset for which all three judges were in agreement, from a total of 6000 files judged.

Test data was in a similar form, consisting of 996 files with an identifier, prompt, soundfile name and speech recognition output for each one (Table 2).

[2] In language teaching and testing, a prompt is the stimulus to which the learner must respond. It often takes the form of a question, but could alternatively be an image, audio recording, video recording, etc.
Table 1: Spoken CALL text-processing task: training data format

<table>
<thead>
<tr>
<th>Id</th>
<th>Prompt</th>
<th>Wavfile</th>
<th>RecResult</th>
<th>Transcription</th>
<th>language</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>11336</td>
<td>Frag: rote Stiefel</td>
<td>11336.wav</td>
<td>i’d like red boots</td>
<td>i’d like red boots</td>
<td>correct</td>
<td>correct</td>
</tr>
<tr>
<td>7068</td>
<td>Frag: Wie viel kostet es?</td>
<td>7068.wav</td>
<td>how many is it</td>
<td>how many is it</td>
<td>incorrect</td>
<td>incorrect</td>
</tr>
<tr>
<td>8774</td>
<td>Frag: Ich möchte die Rechnung</td>
<td>8774.wav</td>
<td>i want the bill</td>
<td>i want the bills</td>
<td>incorrect</td>
<td>correct</td>
</tr>
</tbody>
</table>

Table 2: Spoken CALL text-processing task: test data format

<table>
<thead>
<tr>
<th>Id</th>
<th>Prompt</th>
<th>Wavfile</th>
<th>RecResult</th>
</tr>
</thead>
<tbody>
<tr>
<td>11336</td>
<td>Frag: rote Stiefel</td>
<td>11336.wav</td>
<td>i’d like red boots</td>
</tr>
<tr>
<td>7068</td>
<td>Frag: Wie viel kostet es?</td>
<td>7068.wav</td>
<td>how many is it</td>
</tr>
<tr>
<td>8774</td>
<td>Frag: Ich möchte die Rechnung</td>
<td>8774.wav</td>
<td>i want the bill</td>
</tr>
</tbody>
</table>

The challenge faced by participants in the shared task was to add a column to the test metadata (e.g. Table 2) indicating whether the file should be accepted or rejected, taking the given prompt and response into account. In this respect the language and meaning dimensions, distinct in training, were flattened into a single binary decision. Nevertheless the distinct dimensions played an important role in evaluation.

3.1. Evaluation

The metric for evaluation submissions to the shared task is a differential response score (D) calculated as follows:

- **CORRECT REJECT**: the student’s answer is incorrect; the system rejects (CR).
- **CORRECT ACCEPT**: the student’s answer is correct; the system accepts (CA).
- **FALSE REJECT**: the student’s answer is correct; the system rejects (FR).
- **PLAIN FALSE ACCEPT**: the student’s answer is correct in meaning but grammatically incorrect. The system accepts (PFA).
- **GROSS FALSE ACCEPT**: the student’s answer is incorrect in meaning; the system accepts (GFA).

Overall false accepts (FA) are then calculated as the sum of PFA and GFA, in which gross false accepts are heavily penalised by a weighting factor $k$ where $k = 3$. The intuition here is that a CALL system accepting an utterance which does not make sense is a graver error than a CALL system accepting an ungrammatical utterance.

Finally, the $D$-score is the ratio of the reject rate on incorrect answers to the reject rate on correct utterances:

$$D = \frac{(CR/(CR + FA))}{(FR/(FR + CA))} = \frac{CR(FR + CA)}{FR(CR + FA)}$$  \hspace{1cm} (1)

As shown in 1, the higher the $D$-score the better, as a high proportion of true rejections represents the numerator, which would ideally be divided by a small proportion of false rejections – the denominator.

4. System Description

In our approach to the shared task we had two overarching aims in mind: to do as well as possible in the task, and to build a system which would generalise to other spoken CALL scenarios beyond this task. While we prioritised the latter, since we’re involved in a long-term spoken CALL project as part of the ALTA Institute\(^3\), we of course still wanted to successfully address the given task. Thus we made two submissions both fundamentally the same but with one key difference: use of the referenceGrammar.xml resource, which lists permissible responses to each prompt.

We refer to our general system as Caines\_GENERAL and the system referring to the grammar specific to this task as Caines\_SPECIFIC. We expect the task-specific system to perform better than the general one, but the general system remains of interest to us in the big picture.

We begin with the fundamentals shared by both Caines\_GENERAL and Caines\_SPECIFIC:

- **We preprocessed** the training metadata (text-processing track, Nuance version) so that for each input file we had a translation of the prompt, retrieved from referenceGrammar.xml and an overall judgement label based on the language and meaning ‘correct/incorrect’ labels\(^4\), in this way folding the two assessment criteria into one as required by the task.

- **For feature extraction** we tried to encapsulate the intuitions underlying the language and meaning ratings – that a good answer should be grammatical and occupying a similar semantic space to the prompt. To capture the language dimension we did the following:

  1. Obtained a ‘perplexity’ score for each input transcription: the inverse probability of the response, normalised by length, according to a language model trained on the 10-million word spoken section of the British National Corpus [10]. Perplexity scores were obtained using KenLM [11], and included out-of-vocabulary words.

  2. Obtained the likelihood score assigned by the RASP System [12] to its top-ranked parse analysis of the input transcription, normalised by the number of nodes in the tree (since parse likelihoods are

\(^3\)http://alta.cambridgeenglish.org

\(^4\)If either language or meaning is labelled ‘incorrect’ then the overall label is ‘reject’; else it is ‘accept’.
known to decrease with increasing complexity of tree structure).

3. Obtained a measure of response length by counting the number of tokens in the input transcription. It has been shown in other work that utterance length correlates with assessment scores [13].

- We took several measures to try and capture meaning:

  1. Semantic similarity between prompt and response as the cosine of their word vectors over $N$, where $N$ is the vocabulary of the training corpus (or the test corpus, in the test phase) and word frequencies are expressed as ‘tf-idf’: the product of term frequency (count in given document) and inverse document frequency ($\log\left(\frac{n}{df}\right)$ where $n$ is the number of documents in the corpus, and $df$ is the number of documents in which the given term occurs).

  2. Modified unigram precision of the response compared to the prompt, in a move inspired by the BLEU score used in evaluation of machine translation [7]. This is the proportion of unigrams in the response which are also found in the prompt, ‘clipped’ at the maximum number of times each unigram occurs in the prompt so that pathological responses (e.g. ‘Boots boots boots’) do not score unduly highly.

  3. Another metric borrowed from machine translation is ‘Meteor’ [8], which aligns translation hypotheses to reference translations and calculates a sentence-level similarity score based on exact matches, stemmed matches, synonyms from WordNet\(^7\), and paraphrases. We treat the response as the hypothesis and the prompt as the reference in the input parallel corpus for the Meteor system.

- These features were extracted for both Caines GENERIC and Caines SPECIFIC. Additionally for Caines SPECIFIC, we checked if the learner’s response matched any of the possible answers given in referenceGrammar.xml for the binary feature inGrammar. We did not extend the file in any way, and heed the organisers’ warning that it is incomplete, but nevertheless recognise that if the response is in the grammar it’s a strong indicator of acceptability.

4.1. Implementation

We employed the aforementioned features in a binomial logistic regression model and a support vector machine. Despite many feature combinations and attempts to tune the SVM, we found logistic regression to be more accurate and therefore devote the remainder of this section to the regression model.

We use ten-fold cross-validation to segment the training corpus of 5222 items into tenths (nine segments of 522 items, one of 524), evaluating a regression model on each fold, having trained the model on the other nine-tenths of the corpus. Using R [14] we fitted a generalised linear model (family: binomial, link: logistic) to the training data with judgement as the dependent variable and perplexity, parse likelihood, length, cosine, unigram precision and meteor as the independent variables.

We repeatedly found through inspection of analysis-of-deviance tables that perplexity and unigram precision do not greatly improve the model and have high $p$-values (test: $\chi^2$-squared). For the sake of model parsimony we dropped these features.

4.2. Performance on training data

Therefore with just parse likelihood, length, cosine and meteor we trained a new regression model, Caines GENERIC, which over ten-folds returned a mean accuracy of 77.4%, where accuracy is the proportion of true labels out of all items. However, as explained in §3.1, not all false labels carry the same weight: ‘gross false accepts’, a false accept when the meaning is incorrect, are penalised more heavily (+3) than other false labels.

Table 3: Spoken CALL text-processing task: logistic regression performance on training data (iRej: rejections on incorrect responses, $\frac{CR \cdot PP}{CR + PP}$; cRej: rejections on correct responses, $\frac{FR \cdot CR}{FR + CR}$)

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Caines GENERIC</th>
<th>Caines SPECIFIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR</td>
<td>885</td>
<td>409</td>
<td>914</td>
</tr>
<tr>
<td>PFA</td>
<td>358</td>
<td>663</td>
<td>293</td>
</tr>
<tr>
<td>GFA</td>
<td>99</td>
<td>270</td>
<td>135</td>
</tr>
<tr>
<td>FR</td>
<td>1235</td>
<td>250</td>
<td>635</td>
</tr>
<tr>
<td>CA</td>
<td>2645</td>
<td>3630</td>
<td>3245</td>
</tr>
</tbody>
</table>

The workings for the $D$-score for this first system, Caines GENERIC, are given in Table 3. We also report the performance of a baseline system provided by the organisers – one which seeks out the response in referenceGrammar.xml, performing at 67.5% accuracy on the training data.

We note that our system Caines GENERIC outscores the baseline in large part thanks to a large decrease in false rejects, despite a large increase in false accepts including three times the number of GFAs. Nevertheless the results are encouraging as the method of feature extraction is transferable to other learner corpora where the prompt is known.

With our second system Caines SPECIFIC, we restore the baseline check for a matching response in referenceGrammar.xml. As a result, overall accuracy increases to 79.6% and the $D$-score for the training corpus is reported in Table 3. We see that Caines SPECIFIC only marginally outperforms Caines GENERIC in terms of $D$-score, an improvement which comes in large part from fewer gross and plain false accepts and more correct rejects (though more false rejects). The best-performing system for each evaluation facet is shown in bold in Table 3.

Finally, we trained our two models Caines GENERIC and Caines SPECIFIC on the whole training corpus, using parse likelihood, length, cosine and meteor, plus inGrammar for the task-specific model. Coefficients and $p$-values are given in Tables 4 and 5.

4.3. Performance on test data

It transpires that our systems do not perform as well on the test set as they did on the training set. Table 6 shows precision, recall, $F$-measure (harmonic mean of $p$ and $r$) and differential response score ($D$) for our two submissions, Caines GENERIC

\[^7\text{http://wordnet.princeton.edu}\]
Table 4: Spoken CALL text-processing task: logistic regression feature analysis for Caines_GENERIC, training data

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>e</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>0.397</td>
<td>0.157</td>
<td>0.012</td>
</tr>
<tr>
<td>cosine</td>
<td>2.276</td>
<td>0.157</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>length</td>
<td>-0.284</td>
<td>0.021</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>p.likelihood</td>
<td>0.280</td>
<td>0.067</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>meteor</td>
<td>2.764</td>
<td>0.204</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 5: Spoken CALL text-processing task: logistic regression feature analysis for Caines_SPECIFIC, training data

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>e</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>-0.596</td>
<td>0.181</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>cosine</td>
<td>1.32</td>
<td>0.188</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>length</td>
<td>-0.075</td>
<td>0.023</td>
<td>0.001</td>
</tr>
<tr>
<td>p.likelihood</td>
<td>0.419</td>
<td>0.081</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>meteor</td>
<td>1.003</td>
<td>0.241</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>inGrammar</td>
<td>3.059</td>
<td>0.105</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 6: Spoken CALL text-processing task: logistic regression performance on test data (iRej: rejections on incorrect responses, CR: correct responses, PFA: parent false accept, GFA: grand parent false accept, FR: false rejections, C: check in grammar)

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Caines_GENERIC</th>
<th>Caines_SPECIFIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.822</td>
<td>0.737</td>
<td>0.622</td>
</tr>
<tr>
<td>R</td>
<td>0.723</td>
<td>0.820</td>
<td>0.897</td>
</tr>
<tr>
<td>F</td>
<td>0.770</td>
<td>0.776</td>
<td>0.735</td>
</tr>
<tr>
<td>CR</td>
<td>210</td>
<td>153</td>
<td>68</td>
</tr>
<tr>
<td>PFA</td>
<td>49</td>
<td>86</td>
<td>123</td>
</tr>
<tr>
<td>GFA</td>
<td>21</td>
<td>41</td>
<td>89</td>
</tr>
<tr>
<td>FR</td>
<td>198</td>
<td>129</td>
<td>74</td>
</tr>
<tr>
<td>CA</td>
<td>518</td>
<td>587</td>
<td>642</td>
</tr>
<tr>
<td>iRej</td>
<td>0.652</td>
<td>0.423</td>
<td>0.148</td>
</tr>
<tr>
<td>cRej</td>
<td>0.277</td>
<td>0.180</td>
<td>0.103</td>
</tr>
<tr>
<td>D</td>
<td>2.358</td>
<td>2.346</td>
<td>1.437</td>
</tr>
</tbody>
</table>

Table 7: Spoken CALL text-processing task: logistic regression feature analysis for Caines_GENERIC, test data

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>e</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>0.745</td>
<td>0.310</td>
<td>0.016</td>
</tr>
<tr>
<td>cosine</td>
<td>2.194</td>
<td>0.321</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>length</td>
<td>-0.256</td>
<td>0.046</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>p.likelihood</td>
<td>-0.210</td>
<td>0.148</td>
<td>0.157</td>
</tr>
<tr>
<td>meteor</td>
<td>0.618</td>
<td>0.422</td>
<td>0.143</td>
</tr>
</tbody>
</table>

Table 8: Spoken CALL text-processing task: logistic regression feature analysis for Caines_SPECIFIC, test data

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>e</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>0.285</td>
<td>0.363</td>
<td>0.432</td>
</tr>
<tr>
<td>cosine</td>
<td>2.15</td>
<td>0.324</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>length</td>
<td>-0.234</td>
<td>0.048</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>p.likelihood</td>
<td>-0.222</td>
<td>0.149</td>
<td>0.135</td>
</tr>
<tr>
<td>meteor</td>
<td>0.499</td>
<td>0.428</td>
<td>0.243</td>
</tr>
<tr>
<td>inGrammar</td>
<td>0.487</td>
<td>0.195</td>
<td>0.012</td>
</tr>
</tbody>
</table>

5. Discussion

Since we are penalised three times more heavily for gross false accepts than any other error, there is a clear incentive to reduce their frequency as much as possible. To that end, we inspected the GFAs produced by our task-specific system Caines_SPECIFIC. Table 9 shows a sample of the 135 GFAs in the training corpus.

We found that our measures of semantic similarity and grammaticality are inadequate in several ways. First, they do not distinguish incorrect selection of wh question words, as in items 5841 and 5874. A dictionary would solve this problem, but this is a more task-specific solution than we’d like to adopt: not all spoken CALL systems require the learner to repeat the stimulus so closely.

Secondly, semantic errors of the kind seen in 5848 and 5984, where the learner says ‘three’ instead of ‘two’ could be addressed with richer meaning representations – obtaining a semantic parse rather than distributional or n-gram matching methods. The shallow semantic representation currently obtained also accounts for 5960 and 6147, in which grammatical words are combined in nonsensical ways. A richer semantic parse of both prompt and response would help prevent GFAs.
such as these, perhaps using methods of semantic representation such as those described in [15, 16].

Finally, it is apparent from 5958, 5991, 6002 and 6255 that ungrammatical responses are not being detected adequately. A character-level language model would prevent ‘r t’ being labelled ‘accept’ (5958), and a hard-coded minimum length requirement may be worth investigating if it helps prevent GFAs such as 5991 and 6255. As for 6002, its ungrammaticality would be detected by richer syntactic features – these could be collected at the same time as the RASP likelihoods.

These new insights came after the submission deadline, but we would incorporate them in future work, along with exploration of different machine learning algorithms for the sake of improved performance and increased flexibility over logistic regression. In any case, we found the shared task a useful exercise for scenarios in which the exam or test prompt is known – as it is in our general research programme.

6. Conclusion and Future Work

In future work we may wish to consider the acoustic signal as well as text features. For example we can draw on previous experiments extracting prosodic features such as speech rate, loudness and pitch values [17]. There has also been phonological work on vowel space showing that higher proficiency learners have a more distinctive set of vowels [18]. And we can seek to represent meaning in a more formal way – for instance, using semantic graphs to verify whether the candidate has answered the question [16, 15], or coherence measures to test how well the response holds together [19].

Above all, we have endeavoured to produce an assessment system which is general purpose and uses open-source tools for two main reasons. Firstly on the grounds that it would be time-consuming and costly in less constrained domains to produce an exhaustive grammar of all possible responses as seen in referenceGrammar.xml. Language is known to be infinitely creative, and learners have access to options of expression which cannot be fully itemised. Secondly so that the system is reproducible as a potentially-useful baseline in other work.

7. Acknowledgements

This paper reports on research supported by Cambridge English, University of Cambridge. I’m grateful to colleagues at ALTA-DIAL and the organisers of the Shared Task for their efforts. I also thank the three reviewers for their helpful feedback.

8. References


The CSU-K Rule-Based Pipeline System for Spoken CALL Shared Task

Nico Axtmann, Carolina Mehret, Kay Berkling

Cooperative State University, Karlsruhe (CSU-K), Germany
nico.axtmann95@gmail.com, caro.mehret@gmail.com, berkling@dhbw-karlsruhe.de

Abstract

This paper presents the set-up and results regarding the Cooperative State University submitted system for the shared spoken CALL task. The data was collected from Swiss teenage students using a speech-enabled online tool for English conversation practice. The tasks consisted of training data of a German text prompt with the associated audio file containing an English language response by the students. Problems therefore consisted of recognizing children’s speech with foreign accent, grammatical and vocabulary problems and a number of false starts due to language learning issues.

The task is to create software that will decide whether each response is appropriate (accept) or inappropriate (reject) in the context of the prompt. Results are reported through a single D-value that is computed specifically for this task.

The contribution of this paper is a detailed analysis of a variety of changes to the baseline system (D = 1.69) and the analysis of their contribution to the overall performance. The paper reports the official result (D = 3.21) on the shared task test files but also goes beyond the originally submitted system (D = 4.79).

Index Terms: CALL, speech recognition, ESL

1. Introduction to Shared Task

The work presented in this paper was performed in response to the shared task described in [1]. For the purpose of completion, we will briefly review the basic idea behind the CALL application under study; further details can be found in the above publication.

The exercise to be recognized and scored is of the type prompt-response, where the German-speaking student is prompted to either respond to a request or translate a request or sentence into L2, which is English. The automated system should ideally accept a correct response or reject a student response if faulty and offer relevant support or feedback. There are 565 types of prompts (given as text in German, preceded by a short animated clip in English), namely to make a statement or ask a question regarding a particular item. (1) at the train station, (2) getting to know someone, (3) at the tube station, (4) at the hotel, (5) shopping for clothes (6) at the restaurant, (7) at the tourist information office, (8) asking/giving directions.

A wide range of answers is to be allowed in response, adding to the difficulty of giving automated feedback. Incorrect responses are due to incorrect vocabulary usage, incorrect grammar, or bad pronunciation and quality of recording. The shared task corpus has been annotated with correct transcription and a correct/incorrect tag regarding grammar, vocabulary, pronunciation and fluency.

In designing the automated system it is important to give accurate feedback of correctness without frustrating the student with false negative feedback or letting the student become overly confident by returning too many false positives.

The system’s mistakes can be mitigated by recognizing its own mistakes with low confidence level or giving precise feedback regarding its diagnostic of inaccuracy. However, this goes beyond the scope of the present paper but a rule-based approach is able to support this future application.

The rest of the paper will describe a 2-way decision system that either rejects or accepts the student answer as correct. The baseline system is briefly reviewed in Section 2. Sections 3-5 describe our additions and variations to the baseline system posed by the shared task resulting in the CSU-K system. Section 6 evaluates the CSU-K system given the training and test data in the shared task. Finally, the paper concludes with insights from the task and proposes some future changes to the system.

2. Baseline System Description

2.1. Shared Task Corpus

The data for the shared task was collected in 15 school classes at 7 different schools in the German speaking areas during a series of experiments in 2014 and early 2015. To compare automated system performance, human annotators judge each interaction in order to determine whether or not the utterance should have been accepted by the system. The training corpus contains 5,000 utterances. The testing corpus contains 996 utterances. The selected responses were balanced across gender, age, proficiency and motivation. The data is challenging due to the recording environment in school and the ensuing background noises.

Examples of the data are given in Table 1.

<table>
<thead>
<tr>
<th>ID</th>
<th>Prompt</th>
<th>Transcription</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>4596</td>
<td>Frag: mein Steak</td>
<td>i would like my steak well</td>
<td>I</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>durchgebraten</td>
<td>down</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3709</td>
<td>Sag: Ich habe 2 jüngere Brüder</td>
<td>I have two young brothers</td>
<td>I</td>
<td>I</td>
</tr>
</tbody>
</table>

2.2. Baseline System

A baseline system is provided for the shared task in the form of the speech recognizer and a language model.

2.2.1. Speech Recognition

Acoustic models, language models and scripts for Kaldi [2], a state-of-the-art open-source recognizer platform are provided as a starting point as described in [3]. The provided model is a triphone DNN-HMM model which has been trained on the
training part of WSJCAM0 [4] and the shared task training data. The DNN-HMM model is acquired by adding a softmax layer on to a pretrained DBN structure [5]. The language model is a backed off bigram model trained on the shared task training data.

2.2.2. Grammar
A grammar, automatically derived from the data collection, is provided in XML format for 564 possible prompts with a total of 11,776 possible responses.

2.2.3. Baseline Performance
The training data was split into 90% training and 10% development test data. Each student answer is scored as accept/don’t accept at the semantic and syntactic level and compared against human annotated truth [6]. The baseline system reached an initial WER of 14.81% and a D-Value (Section 6) of 1.694.

3. Improving Speech Recognition
The speech recognizer was changed by adjusting the acoustic model and the language model.

3.1. Changing the Acoustic Model
The baseline DNN-HMM model with a WER of 14.81% was retrained by applying speaker independent transformations Linear Discriminant Analysis and Maximum Likelihood Linear Transformation (LDA+MLLT) on top of the triphone model. LDA+MLLT has shown improvements in recognizing children speech [7]. The model achieved a WER of 13.80%. In the next step Speaker Adaptive Training (SAT) [8] is applied on top of the previous model using Feature-Space Maximum Likelihood Linear Regression (fMLLR) [9] which achieved a WER of 13.41%.

In the next step we experimented with the number Gausssians and leaves in the phone decision tree and retrained the models described above. Results listed in Table 2 show that the best model with 13.05% WER is achieved with 2500/30000 (Leaves/Gausseans) using MLLT+LDA+SAT.

Table 2: WER for different acoustic models.

<table>
<thead>
<tr>
<th>numLeaves/ numGauss</th>
<th>Model</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000/10000</td>
<td>Baseline</td>
<td>14.81%</td>
</tr>
<tr>
<td></td>
<td>+LDA+MLLT</td>
<td>13.80%</td>
</tr>
<tr>
<td></td>
<td>+LDA+MLLT+SAT</td>
<td>13.41%</td>
</tr>
<tr>
<td>2500/30000</td>
<td>Baseline</td>
<td>13.59%</td>
</tr>
<tr>
<td></td>
<td>+LDA+MLLT</td>
<td>13.12%</td>
</tr>
<tr>
<td></td>
<td>+LDA+MLLT+SAT</td>
<td>13.05%</td>
</tr>
</tbody>
</table>

3.2. Changing the Language Model
The language model in the baseline system is a backed-off bigram model trained on the shared task training data. The following steps were used to improve the model.

1. To improve the language model, a new interpolated tri-gram model was trained with the provided shared task training data and the responses in the reference grammar. The SRILM toolkit [10] was used for this purpose.

2. In the next step (+LM) we extended the language model by adding existing responses to the reference grammar with modified words. The following rules were applied in the following order:
   • I am → I’m (2274 additions)
   • a → one (10287 additions)
   • thanks → thank you (11954 additions)
   • I would → I’d (201 additions)
   • one → a (5776 additions)

Both changes are applied to the front end systems from Section 2 and the results are listed in Table 3, showing a general improvement, with the best model resulting in 10.72% WER.

3.3. Adapting the Pronunciation Dictionary
WER can be improved by adjusting the dictionary to include more expected pronunciations [11, 12, 13, 14]. We experimented with phonological rules to extend the pronunciation dictionary with different variants.

The following list shows the adjustments to the pronunciation dictionary according to typical German mispronunciations in English. After applying some of these rules [15] according to what we know about word-structure [16] to the dictionary, new entries were generated and added.

   • dh (beginning) → d
   • dh (beginning) → s
   • v (beginning) → w
   • dh (end) → th
   • d (end) → t
   • g (end) → k
   • b (end) → p
   • z (end) → s

However, as stated in the literature as well, children have limited linguistic knowledge and pronunciation skills. Therefore, the mismatch between regularity of the dictionary extensions and irregularity of pronunciations could not lead to improvements with this global, generative method but instead confused the recognizer further with too many variants. It seems that these adaptations to the user are better addressed in the acoustic space. No improvements but a higher WER resulted from these changes, so they were dismissed.

4. Preparing the Data
The shared task was supplied with a reference grammar. In order to use it well, several steps are taken to prepare a more robust grammar and to clean up the transcript from Section 3.
4.1. Pre-processing of Transcript

The transcribed utterance is first cleaned for further processing.

**White space:** All irregular white-space is removed and replaced with a single empty space.

**Filler words:** Superfluous words like “yes”, “thanks”, “thank you”, “please” and “also” are removed as they have no influence on meaning and linguistic correctness. Some sentences starting with “no” and “and” fail when matching with the reference grammar. These words are removed as were words at the end of sentences (such as “no” and “is”), both of these may be artifacts resulting from erroneous parsing of noise.

**Abbreviations:** Recognized abbreviations are expanded. For example, “I’m” becomes “I am”. A total of 9 such different abbreviations were changed in this manner.

**Unique Words:** Word duplication due to false starts or repetitions are difficult to match with a regular grammar. They are therefore removed during this pre-processing phase.

**Typically Confused Words:** Some words are very difficult for the speech recognition system. One such example is the word “desert” vs. “dessert”. Similarly, “pm” is recognized by letters “p m”. These types of words were manually mapped into their correct words, given what we know about the task (which did not talk about desert).

4.2. Extending the Reference Grammar

In this system, the reference grammar was adjusted slightly by adding a number of utterances in two steps.

1. Delta Grammar: Adding correct answers from the training data that did not appear in the original reference grammar.
2. By comparing the delta grammar with the original reference grammar new structures were derived. These are listed below.

In step two the following minor changes with major impact were addition of new sentences to the grammar that were generated through a number of word substitutions as follows: “one” → “a”, “a” → “one”, “want” → “need”, “need” → “want”, “this” → “that”, “that” → “this”, “o’clock” → “pm”, “pm” → “o’clock”, “night” → “evening”, “evening” → “night”, “for” → “to”, “to” → “for”, “these” → “those”, “those” → “these”.

4.3. Creating a POS-level Reference Grammar (submitted)

The originally submitted system (OLD) is described here. An index table was created for each prompt by looking at each of the words for capitals and countries, such as “Sag: Die Hauptstadt von der Schweiz ist Bern”. Hard-coded rules as to country and city that have to appear in the answer are applied.

4.4. Language Model Score (NEW - revised)

The revised system (NEW), after the deadline is described here. Using the SRILM Tool kit to train a language model on the augmented reference grammar resulted in a Trigram model that returns a sentence score of log probability. The score reflects the goodness of the sentence syntax, given the reference grammar. It can be used as a cut-off score for accepting syntax. Using Scikit\(^1\), a decision tree was trained with the training and development test sets, resulting in a cut-off score (-29645.3398) for classification.

4.5. Creating Clusters of Prompts

A number of prompts are very similar. Therefore, some prompts are merged according to their similarity at the POS level, resulting in total of five clusters of prompts (not all prompts belong to a cluster). From these clusters, words that carry the meaning are extracted in order to match these against the incoming utterances to these prompts.

The following lists these clusters and shows an example POS grammar for the first cluster. These are used to classify incoming prompts into one of the clusters.

**Pay Cluster:** This cluster combines POS structures for all prompts employing payment options. While syntax structure is given, key words differed by prompt. An example cluster is listed in Table 4 and clearly shows how tightly related these prompts are regarding the syntax structure of the prompt.

**Table 4:** Prompts with similar POS structure (PRP MD VB TO VB IN NNS) for example Paying-Cluster:

<table>
<thead>
<tr>
<th>Pay Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sag: Ich möchte mit Dollars bezahlen</td>
</tr>
<tr>
<td>Sag: Ich möchte mit Euros bezahlen</td>
</tr>
<tr>
<td>Sag: Ich möchte mit Kreditkarte bezahlen</td>
</tr>
<tr>
<td>Sag: Ich möchte mit Mastercard bezahlen</td>
</tr>
<tr>
<td>Sag: Ich möchte mit Postkarte bezahlen</td>
</tr>
<tr>
<td>Sag: Ich möchte mit Visa bezahlen</td>
</tr>
<tr>
<td>Sag: Ich möchte mit Pfand bezahlen</td>
</tr>
<tr>
<td>Sag: Ich möchte mit Schweizer Franken bezahlen</td>
</tr>
</tbody>
</table>

**Restaurant Cluster:** This cluster combines POS structures all prompts restaurant options like “Frag: Ich möchte die Rechnung”, “Frag: Ich möchte die Dessertkarte”.

**Room Cluster:** This cluster combines POS structures for all prompts hotel booking options.

**Capital Cluster:** This cluster contained all prompts asking for capitals and countries, such as “Sag: Die Hauptstadt von der Schweiz ist Bern”. Hard-coded rules as to country and city that have to appear in the answer are applied.

**Ticket Cluster:** This cluster includes all prompts that ask for tickets for particular shows (Mamma Mia) or particular days (Monday night) and certain numbers.

5. Classification

The post-processed transcript uses the constructed reference grammar and clusters described in Section 4 through a series of rule-based expert modules for final classification of syntax and semantics for each utterance.

5.1. Basic Response Matching

Response matching is done at word and Part of Speech (POS) levels as described next. Details of each step will be explained subsequently.

1. If the cleaned transcript is matched by the augmented reference grammar (processed as in Section 4) then both syntax and semantics are classified as correct. Classification is finished.

\(^1\)http://scikit-learn.org/stable/
2. (NEW:) The sentence score (4.4) is applied to the incoming utterance. If the sentence score is below the threshold (negative log probability), i.e. it is low enough to reflect a well-formed sentence given the reference grammar, the utterance is classified as semantically correct.

3. If the utterance belongs to prompts "Pay-Cluster" or "Restaurant-Cluster", then it is represented at the POS-level.
   - If the POS-structure matches within its cluster, the syntax is classified as correct (see Section 5.2).
   - (OLD:) Words in the sentence are indexed by POS and matched against those words indexed for this POS in the corresponding prompt. If the vocabulary matches, the semantics is classified as correct. (see Section 5.3)

While the first two steps are straightforward, the third one is explained below.

5.2. POS-Level Syntax Judgement

The utterance is attempted to be matched by a module, if it did not find a direct syntactic match in the reference grammar or pass the cutoff-score in the second step (Section 4.4) and the utterance belongs to one of Pay- and Restaurant-Cluster.

This step works at the POS-level of the sentence. However, during the process of regenerating an extended reference grammar to improve coverage, a simplified approach was adapted. The idea of generating a reference grammar with extended coverage was applied to sub-problems only.

Looking at a subset of the known prompts, it was possible to determine syntactically correct sentence patterns for some prompts at the POS-level. This would allow some freedom for the speakers at the word level that is not covered by the reference grammar, thereby making the grammar more robust against out of vocabulary answers that may still be correct syntactically.

This approach was implemented for two most frequently occurring clusters, namely the Pay- and Restaurant-Cluster that make up 10-15% of the whole testing data. Other clusters (as described in Section 4.5) exist but did not get implemented yet.

Each POS-level Grammar contains the following components:

1. The POS-level reference Grammar distinguishes between the main categories of question vs. statement phrasing followed by
2. the POS-level content component and
3. the matching of length of incoming utterance with reference sentence.

5.3. Rule-Based Meaning Judgement

Since we are looking for both meaning and syntax separately, this section discusses, how meaning can be judged as correct. In order to judge correct meaning apart from correct syntax, it is useful to extract certain POS features from the incoming utterance.

1. Match Nouns at POS level
2. Match words identified as nouns
   - Knowledge-based Matching of nouns

This combination is shown in Figure 1, a) for their respective clusters. The details are explained below.

5.2.1. Question vs. Statement

Different POS patterns are used to extract questions as shown in Table 5. These were compiled from the data for each of the clusters.

Similarly, POS patterns are used to extract Statement patterns as shown in Table 6.

5.2.2. Content

Content is extracted by matching the content part of the sentence structure. These are usually verb and noun structures.

This approach covers the sentence syntax in combination with the question/statement constructs described above.

5.2.3. Length

Finally, after matching of substrings in both target grammar and incoming utterance, no extra words should remain. Having passed all stages, the utterance is judged to have correct syntax. The semantic judgement is described in the next section.

5.4. Rule-Based Meaning Judgement

Since we are looking for both meaning and syntax separately, this section discusses, how meaning can be judged as correct. In order to judge correct meaning apart from correct syntax, it is useful to extract certain POS features from the incoming utterance.

1. Match Nouns at POS level
2. Match words identified as nouns
   - Knowledge-based Matching of nouns
Table 8: Content POS Examples for Noun Structures.

<table>
<thead>
<tr>
<th>Prepositions:</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN, NN, NN</td>
<td>for friday night</td>
</tr>
<tr>
<td>IN, NN, NNS</td>
<td>with credit cards</td>
</tr>
<tr>
<td>IN, NN</td>
<td>for friday</td>
</tr>
<tr>
<td>IN, NNS</td>
<td>with dollars</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Numbers:</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD, NN</td>
<td>one ticket</td>
</tr>
<tr>
<td>CD, NNS</td>
<td>two tickets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adjectives:</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ, NN</td>
<td>musical ticket</td>
</tr>
<tr>
<td>JJ, NNS</td>
<td>musical tickets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Determinants:</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT, NN</td>
<td>the sweatshirt</td>
</tr>
<tr>
<td>DT, NN, NN</td>
<td>a grocery store</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Personal pronouns:</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRPS, NN</td>
<td>my room</td>
</tr>
<tr>
<td>PRPS, NN, NN</td>
<td>my master card</td>
</tr>
</tbody>
</table>

- Meaning Map (using Table generated in Section 4.3)

5.3.1. Key Nouns

Nouns are extracted by matching substrings at the POS-level in the incoming utterance as described in Table 8. The words found as nouns are then matched in two ways described below.

5.3.2. Knowledge-Based Matching of Nouns (by cluster)

The knowledge-based approach looks at nouns in the context of the cluster. If the expected nouns are found in the utterance, the meaning is judged as correct.

In the following clusters the noun is extracted from the utterance transcription and matched with the nouns in the responses defined by the reference grammar for the relevant prompt. The nouns are extracted according to the algorithm described in Section 5.1.

**Ticket Cluster:** The ticket cluster always asks for a number of tickets for a particular show or musical. The correct response must include the name of the show as well as the word for ticket. This cluster always asks for a certain number of tickets for a particular evening. For example, the Prompt “Frag: 2 Tickets für König der Löwen” necessarily contains the nouns “tickets” and “the lion king” or “lion king”.

**Capital Cluster:** A number of prompts ask for capital cities of certain countries. A method of extracting both of these nouns and checking against a general knowledge base has shown to be helpful in identifying correctness of meaning. For example, the list of nouns required for the Prompt “Sag: Die Hauptstadt von der Schweiz ist Bern” necessarily must contain “switzerland” and “bern” and “capital”.

**Other:** Pay, Restaurant and Room Cluster work in similar ways.

5.3.3. Meaning Map for Noun Matching (prompt based)

If a prompt was not matched with the above rules within their cluster, it uses a meaning map to match the utterance nouns against a list of nouns that were established for the corresponding prompt from the reference grammar. If there is a match, the meaning is judged to be correct.

5.3.4. False Friends

A list of false friends that can be extended in future work was added to the system to identify false meaning. This list currently holds only the word “dessert card” for “dessert menu”, which was frequently mistakenly used.

5.3.5. High-Frequency Errors

A number of high-frequency errors that students commit could be added as specific rules. For the present system, we implemented only one rule to detect incorrect usage of singular/plural. Future work can go into more detail here to extend coverage.

5.4. Pipeline-Based Final Classification

The overall system is depicted in Figure 2. Before entering the pipeline of modules, meaning and syntax are both set to false. As the utterance makes its way through the pipeline, intermediary decisions can be overwritten.

1. Pre-processing: The transcripts are formatted and syntax and semantics are set to FALSE (see Section 4.1).
2. Language Model: According to the score, the utterance syntax can be overwritten with TRUE (see Section 4.4).
3. Reference Grammar: A match with the augmented reference grammar can result in both syntax and semantics to be set to TRUE (see Section 4.2).
4. Cluster: According to clusters (as described in Section 4.5) both syntax and semantics to be set to TRUE for Pay and Restaurant Clusters and semantics can be set to TRUE for other clusters.
5. Meaning Map: Finally, the meaning map that works on each prompt (regardless of cluster) can be use to set semantics to TRUE.
6. Singular/Plural mismatch can reset the meaning to FALSE.
7. False friends match can reset the meaning to FALSE.
8. POS-level sentence matching was used in the system that was submitted to the competition and has since been replaced (see Section 4.3).

The final classification is written into the results file. If both syntax and semantics are TRUE the utterance is accepted, otherwise rejected.

6. Evaluation

6.1. Description of Test Data

Test data for the shared task was released two weeks prior to hand-in of the results. It consisted of the following items (similarly to the training data but without the transcriptions or judgements):
6.2. D-Metric

The D-Metric given in Equation 1 is used to evaluate the system performance. The variables in the equation are defined as the number of utterances that fall into each of the following categories.

- **CR** Correct Reject, **CA** Correct Accept, **FR** False Reject
- **PFA** Plain False Accept (the student’s answer is correct in meaning but incorrect English, the system accepts)
- **GFA** Gross False Accept (the student’s answer is incorrect in meaning, the system accepts)

False Accept is defined by $F_A = PFA + k \cdot GFA$, where $k$, a weighting factor that makes gross false accepts relatively more important is set to 3.

$$D = \frac{(CR/(CR + FA))}{(FR/(FR + CA))} = \frac{CR FR + CA)}{FR CR + FA} = \frac{CR}{FR}$$  (1)

6.3. Results

In this paper, we report on two results. First, the system that was submitted to the competition at SLaTE 2017. In this system the OLD module described in Section 4.3 was used instead of the NEW module that was added later. In addition the acoustic model was retrained on the complete training set. Results are given in Table 9.

Table 9: Results, where $Pr$=precision, $R$=recall, $F$=F-measure. (BK=Baseline Kaldi, QS=Our System, PP=Best Submitted Of Our Team, JJ=Best Transcript from competition.)

<table>
<thead>
<tr>
<th>Name</th>
<th>$Pr$</th>
<th>Rec</th>
<th>F</th>
<th>$FR$</th>
<th>CR</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>BK</td>
<td>0.957</td>
<td>0.439</td>
<td>0.602</td>
<td>0.951</td>
<td>0.561</td>
<td>1.694</td>
</tr>
<tr>
<td>BKOS</td>
<td>0.945</td>
<td>0.599</td>
<td>0.733</td>
<td>0.914</td>
<td>0.401</td>
<td>2.280</td>
</tr>
<tr>
<td>PPP</td>
<td>0.838</td>
<td>0.795</td>
<td>0.816</td>
<td>0.660</td>
<td>0.205</td>
<td>3.217</td>
</tr>
<tr>
<td>PPPOS</td>
<td>0.897</td>
<td>0.779</td>
<td>0.834</td>
<td>0.789</td>
<td>0.221</td>
<td>3.578</td>
</tr>
<tr>
<td>JJJ</td>
<td>0.871</td>
<td>0.848</td>
<td>0.859</td>
<td>0.717</td>
<td>0.152</td>
<td>4.710</td>
</tr>
<tr>
<td>JJJOS</td>
<td>0.872</td>
<td>0.839</td>
<td>0.856</td>
<td>0.723</td>
<td>0.161</td>
<td>4.503</td>
</tr>
<tr>
<td>NEW</td>
<td>0.903</td>
<td>0.835</td>
<td>0.868</td>
<td>0.791</td>
<td>0.165</td>
<td>4.799</td>
</tr>
</tbody>
</table>

7. Conclusions and Future Work

A rule-based system lends itself well for giving intelligent feedback to the learner. In this paper, we have attempted to build a rudimentary prototype of such a rule-based system. It is easier to understand where the student needs support, such as vocabulary or syntactic issues. A pipeline architecture allows us to separate meaning from syntax and hone in on problem areas. A lot more work is required to build a helpful feedback mechanism. Many of these rules are also very application dependent and may not generalize well to new problem sets. In future, as more data becomes available new approaches can be added to build hybrid systems. It is interesting to note that most of the system performance was gained by understanding the rule-based modules and using this to extend the reference grammar.

The final system gains most of its leverage from the extended reference grammar. We expect the modules to support robustness against new data.

8. Acknowledgements

This work was performed by Bachelor students as part of their capstone project. The authors would like to thank Xizi Wei for visiting to help with Kaldi. We also thank the support of our “Förderverein”.

9. References


The University of Birmingham 2017 SLaTE CALL Shared Task Systems

Mengjie Qian, Xizi Wei, Peter Jančovič, Martin Russell

Department of Electronic, Electrical and Systems Engineering, University of Birmingham, Birmingham B15 2TT, UK

{mxq486,xxw395,p.jancovic,m.j.russell}@bham.ac.uk

Abstract

This paper describes the system developed by the University of Birmingham for the SLaTE CALL Shared Task on grammatical and linguistic assessment of English spoken by German-speaking Swiss teenagers. Our work focused on automatic speech recognition (ASR) but we also improved the text-processing component of the system. Several approaches to training a DNN-HMM ASR system using the AMI and the German PF-STAR corpus, plus a limited amount of Shared Task data, are described. In cross-validation evaluations on the initial Shared Task data, our final ASR system achieved a word-error-rate (WER) of 9.27%, compared with 14% for the official baseline Shared Task DNN-HMM system. For text processing we expanded the baseline template-based grammar to include additional correct response patterns from the original Shared Task transcriptions. Finally, we fused the outputs of several systems at the text processing stage using linear logistic regression. Our best single and fused systems submitted to the challenge achieved ‘D’ scores of 4.71 and 4.766, respectively, on the final test set.

Index Terms: CALL, shared task, automatic speech recognition, text processing

1. Introduction

Since the 1980s, shared tasks have been a major factor in the development of many areas of speech and language technology, but there has not previously been such a task for Computer Assisted Language Learning (CALL). The 2017 SLaTE CALL Shared Task [1] was led by the University of Geneva with support from the University of Birmingham and Radboud University using recordings of English responses from German-speaking Swiss teenagers interacting with the CALL-SLT system [2]. A development set, ST-DEV, of 5,264 recordings, together with a true transcription, automatic speech recognition (ASR) outputs from a commercial and baseline DNN-HMM system, and a human judgment of grammatical and semantic correctness for each utterance, was released in July 2016. This set was reduced to 5,222 utterances in February 2017. This enabled participating laboratories to develop systems in time for the release of the 996 utterance test set, ST-TST, in March 2017.

This paper describes the three systems that we submitted to the 2017 SLaTE CALL Shared Task. Each consists of two components, automatic speech recognition (ASR) and text processing (TP). Our ASR system was developed using the Kaldi toolkit [3] and builds on the CALL Shared Task baseline ASR system. For ASR training, we replaced the WSUCAM0 corpus of read native English speech [4], that was used to train the baseline system, with a portion of the AMI corpus of unscripted speech [5] and the German PF-STAR corpus of German children reading English [6]. This plus 90% of ST-DEV was used for pre-training and training, followed by a final phase of training using only ST-DEV. The optimum amount of AMI training data (to balance with ST-DEV) and various parameters of the ASR system were determined empirically in cross-validation experiments on ST-DEV. For text processing we expanded the baseline grammar to include word sequence patterns from ST-DEV that were judged correct but were missing from the original grammar.

For the final evaluation on ST-TST we submitted results from three systems:

• Submission 1 consists of our best ASR system (9.27% WER average over cross-validation experiments on ST-DEV) trained on the whole of ST-DEV, plus the expanded TP. The optimal parameters of ASR for Submission 1 were estimated over 10-fold cross-validation experiments.

• Submission 2 is the result of fusing the outputs of six separate systems using linear logistic regression [7]. The systems all use our expanded TP with four variants of the ASR from Submission 1, the Kaldi baseline ASR and Nuance ASR.

• Submission 3 combines Nuance ASR with the expanded TP.

On ST-TST Submission 1 achieved a WER of 15.63% and Submission 1, 2 and 3 achieved D scores of 4.71, 4.766 and 2.533, respectively.

The rest of the paper is organised as follows. In section 2, we provide details of the spoken CALL shared task and the brief structure of our system. Sections 3 and 4 describe the ASR and text processing parts of our system, respectively. Finally, we present our conclusions in section 5.

2. Spoken CALL Shared Task

2.1. Introduction to the Shared Task

The shared task is based on data collected from CALL-SLT [8, 2], a speech-enabled online tool which has been under development at the University of Geneva since 2009. The system helps young Swiss German teenagers practise skills in English conversation. The items of data are prompt-response pairs, where the prompt is a piece of German text and the response is an utterance spoken in English and recorded as an audio file. The challenge of the task is to label pairs as “accept” or “reject”, accepting responses which are grammatically and linguistically correct and rejecting those incorrect either in grammar or meaning according to the judgments of a panel of human listeners [1].

There are two versions of the task: a speech-processing version and a text-processing version [1, 9]. The aim of the two versions are the same, but they have different items provided as system input. In the speech-processing version of the CALL shared task, each item consists of an identifier, a German text
prompt and an audio file containing an English language response. For the text-processing version, there is an extra text string representing the automatic speech recognition result on the audio file, which is obtained from either the official baseline Kaldi ASR system or the Nuance ASR used in the original CALL-SLT system. This paper is mainly concerned with the speech-processing version but we also improve the text-processing version of the task.

### 2.2. Scoring Metric

All the items are annotated by three native English speakers according to their linguistic correctness and their meaning (these are referred to as the language and meaning “gold standard” judgments). For linguistic correctness, both vocabulary and grammar are judged as correct or incorrect. The annotators also judge whether the answer is meaningful or not in the context of the provided prompt, labelling an utterance as “sense” or “nonsense”. It is worse for the system to accept a “nonsense” sentence than it is to accept one which is correct in terms of meaning. Comparing the system’s judgments with the language and meaning gold standards, each response falls into one of the five categories described in Table 1.

<table>
<thead>
<tr>
<th>English</th>
<th>Meaning</th>
<th>Judgment</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>✔</td>
<td>Accept</td>
<td>Correct Accept (CA)</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>Reject</td>
<td>Correct Reject (CR)</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>Accept</td>
<td>Plain False Accept (PFA)</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>Accept</td>
<td>Gross False Accept (GFA)</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>Accept</td>
<td>Reject Rate on Correct Utterances (FA)</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>Accept</td>
<td>Gross False Accept (GFA)</td>
</tr>
</tbody>
</table>

Let CR, CA, FR, PFA and GFA denote the number of utterances in the corresponding categories as given in Table 1. The evaluation of the overall quality of the systems in the Shared Task is performed using a differential response score, \( D \), which is defined [1] as the ratio of the reject rate on incorrect answers to the reject rate on correct utterances, i.e.,

\[
D = \frac{CR/(CR + FA)}{FR/(FR + CA)} = \frac{CR(FR + CA)}{FR(CR + FA)},
\]

where the false acceptance (FA) is defined as \( FA = PFA + k \cdot GFA \), with \( k \) being a weighting factor that causes gross false accepts to have a more prominent effect. In the current evaluation this is set to 3.

### 2.3. Training and Test Corpus

The training (ST-DEV) and test (ST-TST) sets of the Shared Task (ST) were released in July 2016 and March 2017, respectively. The training set contains 5,222 utterances (approx. 4.8 hours of recordings) and the test set contains 996 utterances (approx. 0.89 hours of recordings). The speakers are male and female German-speaking Swiss students ranging in age from 12 to 15 years. No specific information about speakers was released for the Shared Task, so we did not know which utterances are spoken by the same speaker or whether the speaker for a particular utterance is male or female. This has implications for ASR development because Kaldi can exploit this information if it is available. In our cross-validation experiments, the released training set ST-DEV was separated into training data and development data at the ratio of 9:1.

### 2.4. System Structure

The architecture of an automatic system used for the Shared Task is depicted in Figure 1. The system consists of two parts. The first is an ASR system that converts a given audio recording into a text. The second part is a text processor which takes the transcribed audio and makes a judgment of whether the utterance is accepted or rejected according to the language and meaning. A baseline ASR system built using Kaldi and a baseline text processing system were provided by the organisers of the challenge on the website [9], and we will introduce these separately in sections 3.1 and 4.1.

![Figure 1: Structure of the system.](image)

### 3. Automatic Speech Recognition

#### 3.1. Official Baseline System

The provided baseline ASR system is a hybrid deep neural network – hidden Markov model (DNN-HMM) built using Kaldi [3]. The Shared Task data used to develop the baseline ASR is a super-set of ST-DEV, comprising recordings of 5,500 utterances. This corpus is referred to as ST-BASE. Thus ST-BASE includes ST-DEV plus some utterances that were not subsequently released. The baseline Kaldi ASR system is trained on about 18.93 hours of recordings from WSJCAM0 [4] and 90% of ST-BASE. The remaining 10% of ST-BASE is used for testing.

A speech signal frame is represented using 13-dimensional MFCCs with a context of 15 frames (i.e., 7 frames before and after). A neural network with 4 hidden layers and 1024 neurons for each layer was used. The output layer is a softmax layer and each node of this layer represents the posterior probability of the context-dependent HMM states. The initial training of the DNN-HMM is performed using an alignment obtained from a triphone GMM-HMM system, which was trained using an alignment obtained from a monophone GMM-HMM system. For GMM-HMM systems, a speech signal frame was represented using 13-dimensional MFCCs with delta and delta-delta coefficients appended, forming a 39-dimensional feature vector. After training the DNN model, an adaptation is applied by fine-tuning the network using only the ST data. The language model (LM) is a bigram model trained on the reference transcription of the ST data. In cross-validation evaluations, this system achieved an average WER of 14.03%.

#### 3.2. Developed Systems

This section describes the development of ASR systems that formed part of our submission to the Shared Task challenge. All systems were developed using Kaldi. The developed DNN-HMM systems used similar configuration as the official baseline ASR system except of the following differences. We used 13-
dimensional MFCCs with context of 11 frames (i.e., ±5 frames) in most experiments – the use of a slightly smaller context than the official baseline ASR was accidental and was considered to have little effect on results. In addition, some DNN-HMM experiments (see section 3.2.2) were also performed using Mel-scaled filter-bank energies with the same 11 frames context. The neural network with 6 hidden layers and 1024 neurons for each layer was used. In all our experiments, a trigram language model trained with ST training data was used.

3.2.1. Training Data Selection

The first issue we explored was the effect of using different training data on ASR performance. The results of these experiments for monophone and triphone GMM-HMM systems and two DNN-HMM systems are shown in Table 2. In the table, “MonoPhone”, “TriPhone” and “DNN” indicate systems that were pre-trained and trained on the complete training set.

In all of these experiments we also applied the fine-tuning strategy used in the baseline system, re-training the DNN model with only the ST training data after initial pre-training and training with the complete training set. “DNN.reTrain” corresponds to the same DNN-HMM system as “DNN” but after additional training using just the ST data.

Our first ASR system, Sys1 in Table 2, used only ST training data.

The WSJCAM0 corpus comprises recordings of read speech from adults who are native speakers of English. Each of these factors is inconsistent with the Shared Task data. Therefore, as an alternative, we replaced WSJCAM0 with the AMI [5] corpus in our training data. The AMI corpus consists of 100 hours of recordings of unscripted speech from adults participating in simulated meetings. Also, although the recordings are in English, English was not the first language of many of the participants. The AMI corpus was recorded using a wide range of devices, including close-talking and far-field microphone, individual and room-view video cameras. We used 77.3 hours of IHM (Individual Headset Microphone) data from the corpus in our experiments.

Although the properties of the AMI corpus are closer to those of the Shared Task, there is still a miss-match between the ages of the speakers. A model trained with AMI data will be biased towards adults’ speech and will not necessarily represent the speech characteristics of young teenagers. For this reason we explored adding different amounts of AMI data to ST training data: 100% (Sys2), 50% (Sys3) and 20% (Sys4). The results are shown in Table 2.

In a further effort to incorporate more of the characteristics of the ST speakers in our training set, we also added a German English corpus, the PF-STAR corpus of recordings of read English speech spoken by German children [6], to the training set. The complete PF-STAR corpus contains more than 60 hours of speech, including read and spontaneous native language speech in British English, German and Swedish and non-native read English from German, Italian and Swedish children, aged between 4 and 15. The German part of the PF-STAR corpus (PSG) was collected from German children and includes native German recordings and non-native English recordings. The non-native English speech from PSG is used in our experiments. It contains about 3.4 hours of recordings of read speech collected from 57 German children who are aged from 10 to 15. We built a DNN system with ST, AMI and PSG training data using the same methods as those described above. The results are shown as Sys5 in Table 2.

### Table 2: \%WER of development set using models trained on different training data.

<table>
<thead>
<tr>
<th>System</th>
<th>WER (%)</th>
<th>Sys1</th>
<th>Sys2</th>
<th>Sys3</th>
<th>Sys4</th>
<th>Sys5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MonoPhone</td>
<td>31.78</td>
<td>57.95</td>
<td>44.24</td>
<td>41.08</td>
<td>35.69</td>
<td></td>
</tr>
<tr>
<td>TriPhone</td>
<td>17.59</td>
<td>27.68</td>
<td>22.55</td>
<td>21.29</td>
<td>20.29</td>
<td></td>
</tr>
<tr>
<td>DNN</td>
<td>19.50</td>
<td>22.69</td>
<td>18.85</td>
<td>15.76</td>
<td>19.71</td>
<td></td>
</tr>
<tr>
<td>DNN.reTrain</td>
<td>-</td>
<td>14.61</td>
<td>13.61</td>
<td>13.07</td>
<td>14.97</td>
<td></td>
</tr>
</tbody>
</table>

**Sys1**: only ST training data  
**Sys2**: ST training data plus all the AMI data  
**Sys3**: ST training data plus 50% of the AMI data  
**Sys4**: ST training data plus 20% of the AMI data  
**Sys5**: ST training data plus PSG and 20% of the AMI data

From the results in Table 2, we see that adding data to the training set can improve the performance of the DNN model (although it may have disadvantages for the GMM-HMM model). Including 20% of AMI in the training set results in a greater advantage than adding all of the AMI data. The retrained DNN model for Sys2 has 14.61\% WER which is better than that for Sys1 with about 25\% relative improvement. Sys4 achieves a WER of 13.07\%, which corresponds to a 33\% improvement relative to Sys1. We expected that Sys5 would give the best performance, but it only outperforms Sys4 in the case of the GMM-HMM models and does not show an advantage for the DNN-HMM models.

3.2.2. Adaptation

This section presents application of feature normalisation and adaptation, specifically, Cepstral Mean Normalisation (CMN) and feature-space maximum likelihood linear regression (fMLLR). In Kaldi, each utterance is associated with a speaker label and consequently CMN and fMLLR are performed per-speaker. However, the speaker label information is not available in ST dataset. As such, in our Sys4 and Sys5, we used a single speaker-id for all utterances, which resulted in using a globally calculated statistics for CMN and fMLLR. In Sys6 and Sys7, we explored the application of CMN and fMLLR per-utterance basis, i.e., each utterance was considered to be from a different speaker. This was implemented in Kaldi by making the speaker-ids identical to the utterance-ids. In the case of fMLLR [10], the transformation was performed on dimensionality-reduced features. These features were obtained by first applying LDA on the 143-dimensional vector of MFCCs in context to decorrelate and reduce its dimension to 40-dimensional features and then further decorrelating using maximum likelihood linear transform (MLLT). In addition to the use of MFCCs, experiments were also performed using Mel-scaled filter-bank energies.

Experimental results for systems from 4 to 7 are presented in Table 3. It can be seen that the use of features transformed using fMLLR provides considerable performance improvements, e.g., for Sys4, from 15.76\% to 13.82\% and further to 10.77\% after retraining. The use of per-utterance fMLLR transforms, as in Sys6 and Sys7, provided further large improvements over the use of a single global transform. The best system, using per-utterance fMLLR and the training set containing ST and 20% of AMI, achieved 8.90\% WER.

3.3. Final ASR Submitted to the Challenge

In our experimental evaluations before the submission, Sys7 performed slightly better than Sys6. As such, our submissions
Table 3: %WER of development set using models trained with mixed data (ST, AMI and PSG), the first two columns are for ST with one global speaker-id and the last two columns are for ST with different speaker-ids.

<table>
<thead>
<tr>
<th>WER (%)</th>
<th>Sys4</th>
<th>Sys5</th>
<th>Sys6</th>
<th>Sys7</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN (fbank)</td>
<td>17.41</td>
<td>21.83</td>
<td>14.58</td>
<td>15.08</td>
</tr>
<tr>
<td>DNN (mfcc)</td>
<td>15.76</td>
<td>19.71</td>
<td>14.00</td>
<td>13.50</td>
</tr>
<tr>
<td>DNN (fmllr)</td>
<td>13.82</td>
<td>17.49</td>
<td>10.70</td>
<td>10.27</td>
</tr>
<tr>
<td>DNN.reTrain (fbank)</td>
<td>13.68</td>
<td>17.88</td>
<td>12.50</td>
<td>12.28</td>
</tr>
<tr>
<td>DNN.reTrain (mfcc)</td>
<td>13.07</td>
<td>14.97</td>
<td>10.95</td>
<td>11.78</td>
</tr>
<tr>
<td>DNN.reTrain (fmllr)</td>
<td>10.77</td>
<td>11.99</td>
<td>8.90</td>
<td>9.16</td>
</tr>
</tbody>
</table>

Sys4: ST training data plus 20% of the AMI data, one speaker-id for ST.
Sys5: ST training data plus PSG and 20% of the AMI data, one speaker-id for ST.
Sys6: ST training data plus 20% of the AMI data, different speaker-ids for ST.
Sys7: ST training data plus PSG and 20% of the AMI data, different speaker-ids for ST.

were based on Sys7. However, after the submission, we have found a minor mistake in data arrangement, which when corrected resulted in Sys6 actually performing slightly better than Sys7. Note that all results presented in this paper are after the correction was made.

The final system we used for the submission was built by following the procedure of Sys7 but using all 100% of the ST-DEV data for training the system (instead of only 90% as used in cross-validation experiments presented in the previous section). This DNN-HMM system used MLLR-transformed features and was trained first using all the ST-DEV data plus 20% of AMI and PSG data and then further trained using only the ST-DEV data. The values for language model weight (lmwt), acoustic model weight (acwt) and insertion penalty (p) parameters were set based on best performance in our cross-validation experiments.

4. Text Processing

4.1. Official Baseline System

The official baseline text processing system, provided by the ST challenge organisers on their website [9], is based on using a reference template-based grammar. This grammar is generated based on a set of templates of responses for each prompt [11]. The baseline reference grammar includes 565 prompt-units, each prompt-unit consists of a German prompt and a set of possible responses to it. Since a German text prompt is provided for each item in the ST, we could compare the prompt with the prompt-units in the reference grammar and obtain a list of possible valid responses. If an ASR transcription of a given utterance was in the response list, then this utterance would be labelled as “accept”, otherwise, it would be labelled as “reject”.

4.2. Developed System

The main part of our text processing system is based on the baseline system. We expanded the reference grammar using the method described in [11], trying to make it as complete as possible. Apart from this, a pre-processing of the ASR output was included in order to deal with words due to hesitation and word repetitions, which are difficult to handle by the grammar. An extra fusion back-end, which could take advantage of several ASR outputs, was used in our Submission 2.

4.2.1. Expanded Reference Grammar

We found that when using the baseline text processing grammar, there were a large number of false rejections on responses which seemed correct. This led us to realise that the set of responses for some prompts was not sufficient in the baseline grammar. In order to create a more complete grammar, we input the true transcriptions of speech utterances into the text processing. Theoretically, those responses which were labelled as “correct” based on the gold standard should have all been accepted. However, we have found a considerable number of rejections and these could only have been due to transcriptions not being covered by the grammar. We have also found that a few gold standard judgments were actually not correct, thus, some false rejections were actually rejected correctly.

We went through all the false rejections and added the correct transcriptions that had a correct human gold standard judgments to the grammar. We then applied text processing to the true transcriptions with the updated grammar, and went through all the false rejections and false acceptances again and updated the grammar accordingly. This procedure was repeated a few times, at each step taking care that the responses added to the grammar did not cause a large increase of false acceptances. After this, we used the actual ASR output as input to the text processing and applied the same grammar updating procedure.

We also considered adding some commonly occurring incorrect ASR outputs into the grammar in order to reduce the number of false rejections further. One such example of incorrect ASR output was that “london eye” was recognised as “london a“ – such a minor error could be understood easily in real life. We considered adding these texts into the grammar, but we did not do this in our final version of grammar because of a high possibility of increasing the number of false acceptances.

4.2.2. Pre-processing

In providing a response to a given prompt, subjects may often hesitate, be uncertain or want to modify/correct their answer. These result in two main issues when assessing responses, which are difficult to remedy directly in the reference grammar. Hesitations and uncertainty may often result in insertion of words like “um” and “uh” in speech and these may appear at any place in the response. We have also noticed that responses sometimes started with the word “hello”. A sentence should not be rejected just because it contains such words. However, it is difficult to include these words in the reference grammar due to their possibly arbitrary location in the sentence. Thus, we removed these types of words from the ASR output before it was passed to the main text processing.

The other issue is repetition of words or modification/correction of the response. This is also not suitable to be handled by the reference grammar. We assumed that children tend to correct their response during the repetition, so the latter part of the repetition would be better. Thus, when repetition happens, we exclude the former part.

The above two steps of pre-processing provided a considerable performance improvement. However, there are still a few further issues which we have not tackled yet in our current system. One is that there are many false-start words in speech, for instance, “i won’t five tickets”, “brown troll trousers”. The reason why false-start words occur is similar to repetition, but they are harder to be excluded from the texts.
4.3. Fusion

We chose the best ASR system and the best parameters according to cross-validation experiments on ST-DEV. However, the values of the best parameters were inconsistent across different cross-validation partitions. Therefore we were not confident that the same parameters would also be optimal for ST-TST. Hence we built multiple systems, each with different ASR system parameters but the same expanded TP, and fused their outputs. We employed the weighted summation fusion approach with parameters trained on the development set to take advantage of the multiple systems.

The final output of the system is “accept” or “reject”, which is a 2-class classification. For fusion we transferred the judgments into 2-class scores. Let class \( c_1 \) and \( c_2 \) represent “accept” and “reject”, respectively. If the judgment for item \( x \) is “accept”, then the score should be \( score_{c_1}(x) = 1 \), \( score_{c_2}(x) = 0 \), and if it is “reject”, then the score should be \( score_{c_1}(x) = 0 \), \( score_{c_2}(x) = 1 \). In our experiments, we use the log score:

\[
\text{score}_{c}(x) \leftarrow \log(\text{score}_{c}(x) + \epsilon).
\]

Let there be \( K \) input systems where the \( i \)th system outputs the log score vector \( \text{score}_{c,i}(X) \). Then the fused score \( \text{score}_{c}(X) \) is given by:

\[
\text{score}_{c}(X) = \sum_{i=1}^{K} w_{c,i} \cdot \text{score}_{c,i}(X).
\]

The weight, \( w_{c,i} \), can be trained on the training data. After obtaining the fused score, we could assign the class for item \( x \) by:

\[
\text{class}(x) = \text{arg max }_c \text{score}_{c}(x).
\]

Fusion was achieved using the linear logistic regression based fusion module in the FoCal toolkit [7].

4.4. Official Submissions

For the final evaluation on ST-TST data, we submitted results from three systems. These are summarised below together with their achieved \( D \) scores:

Submission 1 (system JJJ) on the official SLaTE CALL Shared Task results table [9] consisted of our best single ASR system and our expanded TP system. The ASR system used values for parameters (lmwt, acwt and p) that were found optimal on 10-fold cross-validation experiments. This submission achieved \( D \) score of 4.710.

Submission 2 (KKK) was the result obtained by fusing the outputs of six separate systems using linear logistic regression. The systems all used our expanded TP with four variants of the ASR from Submission 1 (with different parameter setup), the Kaldi baseline ASR and Nuance ASR. This submission achieved \( D \) score of 4.766.

Submission 3 (LLL) combined baseline Nuance ASR system with our expanded TP system. As such, this enables to evaluate the effect of our expanded TP. This submission achieved \( D \) score of 2.533.

The above results show that fusing multiple systems provided only minor performance gain over the use of the single best ASR system.

Further details of experimental results on the ST-DEV and ST-TST data are presented in Table 4. In the table, \( %\text{Corr} \) and \( %\text{Acc} \) denote percentage words correct and percentage accuracy, respectively – these were obtained using the HResults tool from HTK [12] applied to the output of each ASR system. Results show that our developed system performed considerably better on both datasets than baseline Nuance and Kaldi systems. It can also be seen that performance is considerably lower for all ASR systems on the ST-TST data compared with the ST-DEV data, indicating that there may be a mismatch between these two sets of data. The remaining lines in Table 4 present the \( D \) scores obtained by using the baseline, \( D_{\text{baseTP}} \), and our expanded, \( D_{\text{outTP}} \), text processing. It can be seen that the use of our expanded grammar in the text processing can potentially have a large positive effect on the \( D \) score. The level of improvement seems to be proportional to the quality of the input passed to the text processing. For instance, on the ST-TST, the \( D \) score improved from 4.512 to 27.617 when using the true transcription, while the improvement was only by 0.175 when using Nuance ASR output (i.e., ASR whose speech recognition performance was weak).

Table 4: Recognition performance (%Corr, %Acc) of ASR systems and \( D \) score for the development and test set when using true transcription and output of Nuance and Kaldi baseline recognition systems and our Submission 1 ASR system.

<table>
<thead>
<tr>
<th>True transcr.</th>
<th>Nuance</th>
<th>Kaldi</th>
<th>Our-ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST-DEV:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Corr</td>
<td>100.00</td>
<td>74.40</td>
<td>88.38</td>
</tr>
<tr>
<td>%Acc</td>
<td>100.00</td>
<td>68.25</td>
<td>85.50</td>
</tr>
<tr>
<td>( D_{\text{baseTP}} )</td>
<td>4.231</td>
<td>1.950</td>
<td>2.278</td>
</tr>
<tr>
<td>( D_{\text{outTP}} )</td>
<td>28.976</td>
<td>2.102</td>
<td>3.892</td>
</tr>
<tr>
<td>ST-TST:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Corr</td>
<td>100.00</td>
<td>72.97</td>
<td>79.40</td>
</tr>
<tr>
<td>%Acc</td>
<td>100.00</td>
<td>66.84</td>
<td>74.08</td>
</tr>
<tr>
<td>( D_{\text{baseTP}} )</td>
<td>4.512</td>
<td>2.358</td>
<td>1.753</td>
</tr>
<tr>
<td>( D_{\text{outTP}} )</td>
<td>27.617</td>
<td>2.533</td>
<td>2.379</td>
</tr>
</tbody>
</table>

5. Conclusion

This paper has described the University of Birmingham’s submissions to the 2017 SLaTE CALL Shared Task challenge. We submitted three systems, each comprising an ASR and TP component. Our initial focus was ASR and our best DNN-HMM system, developed with Kaldi using the AMI, PF-STAR (German) and Shared Task corpora, achieves WERs of 9.16% and 15.63% on ST-DEV and ST-TST, respectively. We also improved the TP component by expanding the reference grammar and pre-processing ASR output. We submitted three systems to the challenge. Submission 3 (“LLL” on the official Shared Task results table [9]), combining the baseline Nuance ASR with our expanded TP, achieves a \( D \) score of 2.533 on ST-TST. Submission 1 (“JJJ”), combining our best ASR and expanded TP, achieves a \( D \) score of 4.710. Finally, Submission 2 (“KKK”) is the fusion of six separate systems, each using our expanded TP but with six different ASRs (baseline Nuance, baseline Kaldi and four variants of our best ASR system). This submission achieves the highest \( D \) score of 4.766 on ST-TST. Thus, best performance is obtained by fusing multiple complete systems. However, the performance improvement relative to the system that uses the single best ASR system is marginal.
6. References

task for spoken CALL?” in Proc. Language Resources and Evaluation Conf. (LREC), 2016.

swiss education system: A large-scale experiment on the basis
of english CALL-SLT,” Ph.D. dissertation, Université de Genève,
2015.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek,
N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz,
“The kaldi speech recognition toolkit,” Proc. IEEE Workshop on
Automatic Speech Recognition and Understanding (ASRU), 2011.

CAM0: A british english corpus for large vocabulary continu-
ous speech recognition,” in Proc. IEEE Int. Conf. on Acoustics,

T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal et al.,
“The AMI meeting corpus: A pre-announcement,” in Int. Work-
shop on Machine Learning for Multimodal Interaction. Springer,

M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, “The
PF_STAR children’s speech corpus,” in Proc. European Confer-
ence on Speech Communication and Technology, 2005.

sión and calibration of multi-class recognition scoretutorial

[8] E. Rayner, P. Bouillon, N. Tsourakis, J. Gerlach, Y. Nakao, and
C. Baur, “A multilingual CALL game based on speech transla-
tion,” in Proc. Language Resources and Evaluation Conf. (LREC),
Valetta, Malta, 2010.

[9] “Spoken CALL shared task official website.”

ture processing for deep neural networks,” in Proc. Interspeech,
2013, pp. 109–113.

learning of response grammars in a spoken CALL system,” in
Proc. Workshop on Speech and Language Technology in Educa-
tion (SLaTE), 2015.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu,
175, 2002.
Using an Automated Content Scoring System for Spoken CALL Responses:
The ETS submission for the Spoken CALL Challenge

Keelan Evanini†, Matthew Mulholland‡, Eugene Tsuprun†, Yao Qian‡

Educational Testing Service R&D
†660 Rosedale Rd., Princeton, NJ, USA
‡90 New Montgomery Street, Suite 1500, San Francisco, CA, USA
{kevanini, mmulholland, etsuprun, yqian}@ets.org

Abstract

In this study we investigate the performance of an automated content scoring system for accepting or rejecting learner responses in a spoken CALL application. Specifically, we employed a system based on word and character n-gram features in a support vector machine learning framework that was originally designed for scoring content in written texts and augmented its standard feature set with additional features from the following categories: prompt bias, text-to-text similarity to reference responses, and automatically detected grammatical errors. This system achieved a D score of 4.353 (compared to a baseline score of 1.694) on the test set consisting of Kaldi ASR output in the 2017 Spoken CALL Challenge. In this paper we also provide an analysis of the impact of the size and nature of the training data set (human transcriptions vs. ASR output) on the model’s performance and present the results of feature ablation experiments to demonstrate which of the additional features are most helpful.

Index Terms: Spoken CALL Challenge, automated content scoring, bias features, text-to-text similarity features, grammar features

1. Introduction

Many studies have demonstrated the positive effects of targeted grammar feedback provided by language instructors to language learners in a classroom environment on the learners’ ability to produce grammatically correct utterances [1, 2, 3, 4], and several types of feedback have been investigated, including prompts (repeating the question, providing metalinguistic feedback, etc.) and reformulation strategies (recasting the learner’s response, providing explicit corrections, etc.). Recent improvements in automatic speech recognition (ASR), natural language processing (NLP), and spoken dialog system (SDS) infrastructure have enabled the creation of interactive, speech-based Computer-Assisted Language Learning (CALL) applications that attempt to automate the process of providing grammar feedback to language learners [5, 6, 7]. In order for these automated spoken CALL applications to provide valid grammar feedback to language learners, it is necessary for them to accurately detect erroneous responses. Since this field of research is relatively new, and since few shared resources exist for comparing various error detection methodologies on a common data set, [8] proposed a shared task for spoken CALL, in which spoken English responses provided by native speakers of German while using a CALL application would be released to the community along with annotations about the grammatical and semantic correctness of each response that can be used to train models for predicting whether the responses are erroneous or not.1 This paper describes the system that ETS developed to participate in this shared task at SLaTE 2017.

To address the task of determining whether to accept or reject a spoken response produced in the context of a spoken CALL application, we explore the use of a pre-existing automated content scoring system developed at ETS. This system uses a machine learning approach based primarily on character and word n-gram features and has been applied to score content in a wide variety of tasks. Most research with this content scoring system has been conducted in the context of short answer tasks for the domains of elementary and secondary schools in areas such as science, English language arts, and math [9, 10]; however, it has also been successfully applied to longer written texts from other domains, such as a writing task from a standardized assessment for music teachers [11]. In addition, a recent study explored the use of this content scoring system for automated scoring of non-native spoken English responses provided in the context of a standardized assessment of English speaking proficiency [12]. That study is not directly comparable to the current study, though, since the nature of the spoken responses and the scores across the two studies are quite different. The spoken responses in the Spoken CALL Shared Task quite short (typically a single sentence) and were scored for grammar and meaning only, whereas the spoken responses in [12] consisted of spontaneous speech of approximately one minute in duration and were scored by human raters based on scoring rubrics that contained additional aspects of speaking proficiency, such as pronunciation and fluency.

This paper is organized as follows: in Section 2 we describe the features contained in the content scoring system that was used to train the prediction models as well as the additional types of features that were investigated; Section 3 presents two experiments that were conducted on the training set in order to determine the relative importance of the various additional features as well as the impact of the quality and quantity of training data on the model’s performance; the official test results that were obtained by our submission to the Spoken CALL Shared Task are presented in Section 4 along with additional experiments on the test set that were conducted after the official submission deadline for the shared task; finally, Section 5 discusses the main findings and suggests steps for future research.

2. Features

We explored the following four types of features in our models: features extracted using the automated content scoring system,
bias features based on the prompts and prompt categories, features based on the similarity between a speaker’s utterance and the sample responses for each prompt, and features produced by a grammar checker. These features are described in more detail in the following sections.

2.1. Content Features
As described in Section 1, we used a pre-existing system for automated content scoring to extract four classes of features:

- Character n-grams for \( n = 2 \) to 5
- Token unigrams and bigrams
- Syntactic dependencies
- Length of response in characters

Each feature is represented in a sparse binary format. For example, the n-gram and syntactic dependency features have a value of 1 if present in a response and an implicit 0 if not. The length feature, however, is composed of a set of length bin features. Specifically, length is calculated for each response using the following formula (where “length” is the number of characters in the response): \( \log_2(1 + \text{length}) \). Thus, there are features for different lengths and these features are represented in the same sparse format as the other features mentioned above. Syntactic dependencies were extracted using the ZPar dependency parser [13].

These features are referred to as “Content” features in this paper. Since the motivation for this study was to examine the performance of the automated content scoring system for the task of accepting / rejecting short responses provided in the context of a spoken CALL application, the Content features were included in all of the models that we experimented with, and the additional features described in the subsequent sections were added to augment the model’s performance. Since it was not possible to train prompt-specific models for this task (see Section 2.2), generic models were trained to apply across all prompts with the expectation that the Content features could capture common patterns of linguistic errors that are shared across different prompts.

2.2. Prompt Bias
When the content scoring system is applied to score responses to test questions, a separate model is typically trained for each prompt [9]. While a few of the prompts in the Spoken CALL Task contain a reasonable number of responses for training robust models in the training data set—four prompts contain more than 100 responses with the largest number, 195, corresponding to the prompt Sag: Ich habe keine Reservation (Say: I don’t have a reservation)—most of the 413 prompts do not contain enough responses per prompt to train models. In fact, 44 of the prompts only contained a single response in the training set, e.g., Frag: 3 Tickets für Mamma Mia (Ask for: 3 tickets for Mamma Mia) and Sag: Ich möchte um 22 Uhr morgen abreisen (Say: I want to leave tomorrow at 10pm). Furthermore, 11 out of the 264 prompts in the test set, e.g., Frag: 2 Tickets für Montagabend (Ask for: 2 tickets for Monday evening) and Frag: pinkige Hosen (Ask for: pink trousers) had no corresponding responses in the training set, so it would have been impossible to train prompt-specific models for these prompts.

Therefore, we adopted an approach in which a single model was trained using data from all of the prompts in the training set. In order to enable the model to be able to leverage information about prompt-specific grammar and vocabulary patterns that correspond to different scores, we explored the use of prompt bias features in the model. The prompt bias features consisted of a single binary feature per prompt per response represented in a sparse matrix. These features were inspired by the approach taken by [14] which used the content scoring system for Task 7 in the SemEval 2013 shared task.

In addition to the prompt bias features, we also explored the use of bias features based on categories of prompts that are similar to one another. The motivation for this approach was the observation that many of the prompts are expected to elicit responses that are similar to each other in form and differ only based on a small number of key words. These similarities among expected responses for certain prompts could be leveraged by the bias features to reduce data sparsity; in addition, the prompt category bias features could help address the fact mentioned above that some responses in the test set were drawn from prompts that were not contained in the training set. An example of one of the prompt categories that was designed for this study can be seen in the many prompts in the data set that are related to the communicative task of buying clothes, specifically, making a request to buy a particular item of clothes. All of the sample responses listed in the reference grammar for these prompts are identical except for the specific item of clothes that was specified in the prompt. For example, Table 1 shows some of the 161 sample responses for two prompts in this category: Frag: blaue Sandalen (Ask for: blue sandals) and Frag: braune Stiefel (Ask for: brown boots); in the table, the content words that are specific to each prompt and are expected to differ are underlined.

<table>
<thead>
<tr>
<th>Prompt Category</th>
<th>Sample Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>buying_clothes</td>
<td></td>
</tr>
<tr>
<td>Frag: blaue Sandalen (Ask for: blue sandals)</td>
<td>Frag: braune Stiefel (Ask for: brown boots)</td>
</tr>
<tr>
<td>blue sandals please</td>
<td>brown boots please</td>
</tr>
<tr>
<td>can i buy blue sandals please</td>
<td>can i buy brown boots please</td>
</tr>
<tr>
<td>yes could you please give me blue sandals</td>
<td>yes could you please give me brown boots</td>
</tr>
<tr>
<td>yes i’d like blue sandals please</td>
<td>yes i’d like brown boots please</td>
</tr>
</tbody>
</table>

Table 1: Sample responses for two prompts in the buying_clothes category

In total, 49 different prompt categories were defined manually by the authors for the 487 prompts listed in the reference grammar provided by the task organizers. A few of these prompt categories contained a large number of responses; Table 2 lists the 5 most frequent along with the number of prompts associated with each and a few examples.

On the other hand, some of the prompts were not closely related to any of the other prompts, such as Frag: Welcher Bus fährt dorthin? (Ask: Which bus goes there?) and Sag: Mir gefällt die Farbe nicht (Say: I don’t like the color). In these cases, the prompts were not grouped together with any other prompts and the prompt category label was identical to the prompt label; 22 prompts were treated in this manner for
Table 2: Five most frequent prompt categories

<table>
<thead>
<tr>
<th>Prompt Category</th>
<th>N</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>buying_tickets</td>
<td>106</td>
<td>Frag: Tickets für Sonntagabend (Ask for: tickets for Sunday evening), Frag: ein Ticket für heute Abend (Ask for: one ticket for this evening), ...</td>
</tr>
<tr>
<td>departure</td>
<td>55</td>
<td>Sag: Ich möchte um 9 Uhr morgen abreisen (Say: I want to leave tomorrow at 9am), Sag: Ich möchte am Sonntagabend gehen (Say: I would like to go Sunday night), ...</td>
</tr>
<tr>
<td>buying_clothes</td>
<td>31</td>
<td>Frag: blaue Sandalen (Ask for: blue sandals), Frag: braune Stiefel (Ask for: brown boots), ...</td>
</tr>
<tr>
<td>ordering_food</td>
<td>48</td>
<td>Frag: Apfelkuchen (Ask for: the apple pie), Frag: ein Wasser ohne Kohlensäure (Ask for: a glass of still water), ...</td>
</tr>
<tr>
<td>directions</td>
<td>24</td>
<td>Frag: Wo ist der Coiffeur? (Ask: Where is the hairdresser?), Frag: Wo ist die Tate Modern? (Ask: Where is the Tate Modern?), ...</td>
</tr>
</tbody>
</table>

2.3. Similarity to Reference

We also investigated several text-to-text similarity features to measure the similarity between an utterance to the sample responses for each prompt provided by the task organizers in the file referenceGrammar.xml. These features are motivated by the fact that utterances that are not contained in the grammar but that are similar to the sample responses are more likely to be correct. This set of features is intended to address two potential sources of false rejects by the system, namely correct responses that are not listed in the grammar and correct responses that are in the grammar but that are misrecognized by the ASR system, thus causing the input to the scoring module to be out-of-grammar. The following four text-to-text similarity features were explored in this study:

- Minimum Word Error Rate (WER) between the utterance and each reference response for the prompt
- Maximum WER between the utterance and each reference response for the prompt
- Mean WER across all reference responses for the prompt
- BLEU score between the utterance and the reference responses for the prompt

The WER is obtained by first calculating the edit distance, i.e., the minimum distance obtained by applying dynamic programming to align two sequences of words with possible insertions, deletions, and substitutions [16], between the speaker’s utterance and one of the responses in the reference grammar and then dividing the distance by the length of the sample response. We calculated the WER for each utterance/sample-response pair for a corresponding prompt and then calculated the minimum, mean, and maximum WER values across all of the sample responses for a prompt and used these three values as features in the model.

2.4. Grammar

We used the language-check Python wrapper2 for Language-Tool, an open-source English proofreading package, to check whether an utterance contains any grammatical errors based on predefined rules; the number of errors detected in each response was then used as a feature. For example, one of the system’s rules (Phrase Repetition Rule) looks for duplicated phrases; an example response from the training set that would be flagged by this rule is i would like a ticket to ticket to pay with dollars (ID #5974). Another example rule (BaseForm Rule) checks whether the base form of a verb is used after certain modal verbs; an example response that would be flagged by this rule is i will boots (ID #5958).3

Several rules in LanguageTool were not appropriate for use with ASR output (for instance, a rule that checks whether a sentence ends with a period) or for spoken responses in general (for example, a rule that checks whether input is a sentence fragment). A total of 5 rules were excluded for these reasons from the feature calculation pipeline, leaving a total of 200 grammar rules; of these, 39 rules were matched by one or more responses in the training set. Overall, only 3.8% (200 out of 5,222) of the utterances in the training set were associated with one or more grammar errors for this feature.

3 Experiments on Training Set

This section describes experiments that were conducted on the training set in order to determine the relative impact of the different types of features on the model’s performance as well as to investigate the influence of the two different sources of input to the model: transcriptions and ASR output. Since the WER of the Kaldi ASR output provided by the task organizers (0.147) was substantially lower than the WER of the Nuance ASR output (0.319), the Kaldi ASR output was used for all of these experiments. The models described in this section (as well as the models that were used to produce results on the test set) use the machine learning approach that is typically used for the automated content scoring system, namely support vector regression with hyperparameter grid search optimization. In order to train models with different combinations of features from the different feature sets described in Section 2, the features from

2https://pypi.python.org/pypi/language-check
3Note that the speaker intended will to be the main verb of the utterance (as in want) and boots to be a plural noun, not a 3rd-person singular verb. Since LanguageTool’s rules are based on pattern matching and do not make use of POS tagging, system errors of this nature are to be expected. Additionally, it should be pointed out that, even though the system labeled the error incorrectly, the utterance does, in fact, contain an error (i.e., use of the incorrect verb), and would therefore be correctly rejected by the application of this rule.
the different sets were combined into a single feature vector for each response which was then used as input to learn the support vector regression parameters for each model.

3.1. Feature Evaluation

We conducted a feature ablation experiment by training several models on the training set to determine the relative contributions of the different types of features for predicting erroneous utterances when added to the features from the content scoring system (this experiment was conducted after the test results were submitted for the shared task, so its findings could not be incorporated into the official submission). We trained a series of models that all included the base set of Content features and added one additional feature type at a time (we also trained a model that used only the base Content features). This resulted in a total of eight models for this experiment: one for each of the three WER features, one for the BLEU feature, one for the grammar feature, one for each of the two bias features, and one for the base Content feature set by itself. The results, in terms of D score, which was the chosen evaluation metric for the Spoken CALL Shared Task and is defined as the ratio of the relative correct reject rate to the relative false reject rate [8] obtained with these models (using the output of the Kaldi ASR system and 10-fold cross-validation on the training set) are reported in Table 3.

![Table 3: Results obtained using the Kaldi ASR output and 10-fold cross-validation on the training set for eight different models based adding individual features one by one to the Content features](image)

After this experiment was completed, a second round of feature ablation was conducted in which each feature was added into the base Content feature set starting with the best-performing features until eventually all features were included; the results of this experiment are presented in Table 4. As the table shows, the model based on the Content features with the addition of the Minimum WER and Mean WER features performed best with a D score of 11.504; the performance of the model then declined with the addition of subsequent features. In addition to the D score, additional performance metrics were calculated on the best performing system on the training data set in order to more fully understand its strengths and weaknesses; these are as follows: 3,719 correct reject (71.2%), 242 plain false accept (4.6%), 235 gross false accept (4.5%), 865 correct reject (16.6%), 161 false reject (3.1%), 87.8% accuracy, 88.6% precision, and 95.9% recall.

![Table 4: Results obtained using the Kaldi ASR output and 10-fold cross-validation on the training set for eight different models based on the step-wise addition of each feature set to the Content features](image)

3.2. Quantity and Quality of Training Data

Since the training data set in the Spoken CALL shared task contains both transcriptions and ASR output for the spoken responses, we experimented with different configurations of these two sources of input to the model. Specifically, we conducted 10-fold cross-validation experiments on the training data set using only the Content features to determine what type of input to the model is most effective for scoring the ASR output; the results of these experiments are presented in Table 5.

![Table 5: 10-fold cross-validation results using different configurations of training and testing data for models consisting of the Content features](image)

As the table shows, a model trained on the transcriptions performs worse when evaluated on the ASR output than a model trained on the ASR output. This is expected, due to the mismatch between the characteristics of the vocabulary and grammar contained in the responses in the training and testing sets. However, Table 5 also shows that the performance improves when the ASR output and the transcriptions are combined together in the training set (this was done by including separate entries in the model training set for the ASR output and transcription, thus doubling the size of the training set from 5,222 to 10,444). This seems to suggest that, despite the mismatch, the addition of the transcriptions to the model can provide some additional information that is helpful for the model’s prediction. However, it could also be the case that the combined model simply performs better because it is based on twice the amount of training data.

---

4For the Spoken CALL Shared Task, a weighting factor was applied in the calculation of the D score to penalize gross false accepts (where the system accepts a response that was annotated as incorrect in meaning) three times as heavily as plain false accepts (where the system accepts a response that was annotated as incorrect in grammar but correct in meaning).

5Note that the value of 8.838 reported for the “Kaldi ASR” training/Kaldi ASR” test row in Table 5 differs from the value reported in Table 4 when using the base Content features on the Kaldi ASR due to the use of an earlier version of the data provided by the organizers and a different ordering of the data (with respect to the cross-validation folds).
To investigate which of these explanations is correct, we first conducted an experiment in which we randomly selected five different halves of the combined training data set to make it equal in size to the original training set, i.e., 5,222 responses consisting of 2,611 responses represented by their ASR output and 2,611 responses represented by their transcription. Then, five different cross-validation experiments with shared folds were conducted (regardless of whether a response was randomly assigned to the ASR or transcription category for a given experiment, the ASR output was always used for the responses in the testing set for each fold). The average D score across these five cross-validation experiments was 8.4 (with a range from 8.1 to 8.9), which is similar to the performance of the model trained using the ASR output on the training set consisting of 5,222 responses. This result likely indicates that the performance improvement observed by combining the ASR and transcriptions in the training set is due primarily to the increased size of the training set, not because the transcriptions themselves are beneficial when testing on ASR output.

Finally, Table 5 shows that a model trained on the transcriptions performs quite well when tested on the transcriptions (D = 52.4); this result represents an upper bound on performance that can be expected if a spoken CALL system had completely accurate ASR.

4. Test Results

Based on the results presented in Section 3.2 about the differences between system performance when the models were trained on ASR output or a combination of ASR output and transcriptions, we decided that one of the official test submissions should be based on ASR output only and another should be based on a combination of ASR output and transcriptions; all of the features described in Section 2 were included in these two submissions. For the third submission, we used models trained only on ASR output with the subset of features consisting of a combination of the Content features and bias features (including both the prompt bias and prompt category bias features), since initial results we obtained suggested that the bias feature set performed well in comparison to the other additional feature sets (these initial results were obtained prior to the controlled feature ablation study that was presented in Section 3.1). Table 6 presents the official test results in terms of the D score for these three submissions (labeled MMM, NNN, and OOO by the task organizers); for all of these systems, the models were trained using all responses from the training set and tested on the Kaldi ASR output provided by the task organizers.

As Table 6 shows, the system with all features substantially outperformed the system trained using the subset of Content and bias features; furthermore, the system trained on ASR output only slightly outperformed the system trained on a combination of ASR output and transcriptions. All three submissions outperform the Kaldi baseline provided by the shared task organizers which checks to see whether the utterance is contained in the reference grammar.

As shown in Section 3.1, a system trained using the Content features plus two features based on similarity to the sample responses in the reference grammar (Minimum WER and Mean WER) performed best on the training set. Similar evaluations were conducted on the test set to determine whether a different combination of features could improve on the official results shown in Table 6; these results are shown in Table 7.

As shown in Table 7, the features based on comparing the test response to the sample responses in the reference grammar (WER and BLEU features) outperform the grammar and bias features; this is similar to the results shown for the training data in Table 3, except that the BLEU features are more effective on the test set. Finally, we also trained models with feature sets added one-by-one in the same order as was done for the training set in Table 4; these results are presented in Table 8.

Table 7: Results obtained using the Kaldi ASR output on the test set for eight different models based adding individual features one by one to the Content features

<table>
<thead>
<tr>
<th>Features</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content + Prompt Category Bias</td>
<td>2.801</td>
</tr>
<tr>
<td>Content + BLEU</td>
<td>4.214</td>
</tr>
<tr>
<td>Content + Prompt Bias</td>
<td>3.248</td>
</tr>
<tr>
<td>Content</td>
<td>4.207</td>
</tr>
<tr>
<td>Content + Grammar</td>
<td>3.674</td>
</tr>
<tr>
<td>Content + Max. WER</td>
<td>4.525</td>
</tr>
<tr>
<td>Content + Mean WER</td>
<td>4.358</td>
</tr>
<tr>
<td>Content + Min. WER</td>
<td>4.204</td>
</tr>
</tbody>
</table>

Table 8: Results obtained using the Kaldi ASR output on the test set for eight different models based on the step-wise addition of each feature set to the Content features

As shown in Table 8, the best performance on the test set was produced by a model that included all of the features except for the Prompt Category Bias feature, with a D score of 4.565. Additional evaluation metrics for this model are as follows: 649 correct accept (65.2%), 51 plain false accept (5.1%), 59 gross false accept (5.9%), 170 correct reject (17.1%), 67 false reject
(6.7%), 82.2% accuracy, 85.5% precision, and 90.6% recall.

5. Discussion and Conclusion

In this paper we described the system that was used by ETS to participate in the Spoken CALL Shared Task at SLaTE 2017. The system is based on a pre-existing automated content scoring system that consists primarily of character and n-gram features. This system substantially outperforms the baseline on both the training and the test data sets, thus indicating the usefulness of this system for the task of accepting/rejecting responses in a spoken CALL application.

In addition, we explored the use of three additional types of features (bias, text-to-text similarity to sample responses, and grammar) in combination with the features from the content scoring system. Since these experiments showed that the WER-based features that compared the spoken response to the sample responses in the reference grammar were most effective for this task, future research could benefit from investigating additional matching approaches that could potentially be more robust to ASR errors than WER. [18] explores several of these, such as regular expressions, BLEU, and LM score, in the context of scoring the content of short responses provided in the context of an assessment of English speaking proficiency for teachers of English.

The difference in performance between the cross-validation results on the training set reported in Section 3 and the results on the test set reported in Section 4 is striking. Further research will be necessary to investigate why the performance is so much lower on the test set. As mentioned in Section 2.2, some of the responses in the test set were drawn from prompts that were not seen in the training set, and it is possible that the model did not generalize well to these responses to unseen prompts. However, these responses only represent a small percentage of the test set, and therefore cannot be the sole explanation for the substantial performance difference. Further error analysis should be conducted to determine whether there are other characteristics of the test set that differ from the training set that could explain the observed difference. As shown in Sections 3 and 4, the accuracy of the best system on the test set was 82.2% compared to an accuracy of 87.8% for the best system on the training set; this difference of 5.6% in accuracy corresponded to a difference of 6.939 in D score. Since the proportion of false rejects on the test set that differ from the training set that could explain this difference of 5.6% in accuracy corresponded to a difference of 5.6% in accuracy of the best system on the test set was 82.2% compared to an accuracy of 87.8% for the best system on the training set; this difference of 5.6% in accuracy corresponded to a difference of 6.939 in D score. Since the proportion of false rejects on the test set that differ from the training set that could explain this difference of 5.6% in accuracy corresponded to a difference of 6.939 in D score. Since the proportion of false rejects on the test set that differ from the training set that could explain this difference of 5.6% in accuracy corresponded to a difference of 6.939 in D score. Since the proportion of false rejects on the test set that differ from the training set that could explain this difference of 5.6% in accuracy corresponded to a difference of 6.939 in D score. Since the proportion of false rejects on the test set that differ from the training set that could explain this difference of 5.6% in accuracy corresponded to a difference of 6.939 in D score. Since the proportion of false rejects on the test set that differ from the training set that could explain this difference of 6.939 in D score.
Deep-Learning Based Automatic Spontaneous Speech Assessment in a Data-Driven Approach for the 2017 SLaTE CALL Shared Challenge

Yoo Rhee Oh, Hyung-Bae Jeon, Hwa Jeon Song, Byung Ok Kang, Yun-Kyung Lee, Jeon-Gue Park, and Yun-Keun Lee

Speech Intelligence Research Group, Electronics and Telecommunications Research Institute, South Korea
yroh@etri.re.kr, hbjeon@etri.re.kr, songhj@etri.re.kr, bokang@etri.re.kr, yunklee@etri.re.kr, jgp@etri.re.kr, yklee@etri.re.kr

Abstract

This paper presents a deep-learning based assessment method of a spoken computer-assisted language learning (CALL) for a non-native child speaker, which is performed in a data-driven approach rather than in a rule-based approach. Especially, we focus on the spoken CALL assessment of the 2017 SLaTE challenge. To this end, the proposed method consists of four main steps: speech recognition, meaning feature extraction, grammar feature extraction, and deep-learning based assessment. At first, speech recognition is performed on an input speech using three automatic speech recognition (ASR) systems. Second, twenty-seven meaning features are extracted from the recognized texts via the three ASRs using language models (LMs), sentence-embedding models, and word-embedding models. Third, twenty-two grammar features are extracted from the recognized text via one ASR system using linear-order LMs and hierarchical-order LMs. Fourth, the extracted forty-nine features are fed into a full-connected deep neural network (DNN) based model for the classification of acceptance or rejection. Finally, an assessment is performed by comparing the probability of a output unit of the DNN-based classifier with a predefined threshold. For the experiments of a spoken CALL assessment, we use English spoken utterances by Swiss German teenagers. It is shown from the experiments that the D score is 4.37 for the spoken CALL assessment system employing the proposed method.

Index Terms: Spoken CALL assessment, DNN based classifier, sentence-embedding, word-embedding, sequence-to-sequence, dependency parse tree

1. Introduction

There are considerable researches on a computer-assisted language learning (CALL) based on speech recognition. Moreover, many researches are related to the pronunciation assessment of an imitated speech such as SRI’s EduSpeak [1]. On the other hand, some researches are related to various kinds of assessments (grammar, semantic, vocabulary, etc) of a speech with the increased freedom of speaking, such as CALL-SLT [2] and GenieTutor [3]. Among them, this paper focuses on a spoken CALL assessment of the CALL-SLT system, as a participant of the 2017 SLaTE CALL shared challenge [4]. To this end, we propose a deep-learning based spoken CALL assessment method, which is performed in a data-driven approach while the baseline method supported by the organizer of the challenge is performed in a rule-based approach.

2. The 2017 SLaTE CALL shared challenge

The challenge aims to assess the utterance obtained from a CALL system in terms of the meaning and the grammar for a given prompt. And, the corpus consists of 5,222 utterances and 996 utterances for a training data and a test data, respectively.

In addition, the annotation of each utterance contains a prompt, a transcription, a meaning evaluation result, and an overall evaluation result.

A baseline assessment method first performs speech recognition on an utterance and then determines the utterance to be acceptable if the recognized text exactly matches one of the reference texts corresponding to the given prompt [5]. Thus, the organizer provides the recognized text (Text\textsubscript{Nuance}) from a Nuance ASR (ASR\textsubscript{Nuance}) and the recognized text Text\textsubscript{Kaldi} from a Kaldi ASR (ASR\textsubscript{Kaldi}) for each speech data. Also, it provides the reference texts corresponding to each prompt (Texts\textsubscript{Ref}). When evaluating the test data using a metric of D, the baseline method achieves 2.35, 1.69, and 4.51 for three CALL assessment systems: (a) a CALL system (CALL\textsubscript{Baseline}) using Text\textsubscript{Nuance}, (b) a CALL system (CALL\textsubscript{Kaldi}) using Text\textsubscript{Kaldi}, and (c) a CALL system (CALL\textsubscript{transcription}) using the transcription texts. In addition, the word error rates (WERs) of ASR\textsubscript{Nuance} and ASR\textsubscript{Kaldi} are 33.1% and 25.1%.

The detailed description of the challenge is explained in [4].

3. Proposed deep-learning based spoken CALL assessment in a data-driven approach

Fig. 1 shows the proposed deep-learning based spoken CALL assessment method. In other words, the speech processing component first performs speech recognition for an input speech using three ASRs such as ASR\textsubscript{Nuance}, ASR\textsubscript{Kaldi}, and an ASR (ASR\textsubscript{SLaTE2017}) developed for the challenge. And then, three recognized texts, Text\textsubscript{Nuance}, Text\textsubscript{Kaldi}, and Text\textsubscript{SLaTE2017}, are obtained from ASR\textsubscript{Nuance}, ASR\textsubscript{Kaldi}, and ASR\textsubscript{SLaTE2017}, respectively.

Next, the text processing component performs a meaning feature extraction, a grammar feature extraction, and a deep-learning based assessment using Text\textsubscript{Nuance}, Text\textsubscript{Kaldi}, and Text\textsubscript{SLaTE2017}, a given prompt, and Texts\textsubscript{Ref}. That is, the meaning feature extraction step generates nine meaning features from each of Text\textsubscript{Nuance}, Text\textsubscript{Kaldi}, and Text\textsubscript{SLaTE2017} and then concatenates the three sets of the nine meaning features. Moreover, the grammar feature extraction step generates twenty-two grammar features from Text\textsubscript{SLaTE2017}; we only...
Figure 1: Main procedure of the proposed deep-learning based spontaneous spoken CALL assessment method in a data-driven approach for the challenge.

3.1. An ASR system for the challenge

An original ASR system \(\text{ASR}_{\text{original}}\) is a common-domain American English ASR system for Korean speakers. In addition, the characteristics of \(\text{ASR}_{\text{original}}\) are a sampling rate of 16 kHz, the target language of American English, the speaker’s mother tongue of Korean, and the target LM domain of a common-domain. Moreover, the acoustic models (AMs) of \(\text{ASR}_{\text{original}}\) are configured as deep neural network hidden Markov models (DNN-HMMs) and the LM of \(\text{ASR}_{\text{original}}\) is configured as a 3-gram LM that is trained with common-domain texts.

In order to provide a better text to the text processing component of the proposed method, we adjust \(\text{ASR}_{\text{SLaTE2017}}\) to \(\text{ASR}_{\text{baseline}}\) by reflecting the characteristics of the speech data of the challenge, such as a sampling rate of 8 kHz, the target language of British English, and the speaker’s mother tongue of German. That is, the AMs are adapted using the training data of the challenge and the LM \((M_{\text{ASR},3\text{gram}})\) is interpolated with a challenge-domain 3-gram LM that is trained with TextsRef. In addition, a pronunciation model is adapted by using English pronunciation variants that are frequently mispronounced by the German people.

When evaluating the test data, the WER is 14.9% for \(\text{ASR}_{\text{SLaTE2017}}\) and the D score is 2.86 for the spoken CALL system \(\text{CALL}_{\text{baseline}}\) employing the baseline assessment method using the recognized text from \(\text{ASR}_{\text{SLaTE2017}}\). Moreover, Table 1 summarizes the performance comparisons of the four CALL systems such as \(\text{CALL}_{\text{Nuance baseline}}, \text{CALL}_{\text{Kaldi baseline}}, \text{CALL}_{\text{transcription baseline}},\) and \(\text{CALL}_{\text{SLaTE2017 baseline}}\).

3.2. Deep-learning based meaning features

As summarized in Table 2, nine meaning features are extracted from an input text and the features are categorized into three levels: CALL-level, sentence-level, and word-level, depending on the scope of the meaning to be analyzed. The CALL-level features are extracted using four LMs to determine whether the meaning of the input text corresponds to the learning scope of a CALL system. The sentence-level features are extracted using two sentence-embedding models to determine whether the meaning of the input text includes the meaning of the prompt. The word-level features are extracted using a one word-embedding model to determine whether the meanings of words contained in the input text are related to the meanings of words contained in TextsRef.
3.2.1. CALL-level meaning features based on LMs

The assumption is that, if the meaning of an input text is within the language learning scope of a CALL system, then the probability of an input text would be larger when using a CALL-domain LM than when using a common-domain LM. Three CALL-domain LMs using challenge-related texts and a common-domain LM using general-domain texts are generated as follows:

- $M_{\text{domain},3\text{gram}}$: CALL-domain, 3-gram LM
- $M_{\text{domain},5\text{gram}}$: CALL-domain, 5-gram LM
- $M_{\text{domain},\text{LSTM}}$: CALL-domain, a two-layer long short-term memory (LSTM) recurrent neural network (RNN) with 200 hidden units
- $M_{\text{common},3\text{gram}}$: common-domain, 3-gram LM

When an input text is entered, the average log-probability per word excluding word boundaries ($\text{ppl}_1$, [8]) is calculated using each CALL-level LM. In particular, predefined penalties under certain conditions are applied in the calculation of the log-probability of a word. As a result, we obtain four CALL-level meaning features ($F_{m1}$, $F_{m2}$, $F_{m3}$, and $F_{m4}$) using $M_{\text{domain},3\text{gram}}$, $M_{\text{domain},5\text{gram}}$, $M_{\text{common},3\text{gram}}$, and $M_{\text{domain},\text{LSTM}}$.

3.2.2. Sentence-level meaning features based on sentence-embedding models

For sentence-level meaning feature extraction, we adopt a sentence-embedding approach that is increasingly utilized in various research such as natural language processing, machine translation, and so on [9]. We generate two sequence-to-sequence (seq2seq) models using challenge-related texts as follows:

- $M_{\text{seq2seq,fwd}}$: a forward LSTM-RNN based encoder-decoder model with 200 hidden units
- $M_{\text{seq2seq,bwd}}$: a backward LSTM-RNN based encoder-decoder model with 200 hidden units

When an input text and its prompt are entered, the reference sentences (Texts$_{\text{Ref,prompt}}$) corresponding to the prompt are first obtained from Texts$_{\text{Ref,f}}$. Then, a cosine similarity is calculated between the two vectors of the hidden state values of the encoder of a sentence-level model for the input text and each of Texts$_{\text{Ref,prompt}}$. Next, we obtain the maximum cosine similarities and the average of up to 100 values, hereafter referred to as $\text{cos}_1$ and $\text{cos}_100$, respectively. As a result, we obtain two sentence-level meaning features, $F_{m5}$ and $F_{m6}$, by selecting $\text{cos}_1$ and $\text{cos}_100$ respectively using $M_{\text{seq2seq,fwd}}$. In addition, we obtain one feature, $F_{m7}$, by selecting $\text{cos}_1$ using $M_{\text{seq2seq,bwd}}$.

3.2.3. Word-level meaning features based on word-embedding models

Similar to Section 3.2.2, we adopt a word-embedding approach for a word-level meaning feature extraction. We first obtain one word2vec model [10] using the challenge-related texts as follows:

- $M_{\text{word2vec}}$: a word2vec model

When an input text and its prompt are entered, the reference sentences (Texts$_{\text{Ref,prompt}}$) corresponding to the prompt are first obtained from Texts$_{\text{Ref,f}}$. Then, we measure the similarity distances between the input text and each Texts$_{\text{Ref,prompt}}$ using the word-level model and obtain the minimum of the similarity distances. The similarity distance between two texts is calculated by performing a dynamic time warping (DTW), wherein the distance between two words is the output of the word2vec model. As a result, we obtain one word-level meaning feature, $F_{m8}$, by selecting the minimum similarity distance using $M_{\text{word2vec}}$.

In order to capture the use of the keywords of the prompt, we also obtain one additional feature, $F_{m9}$, by selecting keywords that are most likely present in Texts$_{\text{Ref,prompt}}$ and calculating the penalized similarity distance with the modified distance value of each keyword.

3.3. Parsed-text based grammar features

We first detect awkward word sequences by examining the linear word sequence of an input text. However, this is not sufficient to detect grammatical errors because the order of sentence constituents is not considered. In order to consider the order of sentence constituents, the input text is first parsed into a dependency parse tree and then each 1-depth subtree is converted into a word sequence, which is referred to as ‘text$_{\text{parsed,sub}}$’. Next, we identify incorrect grammar by examining the word sequence of text$_{\text{parsed,sub}}$.

As summarized in Table 3, we extract two grammar features from an input text for a grammar check based on the linear word sequence. We also extract twenty grammar features from the text$_{\text{parsed,sub}}$ of an input text to conduct a grammar check based on the hierarchical word sequence.

### Table 3: Summary of the twenty-two grammar features of the proposed method

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Model Text</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical-order</td>
<td>Parsed</td>
<td>PPL1/max</td>
</tr>
<tr>
<td>Linear-order</td>
<td>Raw</td>
<td>PPL1/max</td>
</tr>
<tr>
<td>$M_{\text{ASR},3\text{gram}}$</td>
<td>Raw</td>
<td>PPL1/max</td>
</tr>
<tr>
<td>$M_{\text{words,5\text{gram}}}$</td>
<td>Raw</td>
<td>PPL1/max</td>
</tr>
<tr>
<td>$M_{\text{words,3\text{gram}}}$</td>
<td>Raw</td>
<td>PPL1/max</td>
</tr>
<tr>
<td>$M_{\text{word2vec}}$</td>
<td>Raw</td>
<td>PPL1/max</td>
</tr>
</tbody>
</table>

For a grammar check that detects awkward word sequences in an input text, the word sequence of the input text is compared with the general distribution of word sequences obtained from
When an input text is entered, ppl1 is calculated using $M_{\text{ASR,3gram}}$. Moreover, the average log-probability per word including word boundaries (ppl, [8]) is calculated using $M_{\text{ASR,3gram}}$. As a result, two linear-order grammar features of $F_{g9}$ and $F_{g10}$ are obtained by calculating ppl1 and ppl of the input text using $M_{\text{ASR,3gram}}$.

### 3.3.2. Hierarchical-order grammar features from the parsed texts of an input text

For a grammar check that detects awkward word sequences based on the sentence constituents of an input text, each parsed text, textparsed,sub, of the input text is compared with a formal distribution of parsed texts and an informal distribution of parsed texts, respectively. The formal distribution is generated using the textparsed,sub of the formal texts from three kinds of resources: the texts of English text books, TextsRef1, and the ones of GenieTutor CALL system [3]. The informal distribution is generated using the textparsed,sub of the informal texts. The informal texts are obtained by performing three steps: generating texts with grammar errors from the formal texts, obtaining parsed,sub of texts with grammar errors, and excluding textparsed,sub that are overlapped with textparsed,sub of the formal texts. To this end, three correct distribution LMs, one incorrect distribution LM, and $M_{\text{ASR,3gram}}$ are used as follows:

- $M_{\text{formal,3gram}}$: formal, 3-gram, parsed text
- $M_{\text{formal,6gram}}$: formal, 6-gram, parsed text
- $M_{\text{formal,3gram,ext}}$: formal, 3-gram, parsed text, expanded reference responses with the synonym words
- $M_{\text{informal,3gram}}$: informal, 3-gram, parsed text
- $M_{\text{ASR,3gram}}$: a 3-gram LM for ASR_SLATE2017 raw text

In fact, $M_{\text{ASR,3gram}}$ is chosen since the text obtained from a phrase-level grammar chunk can be a part of the linear word sequence of an input text. Fig. 2 shows an example of the textparsed,sub of the text, ‘two ticket to London’ and the comparison of the ppl1 of the textparsed,sub using the formal and informal distribution of $M_{\text{formal,3gram}}$ and $M_{\text{informal,3gram}}$, respectively.

When an input text is entered, the text is decomposed into textparsed,sub using a Stanford parser [11]. Next, ppl1 is computed using an LM for each textparsed,sub and then the maximum, minimum, and average values of the set of ppl1 are obtained. In addition, we similarly obtain the maximum, minimum, and average ppl values for the input text. Thus, we obtain 16 hierarchical-order grammar features by selecting the maximum and average values of ppl1 and ppl for each of the three formal distribution models and $M_{\text{ASR,3gram}}$. We also obtain four features by selecting the minimum and average values of ppl1 and ppl for the informal distribution model.

![Figure 2: Example of the textparsed,sub of the text, 'two ticket to London', and the comparison of ppl1 using the formal distribution of $M_{\text{formal,3gram}}$ and the informal distribution of $M_{\text{informal,3gram}}$.](image)

![Figure 3: Performance comparisons based on the ROC AUC of the proposed method by increasing each feature by one for the 9 meaning features](image)

### 3.4. Deep-learning based classifier for the spoken CALL assessment

In order to assess an input utterance using various kinds of meaning and grammar features, we adopt a deep-learning based approach that is a state-of-art method in many applications. To this end, we empirically configure a six-layer DNN of one input layer, four hidden layers, and one output layer. The input layer consists of 49 linear units for the 27 meaning features and 22 grammar features; the output layer is a softmax layer with two units that correspond to the final targets, accept or reject. Each of the first, second, and fourth hidden layers is a fully-connected (FC) layer that contains 256 units with rectified linear unit (ReLU) activation, while the third hidden layer is a dropout layer that prevents an overfitting.

When the 49 meaning and grammar features are entered, we calculate the output unit probabilities of the 6-layer DNN based classifier. And then, the input data is decided as accept when the probability of the accept-labeled output unit is greater than a predefined threshold. The predefined threshold is selected by maximizing the D value throughout the training data in the constraints of the minimum values of $D$, $Fr$, and $Cr$ as 0.04. In particular, the constraints are empirically selected in order to compensate for the fact that the D value becomes too larger when the $Fr$ value is getting close to zero.

### 4. Experiments on the CALL assessment

We first evaluated the proposed method using a transcription text in order to eliminate the effect of ASR errors in Section 4.1. Next, we compared the performances of the proposed method using recognized texts from three ASR systems in Section 4.2. Especially, we additionally used a common binary classification performance metric, a receiver operating characteristic area under curve (ROC AUC).

#### 4.1. Performance on the proposed method using a transcribed text

First, we validated the nine meaning features of the proposed method using transcription texts by entering the features one by one into the proposed method. In other words, we extracted $F_{m1}$ from each transcription, trained a DNN-based classifier which input layer has one unit corresponding to $F_{m1}$, and evaluated the performance based on ROC AUC. Next, we extracted $F_{m1}$ and $F_{m2}$ from each transcription, trained a DNN-based classifier which input layer has two units corresponding to $F_{m1}$ and $F_{m2}$.

1The descriptions of the $D$, $Fr$, and $Cr$ are explained in [4].
and $Fm_2$, and evaluated the performance. This process was performed until the nine meaning features were used. It was shown from the Fig. 3 that the performance of a CALL assessment system was generally improved each time the feature was increased by one for the meaning feature.

Second, we validated the 22 grammar features of the proposed method using transcription texts in the same way to the verification of the meaning features. It was shown from the Fig. 4 that the performance was generally improved each time the feature was increased by one for the meaning feature.

Third, we evaluated the proposed method using both the meaning and grammar features extracted from transcription texts. When evaluating the test set, the ROC AUC score was 0.921 for a CALL assessment system (CALLtranscription) employing the proposed method using both meaning and grammar features, whereas the ROC AUC scores were 0.893 and 0.877 for a CALL system employing the proposed method using meaning features and a CALL system employing the proposed method using grammar features, respectively.

From the experiments on the proposed method using transcription texts, we concluded that each meaning feature was positively correlated for the proposed method. Similarly, each grammar feature was positively correlated for the proposed method. In addition, the performance of the proposed method was considerably improved when combining the meaning features and the grammar features.

### 4.2. Performance on the proposed method using the recognized texts from three ASR systems

Fig. 5 shows the ROC AUCs of the four CALL assessment systems: (a) a CALL assessment system (CALL\textsubscript{Nuance\_proposed}) employing the proposed method using Text\textsubscript{Nuance}, (b) a CALL assessment system (CALL\textsubscript{Kaldi\_proposed}) employing the proposed method using Text\textsubscript{Kaldi}, (c) a CALL assessment system (CALL\textsubscript{SLaTE2017\_proposed}) employing the proposed method using Text\textsubscript{SLaTE2017}, and (d) a CALL assessment system (CALL\textsubscript{Nuance,Kaldi,SLaTE2017\_proposed}) employing the proposed method using Text\textsubscript{Nuance}, Text\textsubscript{Kaldi}, and Text\textsubscript{SLaTE2017}.

First, the ROC AUC of CALL\textsubscript{Nuance,SLaTE2017\_proposed} was 0.81 whereas the ROC AUCs of CALL\textsubscript{Nuance\_proposed} and CALL\textsubscript{Kaldi\_proposed} were 0.75 and 0.77, respectively. Therefore, it was notable from Table 1 and Figure 5 that the performance of the proposed method could be improved by combining the recognized texts from various kinds of ASR systems.

Moreover, we evaluated the assessment performance of the proposed method using the metric $D$. To this end, we first selected an optimal threshold to maximize the $D$ score using the training set. Especially, we used the constraints of the minimum values of $Fr$ and $Cr$ of 0.04 as described in Section 3.4. After that, the $D$ score of the test set was calculated using the selected optimal threshold. Fig. 6 shows an example of the optimal threshold selection for CALL\textsubscript{Nuance,Kaldi,SLaTE2017\_proposed}, where the straight and dotted lines represent the $D$ scores according to thresholds when evaluating the training data and the test data, respectively, and the shaded area presents the corresponding range to the constraints.

Finally, we obtained the $D$ scores for the five CALL assessment systems, CALL\textsubscript{transcription}, CALL\textsubscript{Nuance}, CALL\textsubscript{Kaldi}, CALL\textsubscript{SLaTE2017}, and CALL\textsubscript{Nuance,Kaldi,SLaTE2017\_proposed} when evaluating the training set and test set, respectively.
Table 4: Performance comparison based on the metric D of the baseline method and the proposed method using a transcription, TextNuance, TextKaldi, and TextSLaTE2017.

<table>
<thead>
<tr>
<th>Transcription</th>
<th>Baseline method</th>
<th>Proposed-method</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextNuance</td>
<td>4.51</td>
<td>5.38</td>
</tr>
<tr>
<td>TextKaldi</td>
<td>2.35</td>
<td>2.56</td>
</tr>
<tr>
<td>TextSLaTE2017</td>
<td>1.69</td>
<td>2.28</td>
</tr>
<tr>
<td>TextNuance,Kaldi,SLaTE2017 (submitted)</td>
<td>-</td>
<td>4.37</td>
</tr>
<tr>
<td>TextNuance,Kaldi,SLaTE2017 (fixed)</td>
<td>-</td>
<td>4.49</td>
</tr>
</tbody>
</table>

In short, the proposed method considerably improved the performance of the spoken CALL assessment system by adopting several deep-learning based methods when compared to the baseline method. Moreover, the proposed method reduced the manual efforts by using a data-driven approach. However, there are several future works of the proposed method such as the investigation of the significant features among the proposed 27 meaning features and 22 grammar features, the development of the hybrid method combining the proposed method and the baseline method, etc.

5. Conclusion and discussion
This paper proposed the deep-learning based spoken CALL assessment method for the 2017 SLaTE CALL shared challenge. Especially, we focused on the generation of the spoken CALL assessment system with minimal manual efforts, by using a data-driven approach rather than using a rule-based approach. Moreover, we tried to adopt deep-learning method, which was known as the state-of-the-art machine learning method.

The proposed method consisted of the speech processing component and the text processing component. In other words, when an input speech was entered, speech recognition was performed using each of ASRNuance, ASRKaldi, and ASRSLaTE2017 in the speech processing component. After that, the three texts recognized from three ASRs were passed into the text processing component. In the text processing component, the twenty-seven meaning features were generated by extracting nine features from each of the recognized texts and by concatenating them. Moreover, the nine meaning features consisted of four CALL-level features using Nuance, Kaldi, and SLaTE2017, and twenty features were for a hierarchical-order text. Finally, the forty-nine features were fed into the fully-connected 6-layer neural network for the classification of acceptance. And then, an assessment was performed by comparing the output probability of the DNN based classifier with a predefined threshold. We first validated the designed meaning and grammar features of the proposed method and then compared the performances of the several CALL assessment systems, CALLtranscription, CALLKaldi, CALLproposed, CALLproposed , and CALLproposed . It was shown from the spoken CALL assessment experiments that the D score was 4.37 for the proposed method whereas the D score was 2.35 or 1.69 for the baseline method. In addition, after modifying some mistakes, we obtained the D score as 4.49 using the proposed method.

6. Acknowledgements
This work was partially supported by the ICT R&D program of MSIP/IITP. [R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for the language learning] and by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [17ZS1800, Development of self-improving and human-augmenting cognitive computing technology]

7. References
Syntactic and semantic features for human like judgement in spoken CALL

Ahmed Magooda\textsuperscript{1}, Diane Litman\textsuperscript{2}

\textsuperscript{1}University of Pittsburgh, PA, USA
\textsuperscript{2}University of Pittsburgh, PA, USA
amagooda@cs.pitt.edu, litman@cs.pitt.edu

Abstract

Educational applications of Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) have included providing learners with helpful and accurate feedback. In this paper we present a system that takes a first step towards providing feedback during spoken Computer-Assisted Language Learning (spokenCALL). We propose a machine learning based approach that combines syntactic and semantic features in order to accept or reject a textual response given a provided prompt. Our approach was evaluated as part of the SpokenCALL shared task, ranking third place among the submitted systems and outperforming the provided baselines.

Index Terms: Natural Language Processing, Language modeling, Word embedding, Machine learning

1. Introduction

Any educational application has to deal with learner answer/response variability, and spoken Computer-Assisted Language Learning (spokenCALL) is no exception. For any system question or prompt, it is expected that learners will not all provide the exact same answer/response, and it is the system’s responsibility to deal with this variability. To provide accurate feedback, a system should thus be able to tolerate different language aspects (e.g., word synonymy, paraphrasing) in learner responses. The process of giving response feedback can be done in multiple stages (e.g., accept/reject, highlight errors, propose more accurate answers). In this paper we focus on the first stage (accept/reject), since the first stage has to be the most accurate. It is meaningless for a system to highlight errors or propose corrections if the answer/response is correct to begin with, while it can be less severe to just reject a wrong answer/response without proposing any corrections.

In this context, this paper targets the task of providing accept/reject feedback for data collected from CALL-SLT \cite{1}, a speech-enabled Computer-Assisted Language Learning application that is based on prompts and associated responses. The application was developed to help Swiss German teens practice English conversation. A prompt is a piece of German text that is introduced to the learner, while the response is an English response recorded using an audio capturing device. The response is supposed to be an English sentence that follows the German prompt request. Based on data collected from users of the CALL-SLT tool, Baur et al. \cite{2} proposed the SpokenCALL shared task\textsuperscript{1}. Given a German prompt and an English response, the shared task is to accept linguistically correct and meaningful responses and to reject incorrect responses. The shared task consists of both a text track and a speech track. The speech track uses the recorded response as an input, and based on this input, a system submitted to this track should accept or reject the response. In contrast, the text track uses text responses generated from two state of the art speech recognition systems (Kaldi \cite{3} and Nuance \cite{4}). Systems participating in the text track can use the output text of either of these speech recognition systems, rather than the originally recorded audio, to accept or reject the response. In this paper we target the text track, where we are supposed to give accept or reject feedback for a text response, based on meaning and language quality.

This paper describes three systems we developed and incorporated in this shared task, where our best system achieved the 3\textsuperscript{rd} position in the final evaluation. Our classifiers were developed by using different machine learning techniques to combine syntactic and semantic features extracted from learner responses. The paper is organized as follows: Section 2 presents related work, Section 3 describes the dataset, Section 4 details our proposed approaches, Section 5 summarizes the experiments, Section 6 presents the results, and Section 7 concludes and suggests future work.

2. Related work

Accepting or rejecting a targeted response based on both language quality and meaning can be seen as a hybrid of two tasks: 1- error detection. 2- similarity measurement. It is not enough to just detect grammatical errors, since a meaningless sentence can be correct grammatically. Conversely, detecting similarity is also not enough, as some tasks require grammatically sound responses. We thus tackle the shared task by combining prior work in error detection and similarity measurement.

Multiple shared tasks have previously targeted the problem of error detection and correction (CONLL 2013 \cite{5}, CONLL 2014 \cite{6}, HOO 2011 \cite{7} and HOO 2012 \cite{8}). Sometimes the task is to only correct errors, while other times the task is divided into error detection and correction. A variety of approaches have been based on classification. Cahill et al. \cite{9} used a classification model based on logistic regression to develop a system for error detection. Cahill et al. mined Wikipedia revisions to produce a large error-annotated training corpus, and evaluated the system using other publicly available datasets (HOO 2011 \cite{7} and FCE \cite{10}). Another system that used a classification-based approach for error detection and correction is the system proposed by Rozovskaya et al. \cite{11, 12}. On the other hand, Gamon \cite{13} argued that using language modeling outperforms classification-based models, and also proposed combining language modeling and classification for error detection and correction. As the task we are dealing with targets English learners, responses are expected to be very basic short English sentences and very limited in vocabulary. Due to these aspects, we decided to follow Gamon and adopt language modeling as the core approach for our error detection. Moreover, we adopted Cahill et al.’s idea of using external error resources.

With respect to detecting text similarity, different approaches have also been introduced. Šarić et al. \cite{14} used a

\textsuperscript{1}https://regulus.unige.ch/spokencallsharedtask/
combination of unigram, bigram and trigram overlap features in addition to other corpus and knowledge-based similarity measures to detect similarity. Magooda et al. [15] combined different word embedding models to measure sentence similarity, where sentences are represented using summation of word vectors and similarity is measured using cosine. In our work we use machine learning to combine features based on this prior work in measuring similarity with features based on the error detection work described above.

3. Dataset

The dataset we used during this work is the dataset introduced by the shared task competition committee, which consists of annotated training data as well as test data in which the annotations were hidden until the end of the competition. Table 1 shows the number of samples in the training and test sets. Three native English speakers independently annotated the data along two dimensions (Language and Meaning). Table 2 and Table 3 show the annotation distribution over both training and testing data, respectively. Each training sample in the training set consists of the following:

- Prompt.
- Response “result of speech recognition”.
- Human transcript of audio response.
- Language Annotation.
- Meaning Annotation.

While a transcript is provided for the training samples, no transcript is provided for the test samples. Without transcription and with errors due to speech recognition, the text track was quite challenging. Another useful resource that was publicly available for use was a sample XML grammar file that has a set of possible correct responses for each of the prompts in either the training or test sets.

<table>
<thead>
<tr>
<th>Data</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>5222</td>
</tr>
<tr>
<td>Test</td>
<td>996</td>
</tr>
</tbody>
</table>

Table 1: Training and testing samples

<table>
<thead>
<tr>
<th>Language</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>3880</td>
<td>802</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>240</td>
</tr>
</tbody>
</table>

Table 2: Training set annotations

<table>
<thead>
<tr>
<th>Language</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>716</td>
<td>159</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>121</td>
</tr>
</tbody>
</table>

Table 3: Test set annotations

4. Proposed approaches

This section details the three systems we submitted to the competition. The systems are based mainly on machine learning techniques and combine a set of syntactic and semantic features. To introduce the systems in detail, we will divide our description into 3 parts: feature extraction, feature selection, and classification.

4.1. Feature extraction

During our system development we extracted a wide set of features, which can be divided into two general types:

- Features used to detect inconsistency on the language level (spelling mistakes, part of speech score, language modeling).
- Features used to detect inconsistency on the meaning level (syntactic relatedness, semantic relatedness).

The intuition behind this separation is mimicking the annotation process. Since samples are annotated based on two criteria, we decided to adopt the same criteria for feature engineering. We employed both the idea of using language modeling [13] combined with the idea of using external error resources [9]. We also adopted some features to detect similarity within sentences [14, 15].

4.1.1. Language-related features

To detect inconsistency on the language level, we use three types of features to differentiate between sound English responses, and responses that have grammatical mistakes or that make no sense regardless of the response’s relatedness to the prompt.

Spelling mistakes (F1). To capture the presence of spelling mistakes, we use a binary feature (which we will refer to as F1) indicating if any spelling mistakes exist in the response or not. To check for spelling mistakes we use the NLTK English spell checker [16].

Part of speech score (F2). A score is given to each sentence by a part of speech (POS) tagger and represents how likely the sentence is to be a real sentence. For the tagging process we use the Stanford part of speech tagger [17].

Language modeling (F3-F10). To extract these features we use multiple language models (LMs). All language models are 5-gram language models trained using the SRILM language modeling toolkit [18]. Since the amount of data we have for this task is not much, it was easy to go beyond trigrams without any noticeable time complexity. We tried multiple values of n for training n-gram models and decided to use 5-grams as they achieved the lowest perplexity over the training data. We also found that going beyond 5-grams didn’t yield any improvement in perplexity. We used interpolation and the unknown tag (“unk”) to deal with out of vocabulary and sparsity issues. While it is not a best practice to evaluate models using training data, we did that due to the lack of data. To distinguish between linguistically sound and unsound responses at an abstract level, we then trained two types of 5-gram language models:

- Models trained using correct English sentences
- Models trained using incorrect English sentences

The intuition behind using these two different types of language models is that incorrect sentences are more likely to get higher probability from language models trained on incorrect
sentences compared to language models trained on correct sentences. The same is plausible for correct sentences, which are more likely to get higher probability from language models trained using correct sentences.

Specifically, within each of these two types, we train four different language models. Two of the four models are trained using words, while the other two are trained using part of speech tagged versions of the same data. For the language models based on correct sentences, word and POS models are trained from the following two sources of data:

- F3, F4: Speaker responses “Transcripts” annotated as correct on the language level.
- F5, F6: All sample responses from the grammar XML. Since these are proposed answers, they are guaranteed to be correct English sentences.

For the language models based on incorrect sentences, the word and POS models are instead trained on the following two data sources:

- F7, F8: Speaker responses “Transcripts” annotated as incorrect on the language level.
- F9, F10: Incorrect sentences collected from two previous error detection and correction shared tasks (HOO 2011 [7] and CONLL 2014 [6]). These shared tasks provided datasets of English sentences with linguistic errors, where participants were asked to detect and propose a correction for these errors. These prior shared tasks yielded a total of 2029 different sentences.

4.1.2. Meaning-related features

The other set of features we use are meant to capture relatedness between the prompt and the speaker response. These features aim to detect both syntactic and semantic relatedness and are extracted using word matching, language modeling, and word embeddings.

**Syntactic relatedness (F11-F12).** To capture syntactic relatedness between prompt and speaker response, we use the sample responses for each prompt that were provided in the grammar XML file and employ the concepts of n-gram matching and language modeling. We decided to use language models trained per prompt, as using such language models should assign high probabilities to similar responses. Moreover, we decided to count the number of n-gram matches as this should also capture a degree of relatedness. In more detail, the syntactic relatedness features we use are as follows:

- F11: Number of matching unigrams, bigrams and trigrams. For a prompt-response pair, we calculate the number of matching unigrams bigrams and trigrams between a speaker response and all the sample responses from the grammar XML file for the prompt of concern. The numbers are then averaged over the number of sample responses.
- F12: Language modeling. For a prompt-response pair, we train a language model using the sample responses from the grammar XML file for the prompt of concern. Once we have a language model trained using the sample responses, we can calculate the probability of the speaker response. The trained language model is expected to assign high probability to responses that are syntactically similar to the sample responses.

**Semantic relatedness (F13-F16).** To capture semantic relatedness, we decided to use the concept of word embeddings. Word embedding is based on representing words with vectors in high dimensional space, where each dimension of the generated space can hold a semantic or a syntactic feature. This high dimensionality representation of words can be utilized to measure semantic relatedness between words or sentences, using a distance measure like cosine. Two very popular models of word embeddings (skip-gram, continuous bag of words) were developed by Mikolov et al. in [19]. These two models have a structure similar to neural networks while using log linear classifiers as the core of the model. The parameters of the trained log linear classifiers are used as word embeddings. Continuous bag of words and skip-gram models are trained differently. The first is trained to predict word given context while the second is trained the other way around to predict context given a pivot word. For a prompt-response pair, we train multiple word embedding models using both skip-gram and continuous bag of words (CBOW) algorithms [20, 19]. We train the models per prompt, and for each prompt we train multiple models over the sample responses for the prompt of concern. In particular, for each prompt we train the following models:

- F13: Skip-gram model (50 dimensions and negative sampling [20])
- F14: Skip-gram model (30 dimensions and negative sampling)
- F15: Continuous bag of words model (50 dimensions and negative sampling)
- F16: Continuous bag of words model (30 dimensions and negative sampling)

We decided to train our models instead of using any of the already trained models like the Google News skip-gram model [19]. Since the data we have uses only basic English vocabulary and short simple sentences, tailoring models only on these data can capture relatedness between words that are specific for this data, which may not be found in news data like the ones used for the Google News model. As the data we have is not much, we trained models with a small number of dimensions (30 and 50 dimensions) compared to the one trained on Google News (300 dimensions). To use each of the word embedding models, cosine similarity is measured between the speaker response vector representation and the vector representation of each of the sample responses. We used cosine to be consistent with the word embeddings training objective. Since these word embeddings are trained to maximize the cosine similarity between classification output and expected word, we think it is better to follow the same training objective. Additionally, following Mikolov et al. [20] in employing the additive compositionality property of the word embeddings, a sentence vector is formed using the summation of the constructing words’ vectors. The final cosine similarity is the maximum of all the cosine values calculated between response and sample responses. In this context, we tried using average cosine, maximum and minimum; selecting maximum cosine got the best results during validation.

**Sentence ratio (F17).** This final feature is neither semantic or syntactic based, this feature is used to make sure responses have a reasonable length. Sentence ratio, is the ratio between the length of speaker response, and the average length of sample responses assigned to that specific prompt. As the prompts are asking for very basic English responses, it is safe to assume that responses are expected to have almost the same length. This feature can capture responses that are not reasonable, responses that are too short or too long.
4.2. Feature selection

After extracting the features we discussed in the previous section, we performed feature selection using PCA to select a subset of the features that can achieve the best accuracy. To select features using PCA, we generated two versions of the features, one version with the raw values, the other with a normalized version of the values to be between [-1, 1]. Tables 4 and 5 show the set of features selected using raw and normalized values, respectively.

<table>
<thead>
<tr>
<th>Features</th>
<th>Language</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part of speech sentence score (F2)</td>
<td>LM trained on correct responses (F3)</td>
<td>LM trained on sample responses (F12)</td>
</tr>
<tr>
<td>LM trained on incorrect responses (F7)</td>
<td>LM trained on incorrect sentences (F9)</td>
<td>Skip-gram 50 dimensions (F13)</td>
</tr>
<tr>
<td>LM trained on incorrect sentences POS (F10)</td>
<td></td>
<td>Sentence ratio (F17)</td>
</tr>
</tbody>
</table>

Table 4: Selected features using PCA with “raw values” (feature set 1)

Table 5: Selected features using PCA with “normalized values” (feature set 2)

<table>
<thead>
<tr>
<th>Features</th>
<th>Language</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM trained on correct responses (F3)</td>
<td>LM trained on incorrect responses (F7)</td>
<td>Unigram, bigram, trigram matches (F11)</td>
</tr>
<tr>
<td>LM trained on incorrect responses (F7)</td>
<td></td>
<td>Skip-gram 30 dimensions (F14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CBOV 50 dimensions (F15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CBOV 30 dimensions (F16)</td>
</tr>
</tbody>
</table>

4.3. Classification

For the classification, we tried two different widely used machine learning classifiers:

- K-nearest neighbor (KNN) [21].
- Support vector machines (SVM) [22].

We trained SVM models using the normalized set of features. SVM works best with normalized data because it avoids prior whitening for the dimensions, which in turn tries to avoid dimension domination. On the other hand we trained KNN models using the raw set of features. KNN depends on pure Euclidean distance between samples, so using raw values can produce a wider spectrum of distances between samples, while normalizing can end up having very small differences in distances due to the tightened values allowed. That is why we decided to use the raw feature values with KNN and normalized versions with SVM. The next section will describe the experiments and tuning carried out to use both classifiers.

5. Experiments

To select the best performing classifier and hyper parameters, we performed multiple experiments to train and validate both classifiers. To train and tune both classifiers, we used 10 fold cross validation over the training data provided by the shared task. However for the 10 folds we used 3 different data splitting paradigms.

- Fixed data splitting, where data is split into 10 parts based on the order of training samples in the file.
- Random data splitting, where data is randomly split into 10 parts. The accuracy of random splitting is calculated by averaging the score of performing 10 iterations, of random 10 fold cross validation.
- Per prompt data splitting, where we split the responses for each prompt into 10 parts. In this paradigm we are making sure that responses for the same prompt appear in training and validation folds.

For KNN we tuned the K (number of neighbors) value, and for SVM we varied the cost, gamma and the kernel. We selected the parameters that maximized the D-scores (equation 1 & 2) over the 3 splitting paradigms.

\[
D = \frac{C_R(F_R + C_A)}{F_R(C_R + F_A)} \quad (1)
\]

\[
F_A = PF_A + k.GF_A \quad (2)
\]

where:

- \( C_A \) = Correct Accept, the student’s answer is correct, the system accepts.
- \( C_R \) = Correct Reject, the student’s answer is incorrect, the system rejects.
- \( PF_A \) = Plain False Accept, the student’s answer is correct in meaning but incorrect English, the system accepts.
- \( GF_A \) = Gross False Accept, the student’s answer is incorrect in meaning, the system accepts.
- \( F_R \) = False Reject, the student’s answer is correct, the system rejects.
- \( k \) = A weighting factor, default value is 3

Table 6: Accuracy over training data using 10 fold cross validation

<table>
<thead>
<tr>
<th>Features</th>
<th>D-Score</th>
<th>Fixed</th>
<th>Random</th>
<th>Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1: Raw</td>
<td>KNN</td>
<td>23.034</td>
<td>18.758</td>
<td>17.981</td>
</tr>
<tr>
<td>Set 2: Normalized</td>
<td>SVM</td>
<td>13.101</td>
<td>18.206</td>
<td>16.491</td>
</tr>
<tr>
<td>Combined: Raw</td>
<td>KNN</td>
<td>26.354</td>
<td>18.683</td>
<td>17.707</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>2.06</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6 summarizes the results obtained by using 10 fold validation on the training data. Raw Features are the features selected by PCA using the raw feature values, while normalized are the features selected using normalized values. The combined features are the feature set selected from normalized values and the feature set selected from raw values, but all values then kept without normalization. We submitted three systems to the shared task competition, where each system corresponds to one of the entries in Table 6. First entry corresponds to using KNN as a classifier with features set 1 “raw values”, entry
2 corresponds to using SVM as a classifier with features set 2 “normalized values”, while finally entry 3 corresponds to using KNN as classifier with features set 1 and set 2 combined without normalization “raw values”. The next section will present the results achieved for each of these submissions over the test set.

6. Results

Out of the submitted 20 entries made by 9 teams, one of the entries we submitted achieved the 3rd position. Table 7 shows the results of the highest 10 submissions, sorted by D-score over the test data. The results are anonymous, however our submissions are underlined in the table. Besides showing the scores, the table shows which speech recognition system output each submission used for training and testing, where custom means that the team used their own developed speech recognition system and participated in the speech track. Thus, beside achieving 3rd in both tracks combined, we achieved 1st place in the text track.

Note that in contrast to the results we achieved on the training data, our submissions came in reverse order on the test data. That is, the system that got the lowest score on training data (Set 2 - Normalized - SVM) got the highest on test and vice versa, the system that got the highest score on training data (Combined - Raw - KNN) got the lowest on test data. One thing that comes to mind here is that KNN classifiers are more liable to overfit than SVM ones. Another finding that actually agrees with our intuition is as follows. Since D-score penalizes meaning errors more than language ones, we thought that using more features to capture meaning errors would get us higher results. This intuition actually proved to be right on test data, but didn’t prove to be correct on training data, which in turn can weight the scale towards hypothesizing there was an overfitting issue for KNN classifiers.

It is also worth mentioning that all the systems that came in the first 10 positions which aren’t using their own developed speech recognition systems are using the Kaldi system, although Nuance’s baseline achieves a marginally higher score than Kaldi’s baseline. Due to this contradiction, after the competition we tried our 3 set of features in further runs, once using Nuance system output, and the other time using output of the speech recognition system developed by the 1st place team. As the 1st team only released the output of their speech recognition system of test data, we are not able to train a new model using their speech recognition system output and instead used the same Kaldi model we trained. In addition, we trained a new model using the Nuance system output. Using these two models we evaluated two versions of the test data, one using the Nuance version of the test data, the other using the 1st team’s version. Table 8 summarizes the results. Comparing to Table 7 we see that using the Nuance system instead of Kaldi really got us lower scores, which is somehow consistent with scores from other teams. We also didn’t get any score enhancement by using the 1st team’s version of test data. This degraded performance is somehow expected, because we are training and testing using two different speech recognition systems. This conclusion agrees with our intuition as we know that training using a specific system tailors the model to avoid the errors the system makes, so testing using another system with a different set of errors will result in degraded performance as in our case.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Speech system</th>
<th>D-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Custom</td>
<td>4.766</td>
</tr>
<tr>
<td>2</td>
<td>Custom</td>
<td>4.710</td>
</tr>
<tr>
<td>3</td>
<td>Kaldi</td>
<td>4.468</td>
</tr>
<tr>
<td>4</td>
<td>Custom</td>
<td>4.371</td>
</tr>
<tr>
<td>5</td>
<td>Kaldi</td>
<td>4.353</td>
</tr>
<tr>
<td>6</td>
<td>Kaldi</td>
<td>4.273</td>
</tr>
<tr>
<td>7</td>
<td>Kaldi</td>
<td>3.998</td>
</tr>
<tr>
<td>8</td>
<td>Kaldi</td>
<td>3.678</td>
</tr>
<tr>
<td>9</td>
<td>Kaldi</td>
<td>3.352</td>
</tr>
<tr>
<td>10</td>
<td>Nuance</td>
<td>2.358</td>
</tr>
<tr>
<td></td>
<td>Kaldi</td>
<td>1.694</td>
</tr>
</tbody>
</table>

Table 8: Accuracy of training and testing using systems other than Kaldi

7. Conclusions and future work

In this paper we presented a system that uses natural language processing techniques over the output of a speech recognition system in order to provide feedback on spoken responses to a spoken CALL system. Our approach to system development used machine learning to combine syntactic and semantic features based primarily on language modeling and word embeddings, which makes it easy to develop. We provided three different configurations of the system with three different set of features; these systems were also part of the SpokenCALL shared task and our best configuration achieved the 3rd position in both tracks and 1st position in text track only. This result suggests that with a combination of simple and easy to develop features and a basic machine learning classification model, promis-
ing performance can be achieved. Using the data released by the shared task after the competition, we further evaluated our method by training using the output of both Kaldi and Nuance speech recognition systems. However, we hope that the first place team will make the training version of their speech recognition system available for use by others. We would like to try our same pipeline using this data as we think that better results can be achieved by enhancing the speech recognition accuracy. We would also like to further investigate the interplay of speech recognition output, natural language processing features, and machine learning algorithms.

8. Acknowledgements

We would like to thank A. Mahgoub and M. Zahran for their helpful feedback.

9. References


A SYSTEM FOR ASSESSING CHILDREN READINGS AT SCHOOL

Stefano Artuso(1,2), Luca Cristoforetti(2), Daniele Falavigna(2), Roberto Gretter(2), Nadia Mana(2), Gianluca Schiavo(2)

(1) Pedius Srl, via Boezio 4C, Roma (Italy)
(2) Center for Information and Communication Technology Fondazione Bruno Kessler, via Sommarive 18, Trento (Italy)
{artuso,cristofo,gretter,mana,gschiavo,falavi}@fbk.eu

Abstract

In this paper we describe a system for analyzing the reading errors made by children of the primary and middle schools. To assess the reading skills of children in terms of reading accuracy and speed, a standard reading achievement test, developed by educational psychologists and named “Prove MT” (MT reading test), is used in the Italian schools. This test is based on a set of texts specific for different ages, from 7 to 13 years old. At present, during the test, children are asked to read aloud short stories, while teachers manually write down the reading errors on a sheet and then compute a total score based on several measures, such as duration of the whole reading, number of read syllables per second, number and type of errors, etc. The system we have developed is aimed to support the teachers in this task by automatically detecting the reading errors and estimating the needed measures. To do this we use an automatic speech-to-text transcription system that employs a language model (LM) trained over the texts containing the stories to read. In addition, we embed in the LM an error model that allows to take into account typical reading errors, mostly consisting in pronunciation errors, substitutions of syllables or words, word truncation, etc. To evaluate the performance of our system we collected 20 audio recordings, uttered by 8-13 years old children, reading a novel belonging to “Prove MT” set. It is worth mentioning that the error model proposed in this paper for assessing the reading capabilities of children performs closely to an “oracle” error model obtained from manual transcriptions of the readings themselves.

Index Terms: language learning, children’s speech recognition, reading assessment.

1. Introduction

The usage of automatic speech recognition (ASR) systems in language learning applications is not a novelty, although in the past most of the research focused on the development of approaches and tools to aid the learning of the second language [1, 2, 3, 4, 5]. These studies and research resulted into a large number of current commercial products, implementing software tools for computer-assisted language learning (CALL) and assessment, that extensively use ASR technology, mostly for detecting pronunciation errors and for giving feedback to the learner about the quality of her/his pronunciation in the target language. The usage of these products has been largely debated in the language teaching community giving rise, especially in the past, to skepticism [6, 7] by language teachers to accept the results furnished by ASR systems. However, it is a matter of fact that, due to the recent impressive improvement of speech recognition performance on a large set of languages, ASR has become a strategic component in quite all CALL products. Finally, the huge increase in the usage of the Internet and Web based applications has determined an explosion of online courses for second language (L2) learning, even tuned for specific domains (e.g. business, retail, science, etc), that employ ASR as an essential feature.

Despite the large diffusion of the above mentioned L2 CALL tools, little work has been done till now for developing ASR tools for helping language learning by children at school. This is probably due to the fact that commercial ASR products do not perform well with children voice. In fact, it is well known that spectral and temporal characteristics of children’s speech are highly influenced by the anatomic, physiological and developmental changes that occur during the growth and are hence different from those of adult speakers [8, 9, 10]. Therefore, when an ASR system, trained on adults’ speech, is employed to recognize children’s speech, performance decreases drastically, especially for younger children [11, 12, 13, 14, 15, 16, 17, 18, 19].

To compensate for this behavior, assuming having at disposal a small amount of data for training an ASR system for children, we proposed in the past some approaches [18, 20] for adapting the acoustic models of an ASR system trained on adult data to children’s speech. In these papers we demonstrated that the acoustic hidden Markov models (HMMs) adapted to children’s voices are also effective for large vocabulary speech recognition applications.

Starting from that experience we decided to investigate an application of children’s ASR in the language learning domain. More precisely, we developed a tool for both analyzing and assessing the reading errors made by children of Italian primary and middle schools when they are examined for “Prove MT” (MT Reading Test) [21]. This is a standardized reading achievement test created by experts on diagnosis and treatment of learning disorders through a study involving over 8000 students from schools of all over Italy. The test is commonly used for the assessment of reading speed and accuracy in different school grades. Similar text-reading tests in other languages are the GORT-5 [22] for English, the Alouette test [23] for French and the SLRT II [24] for German. MT Reading Test procedure involves asking children, aged between 7 to 13, to read aloud short stories, each formed by some hundreds of words. During the readings, the teacher manually takes note of both reading errors and reading time. Basically, two parameters are computed from teacher’s notes: (i) reading speed, measured in syllables per second, and (ii) accuracy of text reading, measured as the number of errors (e.g.: skipped lines, replaces, omits, adds or reverses letters, makes up words, pauses and hesitations). These
measures are compared to standard reference data and finally used to produce historical records of the examined students, from which to derive statistics specific either of each student or of entire classrooms.

The system we have developed aims to support the work of the teachers both by automatically detecting the reading errors and by estimating the parameters they usually measure during the reading itself. To do this we use an automatic speech-to-text transcription system that employs a LM trained over the texts containing the stories to read. In addition, we embed in the LM an error model that allows to take into account typical reading errors, mostly consisting in words truncation. Finally, we propose a measure, to score each reading, that takes into account both the fluency of the reading itself, basically based on the total duration of the “erroneous” pauses, and the total number of automatically detected errors.

This paper is organized as follows: Section 2 introduces the topic, discussing the main approaches. Section 3 describes the whole architecture of the system focusing on the user interface (see Section 3.1), Section 4 gives the details of the ASR system, and experiments and results are reported and discussed in Section 5. Finally, Section 6 draws some conclusions and discusses directions for future work.

2. Related works

Numerous studies have demonstrated the effectiveness of support technology on a variety of skills related to reading, including vocabulary, phonemic awareness, reading fluency, speed and comprehension [25]. Several applications have been developed not only to support the reading process but also to provide automatic assessment of oral reading by estimating errors [26], mispronunciations [27] or reading (dis)fluency [28].

Recently, the use of the hybrid deep neural network (DNN)-HMMs has been shown effective for children’s speech recognition [29, 30, 17, 18]. However, a shortage of training data is typical issue when developing ASR systems for children. To cope with this problem, in [18] an age adaptation approach is investigated in which a DNN-HMMs system is first trained on a substantial amount of adults’ speech integrated with a small amount (few hours) of children’s data. The resulting acoustic model is then adapted to children’s voices by exploiting the available children’s training data. In a noticeable work [17], a huge amount of children’s speech data, more than 1,000 hours, was collected from the Internet and used to train a DNN-HMMs based recognition system with excellent recognition results. Furthermore, it was shown that training an hybrid DNN-HMMs system using balanced large amounts of speech from adults (more than 1,000 hours) and children (more than 1,000 hours) leads to recognition performance equal or better than training specific acoustic models for adults and children. Contrastive experiments were conducted using different neural network models and configurations, confirming these results. However, such large amount of children’s speech is rarely available for training purposes especially for languages different than English.

In [20] we described an approach, based on objective function regularization [31], for adapting to children a DNN trained on around 300 hours of voices of adult speakers. We showed its effectiveness even when the size of adaptation data is quite small (i.e. less than one hour speech). Moreover, we observed that the ASR performance does not deteriorate too much also when manual supervision of the adaptation data is not available (i.e. the supervision of the data is generated by an ASR).

In [30] a maximum likelihood linear regression (MLLR)-

based speaker adaptive training (SAT) DNN-HMMs system trained on a small amount, 7 hours, of children’s speech was shown effective in a large vocabulary continuous speech recognition (LVCSR) task. Similar results were also achieved in experiments reported in both [18] and [20] therefore, for this work, we have decided to use the ASR system described in [20] (see section 4), which was trained on the above mentioned small children data set (around 7 hours of speech). Moreover, since this work is aimed at detecting errors in children readings, given a “small” set of “short” stories (whose texts are known), the size of the dictionary to recognize results quiet small (a few hundreds of words), thus reducing the impact of the quality of the acoustic model on the ASR system performance. Instead, it is crucial for this language learning task to effectively model the errors made by children during their readings, measuring the ability of the whole system to correctly detect them.

3. System architecture

Figure 1 illustrates the whole client-server architecture of the system developed for analyzing children’s readings. It is formed by several modules distributed along both client and server sides.

On client side, a web browser acquires audio files either from the microphone on the device or reading it from file. Audio files (note that more than one single audio file can be sent to the server) are sent to a web server, built up with Node.js® framework [32], through a websocket connection provided by a “Node” module called SOCKET.IO [33]. Over the websocket protocol, client and server exchange messages and data in a standardized and secure way, facilitating real-time data transfer. The Node.js® server retrieves the audio files sent by the user and creates a job containing all the audio files and their related information, like the corresponding reference texts. The auto-
matic transcription related to a submitted job, obtained through the KALDI toolkit (see section 4), is aligned with the reference transcription of the corresponding audio recording and an HTML page is produced, as will be explained in section 3.1. This page shows, in a single shot, the detected reading errors, the duration of the erroneous pauses (i.e. the pauses that are not correlated with punctuation) and the score gained in the reading.

3.1. User interface

For a given reading, the system will produce an HTML page containing all the information needed for its evaluation. This information is obtained by means of the following processing steps:

- ASR processing, furnishing a file which contains, in each row, a recognized word with the corresponding time boundaries information;
- alignment between the ASR output and the normalized reference text of the read novel, including timing information;
- computation of two scores: taking into account errors in the word pronunciation and “fluency” in reading (measured using the durations of the pauses), respectively.

The generated HTML page (see Figure 2) is a picture that represents, at a glance, the errors made by the child who read the assigned “Prove MT” text. In the Figure, colors other than blue highlight inserted or missed words while pauses, indicated by underscores, are blue or red depending on whether they correspond to some punctuation or not. In this way, the teacher can have an immediate overview of the reading quality and can easily locate where reading difficulties may have occurred. Furthermore, teachers can listen to the audio and, by clicking over a selected word, they can place the audio pointer to that word. At top of HTML page, the scores achieved in the reading are displayed. At present, each reading is assigned an overall score which is the arithmetic mean between two scores, called: “Rhythm Score” (RS) and “Word Score” (WS), respectively. Rhythm Score takes into account: words durations (WD), duration of “good pauses” (GPD, i.e. segments of silences corresponding to punctuation in the text to read) and duration of hesitations (HD, i.e. segments of silences occurring although there is no punctuation to follow in the text). RS is defined as follows:

$$RS = 10 \times \frac{WD + GPD}{WD + GPD + HD}$$

(1)

Word score takes into account the number of words correctly recognized (CRW) and the number of words wrongly recognized (WRW, which sums both insertions and substitutions errors). WS is defined by the following equation:

$$WS = 10 \times \frac{CRW}{CRW + WRW}$$

(2)


Finally, several statistics can be computed from the whole set of collected records, relatively to the single reader, or in aggregated forms, e.g. referring to: age, gender, classroom, etc.

4. ASR system

The ASR system is based on the KALDI open source software toolkit [34]. The latter, largely used to develop state-of-the-art ASR systems for a variety of applications, integrates hybrid DNN-HMMs in a static decoding network built by means of finite state transducers. Context dependent HMMs with tied states, speaker adaptive training via MLLR transformations [35], linear discriminant transformation of acoustic observations are some, among the many other features, furnished by the toolkit.

4.1. Acoustic models

For HMMs training, 13 mel-frequency cepstral coefficients (MFCCs) are computed every 10 ms by using a Hamming window of 25 ms length. These features are mean/variance normalized on a speaker-by-speaker basis, spliced by +/- 3 frames next to the center frame and projected down to 40 dimensions using linear discriminant analysis (LDA) and a single maximum likelihood linear transformation (MLLT). Then, a single transform is estimated for each training speaker and applied for normalizing the utterance. Clicking over a blue word locates the audio player to that position, allowing to listen the corresponding part of the reading.

![Figure 2: Output produced by the developed system: the upper Figure corresponds to a "quite good" reading, except the last sentence where there are some hesitations. On the contrary, the lower Figure corresponds to a worse reading: in fact, the number of red pauses and words gives an immediate view of the errors made by the child. Clicking over a blue word locates the audio player to that position, allowing to listen the corresponding part of the reading.](image-url)
ing literature for children, depending on his/her grade. Each speaker read a different set of sentences which included, however, a set of phonetically rich sentences (5-8 sentences) which were repeated by several speakers. Speech was acquired at 16 kHz, with 16 bit accuracy, using a Shure SM10A head-worn microphone. The number of utterances in the training set is 7,020, their total duration is 7h:16m.

First, triphone HMMs with gaussian mixture model (GMM) output densities are trained and used to align acoustic observations with tied HMMs states, obtained by means of a phonetic decision tree. Then, a DNN with output nodes univocally associated to tied HMMs states is trained using the resulting alignment. To do this, an eleven frames context window of LDA+MLLT+MLLR features (5 frames at each side of the current frame) form a 440 dimensional input feature vector for the DNN. This is trained in several stages [36], including: i) pre-training consisting in layer-wise training of Restricted Boltzmann Machines (RBM) by Contrastive Divergence algorithm; ii) frame classification training based on mini-batch Stochastic Gradient Descent (SGD), optimizing frame cross-entropy; iii) sequence discriminative training [37] consisting in SGD with per-sentence updates, optimizing state Minimum Bayes Risk (sMBR). Outputs of hidden layers are transformed by sigmoid functions, while softmax normalization is applied to the output layer. The DNN has 4 hidden layers each with 1536 neurons and 2410 output nodes (i.e. the same number of HMMs tied states).

During decoding the computation of MLLR features is done in two steps. First, a word lattice is produced for each input utterance by using the baseline speaker-independent GMM-HMM. A single MLLR transform for each speaker is then estimated from sufficient statistics collected from word lattices with respect to speaker-adaptively trained triphone HMMs. These transforms are used with SAT triphone HMMs to produce new word lattices. A second set of fMLLR transforms is estimated from new word lattices and combined with the first set of transforms. The resulting transforms are used for features normalization. Finally, output probabilities of HMMs states are computed by dividing the DNN output node posterior probabilities by their corresponding prior probabilities, as explained in [36].

4.2. Language models

As mentioned above, the system we developed for assessing the correctness of readings requires the prior knowledge of the reference texts to read, while the readings errors have to be detected.

In the experiments reported below we are given 4 different Italian novels (namely: “The seven kings of Rome”, “Old proverbs”, “The full barrel and the empty barrel“, “The Etruscan sovereigns”). These novels are read by children during their “Prove MT” trials.

First, the texts of all the 4 stories are normalized, i.e.: words are lowercased, punctuation is removed, numbers and acronyms are expanded - e.g.: VI → sexto (sixth); A.C. → avanti cristo (Before Christ). Compound words are split (e.g. romano sabini → romano sabini), special characters are normalized using the “Latin-1” encoding. The total number of words in the 4 stories is 606, the dictionary size is 332, the number of unigrams, bigrams and trigrams resulting after LM training is: 332, 594 and 12, respectively.

Starting from the normalized texts of the 4 stories, we trained the following three different 3-gram LMs with the IRSTLM open source toolkit [38].

• Text To Read (TTR): only the normalized texts are used to train the LM, without including any error model.

• Automatic Error Model (AEM): the previous document text (TTR) is augmented by corrupting it with an error model which includes all possible word beginnings, on a syllable-like basis, due to false starts. For instance, the words vecchi and proverbio can give raise to the following false start words: ve- pro- and prove- that are inserted in the text before the corresponding full words, as shown in Table 1. This approach is a simple, but effective, way to automatically model typical readings errors made by children when they encounter a “not-so-easy” word to read.

• Leave-One-Out (LOO): the previous document text (AEM) is augmented by including in it the manual transcriptions of all the readings made by the children in the evaluation set (see Section 5), except that of the child being evaluated. This is a practical way to insert in the model common but not predictable errors, like for example mispronunciations of uncommon words or names (for instance Tarquinio Prisco often becomes Tarquinio Parisco, proverbio becomes proverbio, etc).

Table 1 reports some samples of the texts used to train the different LMs.

Table 1: Texts used to train the LMs. TTR contain the plain reference texts only; AEM expands TTR with automatically generated syllables derived from word beginnings, LOO expands AEM with manual transcriptions of the utterances in the test set, in a leave-one-out fashion. Pronunciation errors are highlighted in bold.

<table>
<thead>
<tr>
<th></th>
<th>TTR</th>
<th>AEM</th>
<th>LOO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>per la sorpresa e l’amarezza il vecchio proverbio …</td>
<td>pe- per la so- sorpresa e l’ ama- amarezza il ve- vecchio proverbio</td>
<td>per la sorpresa e l’ amarezza del vecchio proverbio …</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pro- proverbio …</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>per la sorpre- sorpresa e l’ amare- amarezza il vecchio prove- proverbio</td>
<td>per la sorpre- sorpresa e l’ amare- amarezza il vecchio proverbio …</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Experiments and results

Experiments were conducted on a data set of readings containing the recordings of the novel “old proverbs” uttered by 20 children, aged between 8 and 13 years. Note that, although the total number of children involved in our experimentation is 20, not all of them read the 4 previously mentioned novels. All the readings were carefully transcribed manually, annotating: words, words fragments and hesitations. Since the text of the novel “old proverbs” consists of 242 words, the number of running words in the test set results to be $20 \times 242 = 4840$.

Performance was computed using two different references:
• **Ideal Reference Text (IRT):** it is the normalized text of the novel. This is the reference to consider when preparing the output of the system for the teachers. Comparing the system output against IRT allows to compute a score that can be used to measure the correctness of the reading. This score should be as similar as possible to the one given by the teacher.

• **Manual Transcription Reference (MTR):** it is the orthographic transcription of the reading, and includes all errors made by the pupil. This is the reference to use to estimate the performance of the ASR system. Typically, a better error model should return a lower word error rate (WER). Note that the MTR texts are the same used to train the LLO LMs defined in Section 4.2. The total number of running words in MTR texts is 5135.

Table 2 reports the performance achieved on the test set described above, in terms of WER, and using as reference the two previously mentioned transcriptions: IRT and MTR.

The upper two blocks report performance achieved using IRT as reference. The first raw (MTR) gives the WER (together with the percentage of insertions, deletions and substitutions) obtained when the same MTR texts (i.e. the reference manual transcriptions) are aligned with the IRT reference. This value provides a sort of WER oracle (13.90%), since it represents the exact number of reading errors made by children (i.e. we are assuming that the ASR system doesn’t make errors). In the second block the 3 LMs defined in section 4.2 are used by the ASR system and allow to score each reading in a completely automatic way. It can be noticed that the proposed error models (both AEM and LOO) give better performance than the baseline LM (TTR, trained without using any error model), and in particular LOO performs closely to the oracle (13.95% versus 13.90%). It is worth noting (see also Table 3) that, in case of hesitation errors, in most cases every LM detects some errors. But while TTR can only insert existing words, AEM and LOO can insert word fragments, often corresponding to the exact error.

Finally, the lower block in Table 2 reports the results when the ASR outputs are aligned with the MTR reference. A perfect ASR system - that knows in advance all possible reading errors and that never makes mistakes - should return 0% WER. In practice, the more accurate the system, the lower the WER obtained. It is important to notice that detecting a word fragment which is similar, but not the same, to the one really pronounced, results in a recognition error. For instance, see hesitation sog- in rows 7 of Table 3: it is an error for TTR (recognized as another word SUD, South) and AEM (recognized as the word fragment SAGGIA-); only LOO could detect it correctly as sog- because some other children did the same hesitation.

### 6. Conclusions

In this paper we have presented an approach for automatically assessing the reading capabilities of children. The developed system makes use of the time information provided by an ASR system for estimating the durations of the pauses uttered during readings, detecting those corresponding to hesitations. In addition, different LMs have been trained, including in their training data fragment words representing typical reading errors made by children. ASR performance, depending on each LM, has been computed on a manually transcribed test set and effectiveness of the proposed error models has been demonstrated.

Future works include the possibility by the teachers to correct errors made by the automatic system. This will allow both to regenerate the HTML page and present to the pupils/parents a more accurate score, and in the long run to obtain more reliable data to be used to retrain and improve the system itself. Furthermore, a focus group with teachers will be organized to better investigate their expectations from the system in term of possible functionalities to be included, as well as to collect their opinions about the current user interface. Finally, a user study will be conducted to evaluate the interface usability and to assess system performance. These activities will open the way for a real usage of the system in the primary school.
7. References


Automatic Assessment of Children’s L2 Reading for Accuracy and Fluency

Kamini Sabu, Prakhar Swarup, Hitesh Tulsiani, Preeti Rao

Indian Institute of Technology, Bombay
{kaminisabu,prkhr,hitesh26,prao}@ee.iitb.ac.in

Abstract
This project targets using state-of-the-art in automatic speech recognition technology, coupled with new work in predicting the relevant prosody ratings, to build an oral reading assessment tool. A reliable automatic system can prove invaluable in helping children acquire basic reading skills apart from facilitating the monitoring of literacy programs at large scale. In the present work, we target middle-school learners of English as a second language in a rural Indian setting. We present the design and observed characteristics of our field-collected oral reading dataset to outline the research challenges faced. Recently proposed solutions to the training of robust acoustic models in the face of limited task specific data are evaluated for the prediction of the child’s word decoding accuracy and for achieved word-level alignments for prosody scoring. A language model is designed to exploit the known text and observed reading errors while being flexible enough to adapt to new reading material without further training. Based on a scoring rubric proposed by a national mission on literacy assessment in India, we present an automatic system that detects reading miscues and computes fluency indicators at the sentence level which are then correlated with fine-grained subjective ratings by an expert.

Index Terms: speech recognition, prosody, children reading, human-computer interaction

1. Introduction

Recent exercises on cross-country assessment of basic literacy and numeracy skills across primary and middle-school students in India have revealed many disquieting facts such as the low proportion of students who can even meet the expectations of their grade-level [1]. Consistent with the finding that nearly 75% of fifth grade students cannot read second grade level texts in rural India, several parts of the country continue to be vulnerable to high school-dropout rates. Annual surveys undertaken by ASER [1] to measure the literacy level of school children follow a prescribed protocol executed by project volunteers visiting schools across rural India. Using prepared texts of letter and word lists, paragraphs and stories in the selected language, they categorize each student in one of the 5 competency levels, viz. ‘story’, ‘paragraph’, ‘word’, ‘letter’ and ‘nothing’ based on the observed word decoding accuracy and reading fluency. For example, word-level mistakes corresponding to common mispronunciations are ignored, and only 3 or more mistakes in a paragraph reading would disqualify a child at the corresponding competency level. The assessment of fluency is carried out on the basis of whether the student reads the words in sentence-style rather than like a “string of words”. Testing is carried out in the regional language (medium of instruction in the school) as well as in English due to the high demand for English as a second language. We can easily see that any automation of the reading assessment can contribute greatly to the efficiency and scale of the literacy testing procedure just described. Apart from this, a system which can automatically evaluate the reading skill of a child in terms of word decoding accuracy, fluency and comprehension, also providing feedback, can alleviate the root cause of the low literacy problem - the dismal teacher to student ratio in rural schools.

There has been a fair amount of research on developing automated systems for reading evaluation and feedback using Automatic Speech Recognition (ASR) technology mainly by specific research groups contributing over the years [2], [3], [4], [5]. Black et al. [2] focused on the reading evaluation of isolated word lists for pre-school children. Other groups [3], [6], [7], [4] have addressed the task of evaluating read sentences in story context, sometimes including tracking the child in real-time. The ASR technology used has traditionally been based on GMM-HMM acoustic models. More recently, Deep Neural Network (DNN) based acoustic models have demonstrated superior performances, especially with the availability of large training datasets [8], [9], [10]. In the context of reading assessment, the language modeling (LM) required in the ASR framework typically makes use of the fact that the text to be read is known a priori [3], [7]. Either N-gram LMs trained on the story text [4], or, alternately, task-specific sentence-level LMs [3] with appropriate parallel paths modeling specific miscues are used.

Much of the research on technology for reading assessment for children has focused on detecting reading miscues, i.e. word decoding errors. Reading fluency, on the other hand, is indicated by the prosody of speech rather than by misread words [11]. Prosody refers to the supra-segmental aspects of speech. It is linked to the smooth delivery of sentences with appropriate chunking into phrases and the proper marking of word prominence. Good prosody has been associated with successful comprehension [12]. Duong et al. [13] compute pitch, intensity, duration and latency contours for each sentence as sequence of average values per word. These contours of children’s speech are correlated with corresponding contours of adult speech to assess reading. To automate the scoring for spontaneous reading by non-native children of age greater than 8 years, [14] tries to combine aspects such as fluency, pronunciation, vocabulary and grammar using various features based on - 1)silence and reading speed, 2)acoustic model score, 3)part-of-speech tags, and 4)number of idioms and/or meaningful phrases. In FLORA [15], for assessment of expressive children reading at paragraph (1-minute reading) level, an SVM classifier is trained on an annotated corpus of children’s read speech using lexical and filled-pause based features with prosodic features like pitch and duration.

In the present work, we consider the automation of assessment, at the sentence level, for a story reading task based roughly on the scoring rubric provided by ASER [1]. We would like to detect reading miscues such as target words that are missed or incorrectly uttered, as well as the fluency in delivery.

1Authors acknowledge support by MeitY, Govt. of India.
The reading miscues would ideally be based on the outcome of ASR based decoding which provides a sequence of hypothesized words best matching the acoustic utterance. A comparison of the target sentence text with the hypothesized words provides an immediate means to detect “mistakes” depending on whatever definition of mistake we might choose. On the other hand, prosodic cues extracted from the decoded word segments can help us identify whether an utterance was read “like a sentence”. Proper phrasing, including sentence ending, and correct word prominences contribute to the perception of a meaningful sentence form.

In the next section, we present the specific scenario of interest, the ongoing field-data collection and manual labeling methodologies. We discuss the characteristics of the data relevant to the reading assessment task. Following this is a review of the ASR framework with a discussion of the acoustic and language modeling aspects. Next, we discuss the acoustic correlates of the relevant prosodic cues and their implementation. Finally, we evaluate our system in terms of its performance in predicting reading miscues and sentence level fluency indicators as obtained by manual ratings.

2. Database Design and Characterization

Our evaluation data is obtained from a rural middle-school setting similar to our final target scenario. We restrict ourselves to English as a second language, selecting suitable texts from readily available animated video stories with text subtitling [16]. In this section, we present the methodology for data collection and labeling required for the development, tuning and testing of system performance. We also discuss the characteristics of our data to highlight our system design challenges and choices.

2.1. Data collection and ground-truth labeling

Encouraged by the potential of the project, a school in the tribal belt of Western India, where a dialect of Marathi is the native tongue, permitted us to organize oral reading sessions for students of grades 5-8 (aged between 10 to 14 years) as a scheduled activity in school hours [17]. An Android application presents the story in video mode and the child can read and record on a tablet with a headset microphone. The stories are displayed in animated karaoke style, one sentence on each screen with the corresponding subtitles highlighted word-wise at a normal reading pace. The tablet application also provides a “listen” mode where the child can hear a narrator read along with the subtitles in a standard Indian English accent [18], [19]. The children are encouraged to listen to the narrator audio before their own recording session.

The sentence level segmentation of the recorded audio is implemented using the karaoke video time-stamps. Manual word-level transcription of each sentence-level audio recording is carried out using a web-based GUI that facilitates marking each displayed target word as one of correct/missed/incorrect. ‘Incorrect’ words correspond to mispronunciations or substitutions, which are transcribed in terms of phoneme sequences. If the utterance/word is gibberish or otherwise indecipherable, it is not transcribed. Hesitations or sound-outs before a word are separately marked by a ‘Disfluency before word’ label. Apart from word-level realizations, we also label the noise, if any, that is audible into broad categories such as environmental sounds, breath and microphone noise. The second stage of sentence-level annotation characterizes the speech delivery aspect in terms of the indicative prosodic events related to phrasing and prominence. Phrasing is rated on a 3-level scale as follows: the absence of chunking (1), some attempts at chunking (2), and the correct grouping of words into phrases (3). Sentence ending is marked separately as not realized (1) or realized correctly (2). Finally, prominence is rated per sentence using 3 levels: absence of prominence (1), some perceivable word-level prominence (2), and prominence realized on the correct target words (3). This relatively fine-grained marking of prosodic events was found to reduce the subjectivity in labeling to a great extent. We therefore relied on the ratings of a single English teacher, checked for consistency in randomly chosen samples by one of the authors.

2.2. Dataset characteristics

While we try to ensure that there is no immediate source of noise in the vicinity of the child who is recording, it is difficult to control the more distant noises such as children playing and falling rain. For the present study, we consider the labeled subset (about 20% of our overall field-collected data) that is relatively free of background noise, disfluencies and untranscribed gibberish; this comprises of 7 stories read by 68 distinct speakers. We separate the data into groups based on story; then a subset of 3 stories, considered the “test dataset”, is set aside for all the system evaluation reported in this paper. The distribution of ground-truth word-level labels in this test dataset (1371 utterances across 52 speakers spanning a duration of 64 minutes) is shown in Table 1 where the substituted words are further partitioned into (i) predictable (i.e. acceptable) substitutions such as common mispronunciations and word inflections, and (ii) out-of-vocabulary (OOV) substitutions. We note that close to 30% of the target words are incorrectly read (i.e. either missed or substituted). Among the predictable substitutions, widely observed substitutions were grapheme to phoneme errors and replacement of English (L2) phones by native language, i.e. Marathi (L1), phones. Observing the mispronunciations at word-level, we found that the children rarely substituted the common function words. It was also observed that unfamiliar content words were more often replaced with OOV words whereas the common content words were more likely to be replaced by inflected forms when misread.

![Table 1: Data characterization in terms of observed miscues for the test set of 1371 utterances comprising 8429 words by 52 speakers across 3 stories](image_url)

The remaining subset, comprising 978 utterances across 4 stories by 38 speakers (spanning a total duration of 58 minutes), is used as task-specific adaptation data for the ASR acoustic models. This dataset serves to tune the ASR acoustic models trained on more general data to the target population speech with its specific L1 influence. This data is partitioned in a suitable manner to ensure that the reported results always correspond to both speakers and stories being non-overlapping in the adaptation and test sets. Prosody-based evaluation is reported on a subset of the test dataset described in Table 1 derived as follows. We discard story title-author utterances and further consider those utterances of the remaining sentences that are devoid of omissions.
(substituted words allowed). Apart from this, we reject text sentences with less than 10 utterances in the dataset. This gives us a total of 688 utterances across 40 unique sentences. Of these, 19 sentences comprise 2 or more phrases such that phrasal breaks can be uniquely specified; further there are two question forms, one a Wh-question, and the other yes/no. The 688 utterances (total duration of 30 min) come from 52 speakers giving an average of 12 utterances per speaker. Table 2 displays the distribution of subjective ratings for each of the 3 prosodic events across the set of 688 utterances. Not all the attributes are rateable for all utterances, e.g. prominence for list-form reading. Sentence endings are not rated for wrongly segmented utterances. We note that a reasonable representation of the different rating levels is available in our data. We further observe that while most sentence endings are realized correctly, improper phrasing (i.e. rating levels 1 or 2) is observed in 32% of the cases. Word prominence is usually not realized at all (rating 1) or placed on the wrong words (rating 2). We expect students to give prominence on same words as narrator. It was observed that sentences that ended with a prominent word were the most prosodically challenging for the children.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Phrasing</th>
<th>Prominence</th>
<th>Sentence Ending</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>86 (13%)</td>
<td>261 (40%)</td>
<td>98 (16%)</td>
</tr>
<tr>
<td>2</td>
<td>124 (19%)</td>
<td>295 (46%)</td>
<td>511 (84%)</td>
</tr>
<tr>
<td>3</td>
<td>445 (68%)</td>
<td>90 (14%)</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>655</td>
<td>646</td>
<td>609</td>
</tr>
</tbody>
</table>

Table 2: Distribution of prosody ratings for the rateable utterances (out of 688) of 40 unique sentences by 52 speakers

3. The ASR Framework

3.1. Acoustic Model Training Data
An important predictor of ASR performance is the quality of training data in terms of how well it represents the expected test data. Given the paucity of usable field data in the present project, we present the considerations that have gone into creating suitable training data by more easily available means. Since the test data is L1-influenced Indian English, we need acoustic models representing both L1 and L2 phones. This motivated the use of a phonetic inventory of 47 Hindi and English phones. Hindi is chosen due to the availability of a Hindi dictionary and its phonetic overlap with several Indo-Aryan languages including Marathi. Considering that our target population is children, we recorded a variety of English and Hindi text (200 distinct phonetically rich sentences drawn from middle-school level material) read by 30 fluent English and 11 fluent Hindi speakers in the age group of 10-14 years who are students in an urban school. A total duration of 5.27 hours of speech was transcribed with minimal effort by rejecting utterances inconsistent with the text in any way (as in [20]). The training speakers are not a completely homogeneous set due to the relatively large age range apart from other speaker dependencies [21]. Most importantly, the test data from the rural school children differs from the training data in both fluency and accent.

3.2. Acoustic Modeling
Due to their capacity for highly nonlinear classification, Deep Neural Network (DNN) based acoustic models have been shown to outperform conventional Gaussian Mixture Model (GMM) based acoustic models on many speech recognition tasks [22]. Two distinct modes in which DNNs are used for acoustic modeling are the Hybrid and the Tandem mode. In the latter, a deep network serves as a nonlinear feature extractor feeding into a conventional GMM-HMM back-end. When speaker-normalized MFCC features are concatenated with the extracted “bottleneck” features, and all features are speaker-normalized by SAT, the resulting Tandem SAT matches or exceeds the performance of a Hybrid SAT system [23], [24], [25]. The former further allows an additional beneficial stage of model adaptation with task-specific data due to its GMM-HMM back-end. The training procedure for the DNN Tandem SAT models is provided below [25].

1. A baseline SAT GMM-HMM system was trained by estimating fMLLR transforms [26] for each speaker in the training set using raw MFCC features.
2. Speaker-normalized features were generated by transforming each speaker’s data through the estimated fMLLR matrix.
3. A DNN with a 40 dimensional bottleneck(BN) layer was trained using these speaker-normalized features. This was used to obtain lower-dimensional discriminative features from the speaker-normalized features.
4. These 40 dimensional BN features were appended with the speaker-normalized features obtained in Step 2 to get “DNN Tandem SAT” features.
5. These features were used to train a SAT GMM-HMM system using the same procedure as in Step 1.

Finally, we use MAP adaptation of the GMM means in our DNN Tandem SAT model [27] with the task-specific adaptation described in Section 2.2. The GMM-HMM model had 1000 context-dependent tied HMM states with 8000 Gaussians shared across them. A single global speaker-specific fMLLR transform is estimated for each speaker in the training set. The decoding process used is the standard two-pass unsupervised fMLLR decoding process which uses the first pass decoding hypothesis of the SAT GMM-HMM system as the transcription labels for speaker-specific transform estimation. During decoding we estimate fMLLR transform at the speaker-story level. The DNN architecture used for extracting BN features consists of 6 hidden layers with 1024 neurons each with the penultimate hidden layer replaced by a 40-dimensional layer to lower dimensionality. It is trained using standard cross entropy loss function on 40-dimensional speaker-normalized LDA-MFCC features with the context of +/-5 frames. All hidden neurons use the sigmoid activation function.

3.3. Language Modeling
To identify word-level miscues, the ASR decoder hypothesis for an utterance must serve to indicate whether each of the text words is omitted or uttered incorrectly. Since we consider only omission and unpredictable substitutions (referred to as OOV substitutions) of a word as mistakes, we would like to discriminate between OOV substitutions and predictable substitutions. A good language model would capture all the expected variations in an utterance from the known text with the appropriate probabilities. To achieve our aim of detecting miscues, we use a sentence specific LM with paths corresponding to the options for each word (DNN) (correct, omitted, substituted with a predicted form or substituted with an OOV) with appropriate probabilities. We choose fragment based modeling of OOVs [28] where
the fragments are phone sequences determined in a data-driven way from an English and a Hindi dictionary [29], [30].

To assign probabilities to various parallel paths around each target word, we use heuristics based on our more general observations of the story reading by children about dependence on word category i.e. whether it is a function or content word (parts of speech) and the complexity of the word, typically the length in phones. For example, our observations indicate that function words are more likely to be missed rather than substituted, whereas the reverse holds for content words. Compared with function words and short content words, long content words have a higher probability of miscues relative to correct utterances. We use a set of heuristic probabilities for each of the above word classes, and employ the probabilities assigned to miscue (omission and OOV combined) relative to correct (including predicted substitutions) to tune the achieved miscue detection versus false alarm rate of the system.

3.4. Reading Miscue Detection

To evaluate the performance of the ASR system, we compute the traditional measure, the Word Error Rate(WER). Here we look for the precise word while considering a correct detection of OOV by the fragment bigram model as a correct recognition. For the task of miscue detection however, we report the results in terms of detection rate/recall (DR) and false alarm rate (FAR) of miscues [2], [3], [7] in a 2-fold cross-validation mode using the total adaptation and test data divided in a manner such that there is no overlap of speakers or stories. A small amount of data was kept aside as a validation set in each of the 2 folds. This validation set was only used to tune the LM weight and Word Insertion Penalty(WIP) in the ASR. The definition of miscue is motivated by [1] as including omission and substitution by OOV. Also, since we are interested in obtaining accurate alignments for the subsequent extraction of word-level prosodic cues, we report our results on a location accuracy based metric (as done in [28]). This is the fraction of words in reference ground-truth (GT) alignments that have some hypothesized word in decoder output whose both start and end boundaries fall within +/- 50ms of the GT word. This metric captures the information about alignments, useful for prosodic evaluation, irrespective of underlying word. For the location accuracy, GT boundaries were obtained by forced alignment with the GT transcripts using MAP adapted DNN Tandem SAT acoustic models.

The reported figures in Table 3 correspond to the test set described in Table 1. We obtain an overall miscue detection rate of 68% with false alarm rate close to 10% which is comparable to the reported performance of reading assessment systems built with significantly higher amounts of training data [3], [31]. The analysis of errors of our system shows that our OOV model (i.e. fragment bigram) sometimes eats up words adjacent to actual OOVs indicating that further topological constraints on phone/fragment bigrams may be warranted.

4. Prosodic Event Detection and Scoring

Duong et al. [13] uses correlation with measured adult speech prosody features for the same text to assess the child’s reading prosody. In the interest of a more general system, and also given that our data is characterised by word omissions and substitutions, we prefer to base our automatic ratings on the known generic acoustic correlates of prosodic attributes. Based on the previous discussion of subjective rating of prosodic attributes (phrasing, sentence ending and prominence) in Sec. 2.1, we investigate acoustic features at the sentence level for the prediction of subjective ratings. This is followed by a discussion of the implementation and evaluation of the automatic scores on the prosody test dataset.

4.1. Acoustic Cues

Correct phrasing refers to the grouping of words so as to indicate phrase breaks between the correct groups as per the text. A phrasal break can be expressed by a pause and/or a pitch reset [32]. The latter refers to a large pitch difference (>15Hz) between end of a phrase and the start of the next [33]. Syllable lengthening at the end of the phrase is also an important cue to phrasal break perception [32]. It has been observed that number of pauses with respect to the expected number as per the text, mean and standard deviation of pauses and word-level average syllable duration are important acoustic cues for the detection of phrasing [11]. We observe in our data that students at beginner level tend to read in list form. Besides monotonous pitch, the list form reading may be perceived through relatively large pauses between consecutive words and/or unusual lengthening of every syllable in the sentence. Our observations show that if the average syllable duration for each word is more than 300ms, the utterance is perceived as list form (i.e. the lowest subjective rating for phrasing). Sentence endings are typically cued by the pitch contour slope and trend (rise, fall or flat) over the segment corresponding to the final word[32]. Next, perceived prominence depends on acoustic features at the word level such as the RMS energy, average intensity, syllable duration, pitch span, maximum pitch, average pitch, and pitch difference across adjacent words [34], [35].

4.2. Implementation

The acoustic features required for automatic prosody ratings are estimated at the word level using the segmentation provided by the ASR decoder.

4.2.1. Feature Estimation

The pitch contour is estimated at 10 ms intervals across the utterance using an autocorrelation based pitch detector over 20ms Hamming windowed speech segments. Unvoiced regions are detected based on low pitch salience and energy. The resulting pitch contour is smoothened further based on [36] where the complete pitch contour is divided into distinct continuous parts.
subject to the conditions: 1) the adjacent pitch values should not deviate more than 12%, and 2) such continuity should exist for at least 5 cent words. Octave errors are corrected by doubling or halving small pitch regions appropriately wherever abrupt pitch changes are observed. We also compute a short-time intensity contour at 10 ms intervals across the utterance.

The ASR decoder hypothesis provides word onset and offset and intervening silence boundaries. For our work, silence regions greater than 150 ms are deemed as 'pause' as in [11]. The number of pauses, mean and standard deviation of pauses, and position of the pauses are calculated using the silence intervals. In order to obtain the average syllable duration for a given word, we divide word duration by number of syllables in the hypothesised word. From the computed pitch and intensity contours, we obtain the maximum value of pitch, mean pitch, pitch span, standard deviation of pitch, RMS energy and average intensity for each word segment.

4.2.2. Prosodic Event Scoring

For phrasing estimation, we first check for "list form" reading. For this, we appropriately threshold each of the following features: at the sentence level, we consider number of pauses, standard deviation in pause duration and average syllable duration; at word level, for words other than phrase-final words, we consider the average syllable duration and pitch span. For the remaining utterances (i.e. those not falling in the "list form" category), we search for the position where the pitch reset or pause is observed. If all the candidates are found at the expected positions as per the target story text, we assign rating 3. If number of phrasal breaks are more or less than the number of expected pauses or if the break is at the wrong position, rating 2 is given.

In order to score sentence ending, pitch shape and span over the last word are examined. The pitch shape is said to be 'flat', if the pitch span across the word is less than 5 Hz. In all other cases, the pitch shape is considered either rising or falling. For the rating level 2 (correct ending), a rising shape is expected for the yes-no question and falling otherwise [33]. In case the pitch slope on last word is close to zero, the pitch declination over the complete utterance is considered. If the latter exceeds 10Hz, sentence ending is rated correct.

For prominence prediction, we train a decision tree using 90 utterances with prominence on the correct words. We obtain 192 prominent and 254 non-prominent words in the training set. The classifier uses the word-level features of mean pitch, standard deviation of pitch and pitch span, all normalized by average pitch at the utterance level. RMS energy and average intensity are also normalized with sentence level RMS energy and average intensity respectively. Average pitch difference among neighboring words and average syllable duration per word are other important features in the prominence classifier. We then test words in the remaining utterances of the prosody test dataset (utterances with subjective rating 1 and 2) using this classifier. If any one word in an utterance is found to be prominent, we assign rating 2 to the utterance; if no word is found prominent in the whole utterance, it is marked as rating 1.

<table>
<thead>
<tr>
<th>Phrasing</th>
<th>Prominence</th>
<th>Sentence Ending</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR(%)</td>
<td>57</td>
<td>55</td>
</tr>
<tr>
<td>RC(%)</td>
<td>71</td>
<td>57</td>
</tr>
<tr>
<td>Acc(%)</td>
<td>64</td>
<td>58</td>
</tr>
</tbody>
</table>

Table 4: Prediction of prosody ratings evaluated in terms of Precision-Recall and Accuracy using decoder alignments of prosody evaluation dataset from Table 2

4.3. Evaluation With Respect to Subjective Ratings

The automatically computed prosodic event scores are compared with the corresponding subjective ratings at the sentence level. Since the goal is to flag reading errors, we report our results in terms of precision-recall and accuracy (shown in Table 4) in the detection of reading errors for each of the three subjectively rated events, viz. phrasing (where both rating levels 1 and 2 constitute reading errors), sentence ending and prominence (where rating level 1 constitutes a reading error in each case).

The sentence ending reading error shows that we can rely on pitch contour of the last word to determine the proper sentence ending. Cases where the sentence ending error goes undetected are typically associated with the occurrence of final word prominence. Another case that needs further investigation is that of flat pitch ending, which sometimes gets perceived as correct sentence ending.

Due to the interdependence of phrasing and prominence, students tend to lengthen the syllables though maintaining proper chunking of utterance. The higher average syllable duration of the sentence then leads to poor phrasing decision (rating 1) by our system. Some other false alarms arise from improper word segmentation by the decoder due to the confusion of recording noise, breath noise, fillers like 'ummm' with phones in adjacent words. This shows that syllable duration and pause related features obtained by our ASR may need refinement for phrasing.

In the prominence estimation, syllable duration lengthening is found to be the most important feature followed by pitch span, RMS energy, and change in average pitch across adjacent words. The pitch span feature is expected to be large for a prominent word. However, for the last word in a sentence, we expect large pitch decline for sentence end realization, and hence large pitch span. Same is true for standard deviation of pitch. The overlapping characteristics of sentence ending and prominence on final word need to be addressed with better features. Further, if even a single word is wrongly marked prominent, the sentence-level decision is affected. Finally, we note that errors in pitch estimation arising from the challenges of signal quality and pitch range can affect the accuracy of the automatic prosodic event ratings.

The whole set of experiments for prosody evaluation is repeated with word-level alignments obtained from known annotated ground truth transcription. The results are quite similar suggesting that the decoder hypothesis are reliable enough for prosody features estimation.

5. Conclusions

We presented ongoing work on the development of a reading assessment system with a discussion of our field data collection and labeling methods. We defined reading errors to match the expectations of a specific literacy monitoring project in terms of word decoding accuracy and fluency. An ASR framework was used to detect word omissions and substitutions, as well as to obtain word-level segmentation for prosodic events of interest. We achieve reasonable accuracy on detection of reading miscues and on the prosody attributes related to phrasing and sentence-ending detection. Word prominence detection needs further work. Future work is targeted towards training and testing with larger datasets on possibly more diverse speakers and environmental conditions.
6. References


Improving fluency of young readers: introducing a Karaoke to learn how to breathe during a Reading-while-Listening task

Erika Godde¹, Gérard Bailly¹, David Escudero², Marie-Line Bosse³, Maryse Bianco⁴, Coriandre Vilain¹

¹ GIPSALab, Univ. Grenoble Alpes & CNRS, F-38000 Grenoble, France
² Department of Computer Science, Univ. of Valladolid, Spain
³ LPNC, Univ. Grenoble Alpes & CNRS, F-38000 Grenoble, France
⁴ LSE, Univ. Grenoble Alpes, F-38000 Grenoble, France

Abstract

We evaluate here an original method for implicitly providing breathing guidance to young readers during a Reading-while-Listening task. Close-shadowing is supplemented by a Karaoke system that highlighted not only the word – as it is jointly spelled by the pre-recorded expert reader and the learner – but also the current and next breath groups. We compare the impact of two highlighting conditions – i.e. word-only vs. word and breath group – on fluency and comprehension. 

Index Terms: reading, reading-while-listening, breathing, fluency, children

1. Introduction

Fluent reading requires the setting up and coordination of multiple processes [1]: visual decoding of letters, letter-to-sound mapping and phonological processing, linguistic analysis, breathing and speech articulation, etc. The setting-up of this complex neural system is thus progressive. Fluency in reading is acquired through several skills: accuracy, automaticity, grouping and expressivity [2]. We are interested in the grouping skill. During the two first years of school, learning to read is mostly focused on accuracy and automacity. Many pupils think then that a good reader is a fast reader, regardless of understanding or being understood. During the third grade, children try to have a more listener-oriented reading. The grouping of words into meaningful units is one of the first challenge of this perspective taking. Well-formed prosody – i.e. regular breathing rhythm and pause patterning – should be thus planned according to the text structure.

The aim of our project is to help young readers in acquiring grouping skills by enhancing their breathing patterns while reading. For this purpose, we propose to use an audio-visual enhanced Reading-while-Listening paradigm (RwL) with close-shadowing: the young readers are instructed to shadow skilled readers while the current spelled units – namely characters, phones, syllables, words, phrases, breath groups, etc – are synchronously highlighted as in Karaoke. We test if the synchronous highlighting of the breath groups of the skilled reader influences breathing patterns of young readers... and further help them to plan meaningful word groupings. This study aims to answer 3 main questions:

Close-shadowing. Are children able to perform close-shadowing in this audiovisual RwL condition?

Implicit monitoring of breathing patterns. Does the highlighting of breath groups by Karaoke impact children’s breathing patterns?

Impact on fluency and comprehension. Does this exposure to enhanced RwL have an impact on text comprehension and further autonomous reading?

2. State of the art

According to Rasinski et al [2], there are four levels of fluency:

Accuracy: the reader is in a decoding phase. He's reading word by word, even syllable by syllable at slow rates.

Automaticity: the reader reads entire words and begins to read primarily in 2/3 words grouping regardless of the context. He can read slowly or fast.

Grouping: the reader groups words in an appropriate phrasing at an appropriate rate.

Expressivity: the reader makes large and meaningful phrase groups with expressive interpretation.

Rasinsky et al. also point out that reading fluency does not only consist in acquiring automaticity and speed but also linguistic and paralinguistic prosody, i.e. the last two stages. Prosody aims at easing linguistic analysis but also adds meaning to the text i.e. "how the speaker feels about what he says" [3]. Rasinsky finally notes that speed is often an obstacle to prosody planning. Moreover, Arcand et al [4] recently showed that comprehension is more correlated to prosody than to speed and accuracy, notably with young readers. A recent study of Alvarez et al [5] also linked comprehension and prosody – particularly pause distribution – with children of 3rd and 6th grades.

2.1. Enhancing grouping by working on breathing

One of the functions of prosody is to ease segmentation and hierarchy, notably by translating punctuation into speech [6]. This means marking inter and intra-sentence boundaries by pauses. Schwanenflugel et al showed that there are significant differences between intra and inter-sentence pause lengths, as well as variability of their respective distributions among young readers, depending on their reading skills. The more skilled readers have shorter pauses and less variability. While inter-sentence pauses are based on punctuation and are therefore easily located by the children, intra-sentence pauses are based on syntax and necessitate an incremental processing of text that is often
beyond children’s capabilities. Arcand et al [4] showed that comprehension is negatively correlated to inappropriate pauses. Several studies on adults [7, 8, 9] also separate the pauses in two categories: breathing pauses and non-breathing pauses. Breathing pauses are longer and located at major syntactic boundaries. Appropriate location of inhalations is thus both beneficial to the speaker (take air regularly and avoid out-of-breaths) and the listener (ease linguistic analysis). Grosjean et al [8] showed that when the readers read faster than normal, the need to breathe takes over and controls the occurrence of pauses, inducing a change in word grouping and making breathing pauses shorter, a phenomenon we experienced with young readers.

2.1.1. Reading-while-Listening

Reading-while-Listening (RwL) has long proved to be a good way to enhance fluency. In this method, the learner listens to someone else reading a text and simultaneously follows his performance in the printed text. Then he tries to imitate the skilled reader he just heard. This method quickly improves reading speed and accuracy [10, 11]. In our study, we used an enhanced version of RwL, using both audio and visual stimuli. This Karaoke method has been used by some teams to improve fluency. Biggs et al [12] used a singing Karaoke with middle school struggling readers during 9 weeks, 3 times a week. The children were very motivated by the exercise of repeated reading in this condition. Their fluency reading scores were remarkably improved during this session. Gerbier et al [13] used a reading Karaoke to improve word memorization and eye movement while reading, with th grade pupils. They showed that the Karaoke improved attention, gaze strategy and semantic recall.

2.1.2. Close-shadowing

In the Gerbier et al experiment, the subject had to read simultaneously with the karaoke, trying to synchronize his speech to the tutor’s one. We hope that the children imitating the karaoke tutor will also be synchronizing their breath with the skilled reader, and so making meaningful breath group. To our knowledge, this close-shadowing has not been studied with children. Stenton [14] augmented L1 listening perception and spoken production through Karaoke with an emphasis on lexical stress visual annotation. But subsequent spontaneous conversation did not confirm the idea of improved oral production. Bailly et al [15] studied close-shadowing with adults and the respiratory shadowing while reading with adults too. They show that, with adult readers, reading rates tend to adapt to each other depending on the degree of involvement in the joint action. More recently, shadowing has been proved useful to improve prosody of Japanese learners of English [16].

2.2. Monitoring breathing

To study the breathing shadowing, we needed to record breathing while Karaoke-enhanced RwL. As the subject were children in a classroom context, we used a system that is light and easy to use and transport. Several studies used respiratory belts on reading adults [9, 15, 17]. We used two belts and a Biopac® system to acquire synchronized audio and respiratory data (see Fig 2).

2.3. Assessing understanding and fluency

In order to assess the impact of the karaoke RwL on the pupils’ reading skills, we chose to use a particular maze test. Maze tests are proved to be good indicators of reading skills, particularly comprehension [18, 19]. While maze is a curriculum test in the US, it is not commonly used with French pupils. In this test, the subject is given a text with one word in seven deleted and substituted with 3 proposals, including the correct word. The subject has 3 minutes to circle as many correct proposals as he can. We used an alternative maze test in which the text is entirely read aloud by the subject. The test is scored on both right answers and reading time [20]. This oral maze test (OMT) enables to assess both fluency and comprehension in a rapid single test [21, 22].

3. Experimental design

The protocol was submitted and approved by the director of the experimental department (Isère) services from the French Ministry of Education. 51 3rd graders (8.6±0.3 years) from 3 classes with volunteer teachers participated to the experiments with the written authorization of their parents: 1 class from a small rural school (15 pupils), 1 class from a small urban school (10 pupils) in the Grenoble area. The average score to the RAVEN [25] test (35.3±7.6) corresponds to the group’s age (35 at 8 years). The average reading level (8.3±1.1 years) estimated via the Alouette test [26] also corresponds to the mean group age. It is to be noted that there are 2 skilled readers among the group (reading age = 12.25, RAVEN = 49).

3.1. Protocol

We contrast between two Karaoke conditions: Word-Only (WO) highlighting vs. Word & Breath Group (WBG) highlighting. Four texts of approximately 100 words were chosen from books used in 3rd grade by French teachers [23, 24]. The texts were rated of equal complexity in terms of average lexical frequencies and sentence lengths. Subjects were first presented with the objectives of the study (monitoring their breathing when reading) and the workflow, where we monitored speech and breathing patterns during a series of RwL, free and OMT readings of two texts among the four. RwL of each text was performed in one of the two conditions we wanted to contrast, WO vs. WBG, in random order. Each session thus consisted in the following tasks:

• Screening their non-verbal IQ (Raven test [25]) and reading age (Alouette reading test [26]),
• Familiarization with the Karaoke display using a sample text with WBG presentation,
• For each text : reading RwL condition, OMT, free aloud reading.

3.2. Subjects

The protocol was submitted and approved by the director of the experimental department (Isère) services from the French Ministry of Education. 51 3rd graders (8.6±0.3 years) from 3 classes with volunteer teachers participated to the experiments with the written authorization of their parents: 1 class from a small rural school (15 pupils), 1 class from a big rural school (26 pupils), 1 class from a small urban school (10 pupils) in the Grenoble area. The average score to the RAVEN [25] test (35.3±7.6) corresponds to the group’s age (35 at 8 years). The average reading level (8.3±1.1 years) estimated via the Alouette test [26] also corresponds to the mean group age. It is to be noted that there are 2 skilled readers among the group (reading age = 12.25, RAVEN = 49).
3.3. Experimental setting
The typical experimental setting is shown in Figure 2. The RAKE (Reading Assistance by KaraokE) system provides the synchronous audiovisual display of texts read aloud by the first author at an average rate of 99 words/min and an articulation rate of 3.6 syllables/s. RAKE is running on a ASUS TFT 100 tablet. Subjects wear a Audio-Technica BPHS1 stereo headset, including both headphones and a cardioid dynamic microphone. Thanks to a mixing board (Micline mixer Mackie 402-VLZ3), the master sound from the tablet is sent to both the headphones and the Biopac MP150. The latter also collects the respiratory patterns by means of two RSP100C belts placed on the abdomen and the thorax of the subject and his/her voice, recorded via the headset microphone.

Figure 2: Recording breathing with the abdominal and thoracic respiratory belts and voice with the headset of a pupil reading the Karaoke.

3.4. Data processing and labeling
Audio signals were automatically aligned with text using context-independent HMM trained on children data. The silence model was especially designed for coping with breath noises, coughs and sniffs. These phonetic alignments were further hand-corrected using Praat because of the large variability on the data quality between subjects. Corrections were made mostly on repetitions, mispronunciation and hesitations. Onsets of phonation were complemented with annotations regarding the time delay between these events and the neighboring inhalation onsets. Note that these delays are occasionally negative, i.e., several phonations start with ingressive sounds performed during inhalation.

3.5. Analysis
The objective was to test the shadowing of voice and breathing between the student and the karaoke tutor and the text comprehension in both conditions (WO vs WBG). We first compared the pauses and ratio between pseudo-syllables and expected syllables in the three tasks (RWL, OMT and free reading) using Praat and AuToBI. In order to globally characterize the close-shadowing performance, we computed the optimal delay between the children (channel 1) and master (channel 2) by cross-correlating the short-term energy profiles. The breathing was characterized by the inhalation-to-phonation delay. A t-test was performed to compare this delay in both condition (WO vs WBG). The delay between children and master’s inhalation was also measured and compared using a t-test. Finally, during OMT, we noted the number of correct answers and the time needed to complete the task. We used an anova to estimate the influence of the condition (WO vs WBG), reading level (Alouette), non-verbal-IQ (Raven) and text, on the OMT scores.

4. Results
4.1. Comparing RwL, OMT and free reading
Figure 4 contrasts the performance of the readers in the different activities. Due to the difficulty of the task, OMT requires more frequent and lengthy pauses: the total pause duration is 59% for OMT compared to 55% and 44% for RwL and free reading. These pairwise comparisons show that the total pause durations significantly differs along all the conditions (p < 0.001). As expected free reading is faster than RwL since Karaoke forces the reader to constitute consistent breathing groups, causing the overall pause duration to increase. There is a higher number of pseudosyllables in free reading and OMT compared to RwL: the ratios between the number of pseudosyllables and the number of expected syllables in the read text are 1.06, 1.04 vs 0.91. Differences between the distribution of RwL ratios vs. free reading and OMT are statistically significant at < 0.001. There is no significant differences between free reading and OMT (p > 0.4). Note that the value of the ratio for free reading and OMT is higher than 1. This is due to the high number of disfluences found in the readings.

Figure 4: Comparing rhythm of RwL, OMT and free reading. Left: percentage of silence in the utterance. Right: ratio between pseudo-syllables and expected syllables. We used Praat [27] for locating the silences and AuToBI [28] for finding the pseudo-syllables.

4.2. Close-shadowing performance
Figure 5 displays the distributions of the delays between master and children and the associated correlation coefficients according to the Karaoke condition. Note that the average shadowing delay is close to 100ms, which is a bit higher than what
Bailly [29] observed for adults reading a well-known passage. While correlation coefficients seem to be higher in the WBG condition, none of the observed differences are statistically significant.

While correlation coefficients seem to be higher in the WBG condition, none of the observed differences are statistically significant. We now plan to study the impact of the highlighting condition on comprehension.

4.3. Impact of the highlighting condition on breathing

Figure 6-left shows the distributions of the inhalation-to-phonation delays according to the Karaoke condition. The condition have a strong impact ($p < 0.0001$) on this feature: children anticipate inhalation much earlier in the WBG condition because of the cueing of upcoming breathing opportunities/dates by the enhanced Karaoke. Five pupils outline the distributions, mainly because of their unstructured phonations that do not provide reliable landmarks for relating phonation onsets to corresponding inhalations. Note that we get an average of 500 ms. This should be compared to the 350ms delay usually performed by adults within paragraphs [7]. We expect that the maturing of the synchronization between breathing and speech production will evolve towards a just-in-time performance, i.e. end of inhalation coincides with start of phonation. Moreover delays between children and master’s inhalations differ significantly between the 2 conditions (cf. right side of Figure 6). The delay for WBG condition is significantly lower ($p < 0.008$) showing that children tend to have a breathing pattern closer to the tutor’s one than in WO condition. It is to be noted that the pupils outlining the distributions are, in this case, the best readers (Alouette score > 10). Their breath groups are meaningful but also longer than the tutor’s ones which are regulated to help poor readers with small breath groups.

4.4. Impact of the highlighting condition on comprehension

The OMT performance measurements show no significant difference between the 2 conditions. Contrary to our expectations, the scores are a little higher in the WO condition. However if the number of correct answers is significantly better for the WO conditions, the time-to-complete is lower for the WBG condition. To summarize, they read faster but make more mistakes in WBG condition. Reading level and non-verbal IQ also have a significant impact respectively on the time and the score ($p < 0.05$). We also had a look at the number of words read and the number of words correctly spelled during the Karaoke. The ANOVA shows also a significant influence of the reading level (Alouette score) ($p < 0.001$) on both number of words correctly spelled, and on the OMT score. It is to be noted that for all those analysis, as showed in Figure 7, the text “Clara” – whose average lexical frequency is slightly lower than the other three – has significantly different results compared to the other 3.

5. Discussion

The RAKE system is planned to be used in classroom as a training program for improving reading fluency. It could be used in repeated reading e.g. 3 sessions a week during a long period e.g. 3 months. In that respect, we are interested in the acceptability of the system. Several experimental variables are expected to impact the subjects’ performance and the acceptability of the RAKE system: the delay between the acoustic onset of a word and its highlighting is set to zero. We have however observed that adult readers prefer an anticipatory highlighting about 300ms before its spelling. This variable should in some way mirror the complex subject-specific patterning of the coordinations between gaze and speech production. We will explore the possibility for subjects to adjust this delay. The corresponding psychometric procedure is to be adapted for children. Another way to increase the acceptability is to adapt the reading speed of the tutor to the fluency of the learner. In fact, several children didn’t manage to follow correctly the Karaoke because it was too fast. We will then explore the possibility for the readers to adjust the reading speed to be comfortable. Pauses are also highlighted with a red marker that presently lights up in red and gradually vanishes to white as a function of relative elapsed duration. This certainly favors the anticipatory inhalation strategy observed in the WBG condition. We will study the impact of the highlighting profile on the reading pattern.

During this long-term training sessions, we expect the different highlighting styles to have an impact on different aspect of reading fluency. The children could then progress from one highlighting style to another while their fluency improves e.g. first children work on WO highlighting to train automaticity and then when reaching a necessary level move to WBG highlighting to train phrasing. To this purpose, the following observations need further investigations. Firstly the number of word correctly spelled during the Karaoke, the OMT scores are impacted by the text complexity and the children’s reading level. We can hypothesize that the impact of the Karaoke condition is different regarding the reading level. Good readers cap in the reading and comprehension task, while the struggling readers have difficulties following the Karaoke, impacting both breath patterns and comprehension. We now plan to study the impact...
of Karaoke condition for the different reading levels. Secondly as the results shows a difference in pauses and breathing between the two conditions, it could be interesting now to have a look at the pitch and F0, to see if the breath group highlighting has also an impact on prosody. Thirdly in this study we tested the comprehension after only one text reading. By increasing the number of Karaoke reading – i.e. with repetitive reading [30] – we may increase the difference between the two conditions.

6. Conclusions

Our project aims at enhancing fluency in young readers by introducing a Karaoke that both highlights the current word and current/next breath groups. The RwL task is supposed to improve their prosody and comprehension skills. In this first study, we tested two different highlighting conditions: word only (WO) vs. word and breath groups (WBG) on 51 3rd grade children. We show that WBG has a significant impact on breathing patterns and fluency (less pauses and less pseudo-syllables). The impact on text comprehension is however not significant.

We will carry on characterizing the children’ performances, notably the free readings. Detailed analysis of segmental and prosodic patterns will be conducted to further examine the impact of the RwL condition on close-shadowing and further autonomous reading. One exposure is of course not sufficient: we will expect repeated RwL to improve over the sole repeated autonomous reading [31].

7. Acknowledgements

This work is supported by the e-FRAN project Fluence, sponsored by the “Investissements d’avenir” program.

8. References

Modeling Discourse Coherence for the Automated Scoring of Spontaneous Spoken Responses

Xinhao Wang†, Keelan Evanini‡, Klaus Zechner‡, Matthew Mulholland‡

Educational Testing Service R&D
†90 New Montgomery Street, Suite 1500, San Francisco, CA, USA
‡660 Rosedale Rd., Princeton, NJ, USA
{xwang002, kevanini, kzechner, mmulholland}@ets.org

Abstract

This study describes an approach for modeling the discourse coherence of spontaneous spoken responses in the context of automated assessment of non-native speech. Although the measurement of discourse coherence is typically a key metric in human scoring rubrics for assessments of spontaneous spoken language, little prior research has been done to assess a speaker’s coherence in the context of automated speech scoring. To address this, we present a corpus of spoken responses drawn from an assessment of English proficiency that has been annotated for discourse coherence. When adding these discourse annotations as features to an automated speech scoring system, the accuracy in predicting human proficiency scores is improved by 7.8% relative, thus demonstrating the effectiveness of including coherence information in the task of automated scoring of spontaneous speech. We further investigate the use of two different sets of features to automatically model the coherence quality of spontaneous speech, including a set of features originally designed to measure text complexity and a set of surface-based features describing the speaker’s use of nouns, pronouns, conjunctions, and discourse connectives in the spoken response. Additional experiments demonstrate that an automated speech scoring system can benefit from coherence scores that are generated automatically using these feature sets.

Index Terms: discourse coherence, spontaneous spoken English proficiency, automated speech scoring

1. Introduction

In recent years, much research has been conducted into developing automated assessment systems to automatically score spontaneous speech from non-native speakers with the goals of reducing the burden on human raters, improving reliability, and generating feedback that can be used by language learners [1, 2]. Various features related to different aspects of speaking proficiency have been explored, such as features for pronunciation, prosody, and fluency [3, 4, 5, 2], features for vocabulary and grammar, as well as content features [6, 7, 8, 9, 10]. However, discourse-level features have, to our knowledge, not been investigated in the context of automated speech scoring, except for our own previous work [11]. This is despite the fact that an important criterion in the human scoring rubrics for a standardized international English language speaking assessment is the evaluation of coherence, which refers to the conceptual relations between different units within a response [12].

This study focuses on modeling coherence cues for spontaneous speech, and its main contributions can be summarized as follows:

• We extended a corpus of coherence annotations for spoken responses drawn from a large-scale standardized assessment of English for the academic domain from 600 to 1440 responses (600 responses are double-annotated, 840 responses are single-annotated).
• We used these coherence annotations on a corpus of spontaneous spoken responses to demonstrate the effectiveness of using coherence cues in the task of automated speech scoring by (1) evaluating correlations between human discourse coherence scores and human holistic spoken proficiency scores, and (2) evaluating the extent to which human discourse coherence scores can improve a baseline scoring model for spontaneous speech.
• We examined two different types of feature classes to automatically model these coherence scores from human annotators, namely features from an NLP system designed to measure multiple aspects of text complexity as well as several surface-based features which were designed to represent the use of nouns, pronouns, conjunctions, and discourse connectives in a spoken response.
• We used these features for predicting human discourse coherence and evaluated (1) the prediction accuracy compared to the human coherence annotations, and (2) the improvement of a baseline speech scoring system when including these automatically predicted discourse features.

2. Related Work

Methods for automatically assessing discourse coherence in text documents have been widely studied in the context of applications such as natural language generation, document summarization, and assessment of text readability. For example, [13] measured the overall coherence of a text by utilizing Latent Semantic Analysis (LSA) to calculate the semantic relatedness between adjacent sentences. [14] introduced a model for the document-level analysis of topics and topic transitions based on Hidden Markov Models. [15] presented an approach for coherence modeling focused on the entities in the text and their grammatical transitions between adjacent sentences, and calculated the entity transition probabilities on the document level. [16] provided a summary of the performance of several different types of features for automated coherence evaluation, including features based on cohesive devices, measurements of adjacent sentence similarity, Coh-Metrix [17], word co-occurrence patterns, and entity transitions. [18] proposed a graph-based approach for modeling transitions among entities in a text (in contrast to previous entity-grid approaches, which only examined transitions between adjacent sentences) and model coherence by calculating centrality measures on the nodes in the graph.
[19] examined the use of Recurrent and Recursive Neural Network models based on word embeddings for modeling coherence and showed that they outperform other approaches on the tasks of sentence ordering and readability assessment.

In addition to these studies on well-formed text, prior research has also investigated the task of evaluating coherence in student essays, which may contain multiple spelling, vocabulary, and grammar errors, especially when produced by non-native speakers of English. Utilizing LSA and Random Indexing methods, [20] measured the global coherence of students’ essays by calculating the semantic relatedness between sentences and the corresponding prompts. [21] combined entity-grid features with writing quality features produced by an automated essay assessment system to predict the coherence scores of student essays. [22] systematically analyzed a variety of coherence modeling methods within the framework of an automated assessment system for non-native free text responses and indicated that features based on Incremental Semantic Analysis (ISA), local histograms of words, the part-of-speech IBM model, and word length were the most effective. [23] applied ideas from Centering Theory to model coherence in short argumentative essays by calculating the percentage of sentences within each paragraph that have the same centers. [24] investigated the use of lexical chains (sequences of related words within a text that contribute to the continuity of meaning) through features related to the length, frequency, location, and quality of the lexical chains in an essay, and demonstrated that these features can effectively model writing proficiency scores for essays written by native and non-native speakers of English.

In contrast to these previous studies involving well-formed written text or learners written texts containing errors, our previous work provided a corpus of coherence annotations on 600 spoken responses from an English language assessment, and examined several features that have been used in the context of learner written essays based on human transcriptions of these spoken responses in the task of automatic prediction of human coherence scores [11]. The current study extends this research effort in the following three ways: 1) obtaining a larger set of spoken responses annotated with coherence scores (N=1,440); 2) extending the construct coverage of an automated speech scoring system by modeling the coherence quality scores using two new classes of features; and 3) employing these two new classes of features based on automatic speech recognition output in an attempt to measure the coherence quality of spontaneous speech.

In a related study, Hassanali et al. investigated coherence modeling for spoken language in the context of a story retelling task for the automated diagnosis of children with language impairment [25]. They annotated transcriptions of children’s narratives with coherence scores as well as markers of narrative structure and narrative quality; furthermore they built models to predict the coherence scores based on Coh-Metrix features and the manually annotated narrative features. The current study differs from this one in that it deals with free spontaneous spoken responses provided by students at a university level; these responses therefore contain more varied and more complex information than the child narratives.

3. Data

3.1. Spoken language corpus

The data used in this study was drawn from the TOEFL® Internet-based test (TOEFL® iBT), a large-scale standardized assessment of English for non-native speakers, which assesses English communication skills for academic purposes [12]. The Speaking section of TOEFL iBT contains six tasks, each of which requires the test taker to provide an extended response containing spontaneous speech. In total, we collected 1,440 spoken responses from the TOEFL iBT assessment, including 240 responses from each of six different test questions. These six test questions comprise two different speaking tasks: 1) providing an opinion based on personal experience (N = 480 responses) and 2) summarizing or discussing material provided in a reading and/or listening passage (N = 960 responses). The spoken responses were all manually transcribed, and the transcriptions include standard punctuation and capitalization. The average number of words per response was 113.8 (SD = 33.6) and the average number of sentences was 4.8 (SD = 2.1).

The spoken responses were all provided with holistic English proficiency scores on a scale of 1 - 4 by expert human raters in the context of operational scoring for the spoken language assessment. The scoring rubrics1 address the following three main aspects of speaking proficiency: delivery (pronunciation, fluency, prosody), language use (grammar and lexical choice), and topic development (content and coherence). In order to ensure a sufficient quantity of responses from each proficiency level for training and evaluating the coherence prediction features, the spoken responses selected for this study were balanced based on the human scores as follows: 60 responses were selected randomly from each of the 4 score points (1 - 4) for each of the 6 test questions.

3.2. Annotation

The coherence annotation guidelines used for the spoken responses in this study were modified based on the annotation guidelines developed for written essays described in [21]. According to these guidelines, expert annotators provided each response with a score on a scale of 1 - 3. The three score points were defined as follows: 3 = highly coherent (contains no instances of confusing arguments or examples), 2 = somewhat coherent (contains some awkward points in which the speaker’s line of argument is unclear), 1 = barely coherent (the entire response was confusing and hard to follow; it was intuitively incomprehensible as a whole and the annotators had difficulties in identifying specific weak points). For responses receiving a coherence score of 2, the annotators were requested to highlight the specific awkward points in the response. In addition, the annotators were specifically required to ignore disfluencies and grammatical errors as much as possible; thus, they were instructed not to label sentences or clauses as awkward points solely because of the presence of disfluent or ungrammatical speech [11].

Two annotators (not drawn from the pool of expert human raters who provided the holistic scores) first made independent coherence annotations for 600 spoken responses, including 25 samples from each of the 4 score levels of speaking proficiency for each of the 6 test questions. The distribution of annotations across the three score points is presented in Table 1 (numbers repeated here from [11] for convenience). The two annotators achieved a moderate inter-annotator agreement [26] of κ = 0.68 on the 3-point scale of coherence scores. Subsequently, the same two annotators provided coherence annotations for the remaining 840 responses in the corpus using the following approach: each annotator provided a single annotation for 420 responses from 3 test questions, i.e., 35 responses from each score

---

Table 1: Distribution of coherence scores from two annotators, where 600 responses receive double coherence scores and 840 responses receive single scores.

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Coherence Scores</th>
<th># 1</th>
<th># 2</th>
<th># 1</th>
<th># 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># 1</td>
<td>160 (26.7%)</td>
<td>278</td>
<td>162</td>
<td>251</td>
<td>225</td>
</tr>
<tr>
<td># 2</td>
<td>125 (20.8%)</td>
<td>251</td>
<td>224</td>
<td>112</td>
<td>152</td>
</tr>
<tr>
<td>Single</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># 1</td>
<td>112 (26.7%)</td>
<td>156</td>
<td>132</td>
<td>95</td>
<td>162</td>
</tr>
<tr>
<td># 2</td>
<td>163 (22.6%)</td>
<td>162</td>
<td>163</td>
<td>163</td>
<td>163</td>
</tr>
</tbody>
</table>

level for each of 3 test questions.

4. Method

In order to verify the effectiveness of the proposed coherence cues in the assessment of speaking proficiency, we conducted the following four experiments:

- correlating human discourse coherence scores with human holistic scores (for each spoken response);
- comparing a baseline scoring model for predicting human holistic scores that does not use coherence information with an extended scoring model using such coherence information (human coherence scores);
- predicting human coherence scores using two classes of low-level features; and
- comparing a baseline scoring model for predicting human holistic scores that does not use coherence information with an extended scoring model using such coherence information which is generated based on two classes of coherence features.

4.1. Correlation between human coherence scores and human holistic speaking scores

For the first experiment, we extracted two types of features based on the human annotations of the spoken responses, namely the coherence scores and the number of awkward points identified. These two features are then correlated with the holistic proficiency scores for each response. For the double annotation set (600 responses), the average coherence scores from two annotators were used; and the union set of awkward points identified by either annotator on each response was used.

4.2. Comparison between baseline scoring model and extended scoring model using human coherence scores

For the second experiment, these coherence cues from human annotations were further investigated within a context of an automated spoken language assessment system, SpeechRaterSM [27, 1]. SpeechRater can automatically generate various features to assess different aspects of spontaneous speech, including pronunciation, prosody, fluency, vocabulary, grammar, as well as content. In this study, we employed 12 of these features to measure different aspects of delivery and language use. These features were either extracted directly from the speech signal or were based on the output of an automatic speech recognition system within SpeechRater. Both the training and evaluation sets that were used to develop the speech recognizer consist of similar spoken responses drawn from the same assessment and do not overlap with the data sets included in this study; its word error rate on a held-out ASR evaluation set is around 28%..

This evaluation was conducted on the double-annotated responses only. Classification models were built to automatically predict the holistic speaking proficiency scores. 10-fold cross-validation was performed using the Random Forest classifier from SKILL2, a Python toolkit that simplifies the running of common Scikit-Learn experiments [28]. The classification accuracy, i.e., the percentage of correctly predicted holistic scores, and the Pearson correlation coefficient ($r$) between the predicted scores and the human scores were used as evaluation metrics. Specifically, we compared (1) the baseline model consisting of the 12 SpeechRater features only; and (2) an extended model where average human coherence scores and the number of merged awkward points were added to the baseline feature set separately.

4.3. Discourse coherence features

In the third and fourth experiments, we describe how we explored ways to model and automatically predict the human discourse coherence scores and then verified that the automatic scoring system for speaking proficiency prediction can also benefit from these automatically produced coherence scores.

In summary, we use two classes of features for the prediction of human discourse coherence scores: (1) features from TextEvaluator (TE) [29], a system originally designed for evaluating the complexity of reading passages, and (2) surface-based features looking at distributions of certain words and word classes that may indicate discourse coherence.

For these experiments, the 600 double-annotated responses were used as a training set and the 838 single-annotated responses were used as the evaluation set for the task of automatic prediction of human coherence scores.$^3$

4.3.1. TextEvaluator features

TextEvaluator$^4$ is a tool that employs a variety of natural language processing techniques, as well as linguistic resources such as word lists, to generate more than 300 features measuring multiple aspects of sentence structure, vocabulary difficulty, connections across ideas, and organization [29]. In this study, we explore the use of these features to model discourse coherence in spontaneous, non-native speech.

TextEvaluator features were first examined on the training set of 600 double-annotated responses, which were extracted based on both the human transcriptions and the automated transcriptions of the spoken responses (ASR output). The Pearson correlation coefficients ($r$) of these features with the average human coherence scores were used as the evaluation metric. Table 2 shows the distributions of the absolute values of feature correlations, which can be grouped into 6 bins. There are around 150 features with very low correlations, $r < 0.1$, and around 30 features with moderate correlations, $r \geq 0.4$, which indicates that these features can potentially contribute to the automatic

---

$^2$Downloaded from https://github.com/EducationalTestingService/skill

$^3$Two responses from the single annotation set of 840 responses were removed from the experimental set due to speech recognition failures.

$^4$https://textevaluator.ets.org
Table 2: Distribution of the absolute values of Pearson correlation coefficients (r) of TextEvaluator features with the average human coherence scores. Features were separately extracted on the human transcriptions and the ASR output.

<table>
<thead>
<tr>
<th>Features</th>
<th>Transcriptions</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>r &lt; 0.1</td>
<td>143</td>
<td>153</td>
</tr>
<tr>
<td>0.1 ≤ r &lt; 0.2</td>
<td>92</td>
<td>84</td>
</tr>
<tr>
<td>0.2 ≤ r &lt; 0.3</td>
<td>36</td>
<td>33</td>
</tr>
<tr>
<td>0.3 ≤ r &lt; 0.4</td>
<td>31</td>
<td>33</td>
</tr>
<tr>
<td>0.4 ≤ r &lt; 0.5</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>0.5 ≤ r &lt; 0.55</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

4.3.2. Surface-based features

In addition to using the NLP-based features in TextEvaluator, we also developed a set of simple features which were designed to be more robust towards errors from the ASR system (as well as ungrammatical speech produced by test takers); these capture the use of nouns, pronouns, conjunctions, and discourse connectives in a test taker’s spoken response. For this purpose, the discourse connective list from the Penn Discourse Treebank (PDTB) [30] was used. Various basic features were computed based on occurrence counts, such as the number of nouns, pronouns, conjunctions, as well as discourse connectives (counted based on word types and word tokens), and the ratios between these counts were also extracted. A preliminary evaluation was conducted to examine the correlations of these features with the average human coherence scores on the 600 double-annotated responses. And only features with absolute correlations greater than 0.1 on both the human transcriptions and the ASR output were included for further experiments. The first 5 features shown in Table 3 are based on word counts.

Loosely inspired by related work [15], we also designed additional features that capture the global coherence, represented by the use of pronouns, conjunctions and discourse connectives across an entire test response. In order to obtain these features, first, a reference corpus containing high-scoring responses was collected, and then a connective chain was extracted from each reference response, where only the pronouns, conjunctions, and discourse connectives were retained and all other words were removed from the response. Given a test response, a similar connective chain can be also extracted. Then by comparing the similarity of the test chain with each of the reference chains, the maximum similarity or the minimum distance can be computed as a feature to measure the proper use of the connective sequence in a test response. The following three evaluation metrics were investigated to evaluate the similarity between two chains: BLEU score [31], edit distance, and word error rate (which is a normalized edit distance).

Furthermore, the reference chains can be built in either an item-specific or generic manner: item-specific references were drawn from responses to the same test question as the test response; generic references were drawn from multiple different test questions. In this work, the reference samples were extracted from a corpus that was used to train the speech recognizer in SpeechRater and did not overlap with the discourse coherence annotation corpus used for this study. Around 200 - 260 responses with the highest human holistic speaking scores were obtained for each test question; in total, 1,395 responses across 6 test questions were collected as references. A preliminary experiment indicated that the BLEU similarity with the item-specific models, i.e., connective_chain_bleu_item, and the edit distance with the generic models, i.e., connective_chain_ed_generic, can achieve moderate correlations that are higher than the other model configurations. The performance of these two features is shown in Table 3.

4.4. Automated prediction of human coherence scores

In order to model the human coherence scores, regression models were built with the TextEvaluator and surface-based features described above. In this experiment, the training set contained the 600 double-annotated responses, and the test set contained the 838 single-annotated responses. Since the average coherence scores were used for model training, a regression model was used instead of a classification model for this experiment, specifically, the Random Forest Regressor from SKLL [28]. The Pearson correlation coefficients of the automatically predicted scores with the human-annotated coherence scores was taken as the evaluation metric. Features were separately extracted on the human transcriptions and the ASR output.

4.5. Automated prediction of speaking proficiency scores

Since this work aims to improve the performance of an automated speech scoring system by modeling the discourse coherence of spontaneous speech, in the final experiment, we further investigated the integration of the automatically predicted coherence scores into the classification models for the automatic prediction of holistic speaking proficiency scores. This experiment was conducted on the 838 single-annotated responses, and 10-fold cross validation was performed. Both the classification accuracy and correlations of the automatically predicted holistic scores with human holistic proficiency scores were evaluated. The baseline system was built by only using the 12 SpeechRater features as described in Section 4.2.

As described in the above Section 4.4, regression models predicting the human coherence scores were trained on the 600 double-scored responses and then used to predict the coherence scores for the 838 responses included in this evaluation, where automatic speech recognition output was used for feature extraction. Finally, results from a system using both
Table 4: Improvement to an automated speech scoring system by adding human-assigned coherence scores, labeled as Coh, and numbers of human-identified awkward points, labeled as nAwk. Both the classification accuracy and the Pearson correlation coefficient $r$ between the experts’ speaking proficiency scores and the automatic scores are reported.

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpeechRater</td>
<td>54%</td>
<td>0.728</td>
</tr>
<tr>
<td>SpeechRater + Coh</td>
<td>58.2%</td>
<td>0.779</td>
</tr>
<tr>
<td>SpeechRater + nAwk</td>
<td>58.5%</td>
<td>0.771</td>
</tr>
<tr>
<td>SpeechRater + Coh + nAwk</td>
<td>59%</td>
<td>0.782</td>
</tr>
</tbody>
</table>

the SpeechRater features and the human-annotated coherence scores were also reported for comparison.

### 5. Results

#### 5.1. Correlation between human coherence scores and human holistic speaking scores

The average coherence scores on the set of 600 double-annotated responses correlate with the proficiency scores at $r = 0.656$, and the number of the union set of identified awkward points correlates with proficiency scores at $r = -0.626$. This indicates that the assessment of spoken language proficiency could benefit greatly from modeling the coherence cues proposed in this study. The correlations with proficiency scores for the coherence scores and the number of awkward points on the single-annotated set of 840 responses were $r = 0.615$ and $r = -0.597$, respectively.

#### 5.2. Comparison between baseline scoring model and extended scoring model using human coherence scores

The average accuracy and correlation across 10 folds are reported in Table 4. The baseline system (with the 12 SpeechRater features) obtained an accuracy of 54%. Furthermore, by adding the average annotated coherence scores or the number of identified awkward points, the accuracy can be improved to 58.2% and 58.5% respectively. These experimental results demonstrate that the automatic scoring system can benefit from the coherence cues directly extracted from human annotations.

#### 5.3. Automated prediction of human coherence scores

As shown in Table 5, the TextEvaluator feature set, the union of word-based and connective chain features, and their combination were examined in this regression task. The TextEvaluator and surface-based and connective chain features extracted from the human transcriptions can achieve correlations of 0.53 and 0.379 respectively. However, when the automatic speech recognition output was used, the correlations decreased to 0.498 and 0.299, respectively, due to the presence of recognition errors. The further combination of all feature sets cannot result in any correlation improvement, which may be due to the fact that the 7 surface-based features can potentially be subsumed in the much larger set of 162 TextEvaluator features.

#### 5.4. Automated prediction of speaking proficiency scores

As shown in Table 6, in addition to the annotated coherence scores, labeled as AnnoCoh, three different sets of automatically predicted coherence scores were compared for classification model building, which were separately generated with TextEvaluator features, labeled as PredCoh(TextEvaluator), with surface-based features (including the connective chain features), labeled as PredCoh(Surface), and with the combination of TextEvaluator and surface-based features, labeled as TextEvaluator&Surface. The experimental results indicate that by adding the human-annotated coherence scores, the classification accuracy can be improved from 56% to 60.7%. In contrast, the automatically predicted coherence scores based on TextEvaluator features can improve the accuracy to 59.4%. When only using 7 surface-based and connective chain features to model the coherence quality, the automatic coherence scores can still improve the classification accuracy from 56% to 58.4% and correlation from 0.736 to 0.752. However, as also shown in previous results, when using the automatic coherence scores predicted with the combination of the TextEvaluator and the surface-based features, the system performance cannot be further improved.

### 6. Discussion and Conclusion

In this study, we presented a corpus of coherence annotations for spontaneous spoken responses, and the analyses on these annotations showed that an automated speech scoring system can benefit from modeling the coherence of spoken responses. Based on this finding, two different sets of linguistic features were employed to model the discourse coherence of spontaneous speech, and the automatically generated coherence scores were further examined in the automatic prediction of holistic speaking proficiency scores. Experimental results showed that the performance of an automated speech scoring system can be improved by automatically modeling the coherence quality scores based on automatic speech recognition output and then introducing the generated coherence cues to measure the coherence
ense of spontaneous speech.

While TextEvaluator features and word-based features can be computed directly for each spoken response, the coherence chain features need a pre-scored corpus of spoken responses to build a reference model, thereby putting this feature class at a disadvantage compared to the other classes. In this study, we only use the first-best ASR hypotheses for further processing (feature generation); however, we could also look into obtaining additional information from the recognizer lattice and/or from an ASR N-best list, thereby potentially improving feature performance for ASR output.

Finally, since the correlations between human coherence scores and human holistic speaking proficiency scores are quite high, it is conceivable that the additional step of human annotation of discourse coherence could be skipped, and instead of using the coherence features to first predicting human coherence scores as an intermediate step, they might be used directly as part of the SpeechRater scoring model for predicting human holistic speaking proficiency scores. In future work, we will also attempt to develop more effective discourse-related features which are more robust to ASR recognition errors.

7. References


CARAMILLA – Speech Mediated Language Learning Modules for Refugee and High School Learners of English and Irish

Emer Gilmartin1, Jaebok Kim2, Alpha Diallo2, Yong Zhao3, Neasa Ni Chiaráin1, Ketong Su1, Yuyun Huang1, Benjamin R. Cowan4, Nick Campbell3

1Trinity College Dublin, Ireland
2University of Twente, Holland
3Vrije Universiteit Brussel, Belgium
4University College Dublin, Ireland

gilmare@tcd.ie

Abstract

In the development of Computer-Assisted Language Learning (CALL) modules there is a growing emphasis on the use of evolving technological developments which include text-to-speech technology (TTS) and automatic speech recognition (ASR). These technologies allow for greater spoken interaction between the learner and the computer, which is particularly useful when native speaker input is not readily available and to elicit learner speech in a quasi-natural environment. They also allow for greater learner autonomy, and facilitate the creation of modules based on effective learning activities at all levels to address needs for accuracy as well as fluency. The present paper reports on two modules which integrate speech technology, (1) a spoken dictogloss, or text reconstruction, module and (2) a pronunciation module. These modules have been implemented in JAVA as part of CARAMILLA, a spoken dialogue-based language learning tool, which is based on an earlier prototype, MILLA. The dictogloss module is being evaluated with two distinct learner groups – adult refugees learning English while living in the host country and learners of Irish, a minority endangered language. The pronunciation module is also being evaluated with the adult refugee group.

Index Terms: CALL, spoken dialogue systems, autonomous learning

1. Introduction

In this paper we describe the design and implementation of two Computer-Assisted Language Learning (CALL) modules - a spoken dictogloss and a pronunciation module intended to add to our existing system [1], and as incremental progress towards our goal of building a spoken language dialogue system to provide flexible practice of second language skills via the Internet. We first briefly review relevant background in the CALL literature, and discuss the challenges and opportunities of provision of language learning resources using CALL. We then describe the user groups we are focusing on, outlining their specific needs and discussing how an online system and the modules described in this paper could benefit them. We overview our work to date with the MILLA system, and then describe the design and development of the dictogloss and pronunciation modules carried out at the ENTERFACE workshop in summer 2016. We describe our plans for evaluation and further development of the modules and discuss plans for future work on the system.

2. CALL applications for autonomous language learning

In this work CALL refers to the use of an artificial environment containing tasks and activities which help learners attain their goals of improving language skills. Speech technology offers interesting opportunities for language education. Such technology has now been extended to the use of speech recognition (ASR) and synthesis (TTS) in addressing specific tasks and in the implementations of complete tutoring systems [2]. Technology has long played an important role in language learning with video and audio courses available since the early days of audiovisual technology. Globalisation and migration, coupled with the explosion in personal device ownership, have increased the need and the opportunities for well designed, pedagogically sound CALL applications.

Many current CALL applications address the receptive skills (listening and reading) with activities much like a traditional listening or reading exercise transferred to a screen [3]. Such exercises offer few advantages over tradition pencil and paper activities with the possible exception of those which include an audio dimension. Pronunciation tutoring applications range from ‘listen and repeat’ exercises without feedback or with auto-feedback (the learner hears a recording of their attempt) to more sophisticated systems where the learner’s utterance is compared with the target utterance and feedback is given on errors and strategies to correct those errors. There are interesting examples of speech-technology-based spoken production training where phoneme recognition is used to provide corrective feedback on learner input. These include CMU’s Fluency [4], and Cabral et al’s MySpeech [5]. Much effort has been put into creating speech activities which allow learners to engage in spoken interaction with a digital conversational partner. This engagement is considered the most difficult competence for a learner to acquire independently. An effective approach to providing practice in spoken conversation (or texted chat) is to use relatively simple chatbot systems based on pattern matching (e.g. Pandorabots) [6, 7]. Dialog systems using written text and more recently speech have long been used to tutor learners in science and mathematics [8, 9]. However, conversation poses challenges as success does not solely depend on a lexical ‘right answer’ but rather on the degree to which the learner manages competent spoken interaction. Early language learning systems such as VILTS [10] presented tasks and activities based on different themes which were chosen by the user, while other systems concentrated on pronuncia-
tion training. Recently interest has grown in more holistic systems where learners engage in spoken interactions relevant to their learning needs such as service encounters or job interviews, particularly for assessment purposes [11]. While many of these systems are being developed with an emphasis on the assessment of communicative skills, and greatly aid fluency and pragmatic competence, there is still scope for activities which combine speech technology with learning tasks which help foster learners’ acquisition of lexical, syntactic and phonetic accuracy. With improvements in technology it is now possible to provide a range of activities from free conversation practice to more focussed activities. The use of gamification in educational software is currently receiving a lot of attention as a method of increasing learner motivation [12]. This approach lends itself very well to language learning activities, such as ‘grammar games’ and ‘information gap exercises’, and has long been a successful element of the language classroom. In CARAMILLA we aim to combine engaging activities which address specific skills and free conversation practice to allow learners to integrate their linguistic skills.

3. Motivation for System

Many language learners today do not follow traditional learning paths, but rather learn when and where they can. For these learners the goal is often communicative competence in the target language for practical purposes. Fundamental challenges are the development of spoken interaction skills, and the integration of lexical and syntactic knowledge to produce accurate and appropriate language in different contexts and to comprehend vocabulary and syntax. In more formal language provision, conversation classes and practice with peers or native speakers are the traditional methods for development of spoken language proficiency, while syntax, lexicon, and pronunciation development needs are met by teacher led instruction. However classes are expensive and learners may not have access to native speakers willing to practice with them. Indeed, even in traditional classroom settings, pronunciation is the poor relation, and although accurate pronunciation and intonation can be greatly accelerated by focussed practice, this seldom forms an integral part of the curriculum.

For practical reasons, language learning providers group learners according to an average of measures of their competence on tests of the traditional four skills - reading, writing, speaking, and listening. The fifth skill, spoken interaction, is often tested in a brief interview. This method of grouping learners by level does not address the fact that many learners, and particularly those who have not worked through a traditional Western modern languages curriculum in a formal learning environment, have ‘spiky’ profiles, with some skills present at much higher or lower levels than others. Access to independent learning resources which allow study and practice of the particular skills at levels suitable to the individual offers several advantages – learners can improve their weaker skills while learning at times and places convenient to them. An important advantage is that such resources can be offered online free or at very low cost to users – many of the learners most in need of language skills for work, education, and training, or integration cannot afford private bricks and mortar language tuition and state provision is often rudimentary at best.

The text (both spoken and written) underpinning language learning resources has usually focussed on topics common to the widest variety of potential learners possible - pop culture, uncontroversial general or human interest stories, sport, and activities useful to those learning a foreign language for tourism purposes. Specialised texts do exist for groups learning a language for use in fields such as business or medicine, but these are narrowly focussed on the target area. Applications where the underlying text can be pulled from a wide variety of sources suitable to a learner’s particular needs can be very helpful to many different learner groups. For an adult migrant, lessons based on texts drawn from social services websites provide the same basis for linguistic or functional learning goals as stories about pop stars, but also motivate learners as there is a clear link to learners’ everyday communicative needs. Similarly, for high schoolers, texts based on school subjects in the target language provide relevant vocabulary and content and language integrated learning.

For this project, we focus on two learner groups - adult refugee ESOL (English for Speakers of Other Languages) learners and learners of Irish.

3.1. Language needs of adult refugees

For an adult newcomer, living in a new country presents many challenges. Social, legal and cultural norms may differ greatly from those of the country of origin. When the language of the host community is unfamiliar, these challenges intensify [13]. Without knowledge of the language, access to training and employment is hampered, – employment and earning levels for US immigrants have been shown to be strongly linked to English language ability [14]. Enabling immigrants to acquire basic knowledge not only of the language but also of the customs and daily life of the host country is recognised as essential to successful integration. (Council of Europe, Common Basic Principles of Integration, 2001).

Adult refugees are a particularly diverse group of language learners. As an illustrative example, in Ireland’s national language and integration training organisation for refugees in 2008, 93 nationalities were present and educational backgrounds varied widely; while some learners held third level qualifications, others had had little or no formal education. Levels of literacy varied with learners presenting with no literacy in their own language, with literacy in a non-Latin alphabet, and fully literate in the Latin alphabet. The length of time that learners had lived in the host county (Ireland) ranged from several years to a few weeks, resulting in different levels of familiarity with Irish life and the English language. In terms of language ability, profiles were very spiky, with some learners having attained enough English to get by, often appearing fluent but with very low accuracy although listening comprehension could often be excellent. Others had studied formally but had comparatively low spoken interaction skills although reading and writing was of a high standard [15].

Learning a language to live in a country where the language is spoken is not a simple matter of attaining an academic understanding of the language. Better results can be expected when the texts used are tailored to practical communicative needs or areas of interest - a parent would benefit far more from an exercise on Present Tense structures based on the local education system than on a description of life on the Space Station. In addition, free or low cost language learning resources which can be accessed from home at any time are particularly helpful to this group. Activities which
allow topics of interest to the learner to be selected as un-
derpinning text can also foster learner autonomy and moti-
vation [16]. Caramilla's pronunciation and dictogloss activi-
ties are designed to be flexible, with pronunciation practice
available for many first languages, while the dictogloss can be
adapted to any text of interest or level of competence, and
thus should be useful to this group. The system is currently
being trialled with learners around the CEFLRL B1 level, de-
 fined as a level where the language learner ‘can understand
the main points of clear standard input on familiar matters
regularly encountered in work, school, leisure, etc. (Council
of Europe, 2001). This level was chosen as it is commonly de-
 fined as the ‘threshold’ for integration and relevant to all new-
comers. Of course the modules can also be tailored to learn-
ers who need to attain higher competences to meet gatekeep-
 ing requirements for entry to higher education or to progress
towards citizenship.

3.2. Irish language learners

Irish is a Celtic language which was spoken widely throughout
Ireland up to the 19th century but has since receded as a com-
munity language into just a few small geographical pockets
known as Gaeltacht areas. In these areas Irish has extended
as a native language in an unbroken chain. Most speakers of
Irish, however, have acquired the language through school-
ing. It has the status of being one of the two official lan-
guages in Ireland and since 2007 has recognition as an official
language of the European Union. It is a compulsory subject of
study in all primary and post-primary schools in the Republic
(for students from 4-18 years old) but motivation and attitude
amongst the learners is very variable [17, 18]. Irish, nonethe-
less, is classified as an endangered language by UNESCO [19].
There are few native speaker models available to the vast ma-
jority of learners of Irish and, anecdotally, the language com-
petence of teachers seems variable as most are themselves
second language learners.

Work on speech technology development for Irish has
been underway in the Phonetics and Speech Lab., Trinity
College, Dublin for some years [20]. Synthetic voices rep-
resenting the three main dialects of Irish have been devel-
oped as part of the ABAIR initiative and are freely available at
www.abair.ie. A significant strand of the ABAIR initiative has
been the development of CALL materials for Irish [21, 22, 7].

As part of the CARAMILLA project, the spoken dictogloss
game has been designed for Irish. It is intended to be a mo-
tivational and challenging CALL tool which goes beyond the
range of materials currently available for the teaching of Irish
insofar as it gives learners the freedom to choose their own
learning materials while integrating speech technology. It is
a highly interactive, task-based language learning exercise,
which gives freedom to learners to choose topics or themes
which are of relevance to them while providing them with ac-
cess to native speaker models. It is intended to have positive
effects on the attitude and motivation of Irish learners at sec-
ond level as well as being a pedagogical tool for the continu-
ing linguistic development of Irish teachers.

4. CARAMILLA System

The CARAMILLA project at the Enterface 16 workshop, held
at the University of Twente's Design Lab, was a follow on to an
erlier project (MILLA) at Enterface 14, held in Bilbao. MILLA
(Multi-modal Interactive Language Learning Agent) is a dia-
logue system providing spoken social chat at different levels
(two speech-enabled web-based Pandora chatbots), pronun-
ciation and traditional grammar training [1]. MILLA was cre-
ated in Bilbao by a team of nine.

The goals of the 2016 project at Twente were to pro-
duce language learning modules to extend the capabilities
of the MILLA/CARAMILLA language learning agent system.
There was an onsite team of four postgrad students, who
designed and implemented a language game module (dic-
togloss), and redesigned and implemented an improved pro-
nunciation module. The CARAMILLA project focused on cre-
ating a speech enabled version of an engaging and adaptable
activity, dictogloss, which integrates all skills in a focussed
game, and on redesigning and improving the system’s pro-
nunciation training module. These activities and the original
MILLA are currently being integrated with a more robust dia-
log platform (CARA) to form CARAMILLA. The system targets
two user groups with differing needs ‒ school language learn-
ers and teachers of Irish, and adult refugee learners of English
living in Ireland.

4.1. Dictogloss

A dictogloss is a well-known text reconstruction game, which
is ideal for implementation in a CALL system as it can be
played by one learner with a tutor (the system) or between a
group of learners with a tutor. This section describes the de-
sign and implementation of the spoken dictogloss module for
the CARAMILLA system.

4.1.1. Dictogloss Game Description

A dictogloss is a complete cloze text reconstruction exercise.
It is widely recognised as a useful language learning game and
the version implemented here is close to that described in
Rivoluci’s seminal ‘Grammar Games’ [23]. In the basic form
of the activity, learners are exposed to a text, either by reading
or listening, and then reconstruct some or all of it ‒ by filling
in blanks on a worksheet or by writing and rewriting fragments
to assemble a coherent text. The exercise aids acquisition of
syntax and vocabulary, and when performed orally, also aids
in developing listening skills. A major advantage of the ex-
ercise is its flexibility ‒ different areas of vocabulary and syn-
tax can be addressed at different levels of difficulty by sim-
ply changing the text used in the exercise. Thus, the game
can be very useful in context and language integrated lan-
guage learning scenarios, where learners learn the language
by learning about something else ‒ for example, for refugees
learning the past tenses while reconstructing an account of
the life of an important figure in the history of the host coun-
try drawn from a history resource or a current news story of
interest scraped from a news website.

The typical procedure for a spoken dictogloss exercise is as
follows:

1. Teacher asks learners to relax, clear desks, and listen ‒ there is no note-taking allowed
2. Learners listen as text is read at normal speed with nor-
mal intonation
3. When text is finished, learners write down words re-
membered from text on a scrap of paper
4. Teacher distributes a blank rectangular grid containing
enough cells for all the words (tokens) in the text, or
draws the grid on the board.
5. Learners take turns guessing words that were in text, teacher tells them grid co-ordinates and learners enter correct guesses into grid. Points are awarded for correct words.

6. Teacher provides support in the form of hints and encouragement as learners complete grid

The game has two distinct phases - in the first few turns the learners use the words they have remembered and written down. After they have exhausted the words from their wordlists, learners start to reconstruct the sentences in the text by inferring what possible words could fill the blanks in the emerging text. Seeing the string ‘I __ a __ car’, they will (implicitly or explicitly) realise that a verb is needed in the first blank and will start guessing verbs appropriate to the world of cars. The second blank will elicit adjectives. This hypothesis and test stage is the core of the exercise allowing learners to consolidate their syntactic and lexical knowledge in an interactive context. The constant rehearsing and hypothesising of possible phrases based on the known words activates latent vocabulary and with immediate feedback on guesses the learner refines their syntactic knowledge in the context of the spoken language rather than in isolated grammar exercises.

Added motivation can be provided by grouping learners into teams, playing against a clock, or the use of more detailed scoring systems to reward good guesses - when a learner proposes a word which is the correct part of speech for example. It is clear that several of these additions can be easily incorporated into an automatic version of dictogloss, with virtual players or additional characters, access to NLP tools such as POS taggers, and the use of scoreboards. The spoken dictogloss was implemented in JAVA as part of CARAMILLA as described below.

4.1.2. JAVA Implementation of Dictogloss

The dictogloss process was implemented in Java, by creating web pages using Java Servlets and JSP. Both an English and Irish version were implemented differing only in (1) where the text was retrieved from, in this case the Simple English and Irish Wikipedia pages, and (2) in the text-to-speech (TTS) systems used to read the text to the user. Simple Wikipedia was chosen as the language used is more suitable for learners in terms of lexical and syntactic complexity. Importantly, the sentences are short enough to sound like plausible spoken text, unlike standard Wikipedia articles where collaborative editing often leads to very convoluted embedding and lengthy sentences. The TTS system used for the English version was CerProc’s Caitlin (Irish-accent English language) [24], which provides a speech model relevant to learners living in Ireland. For the Irish version, the ABAIR synthesiser was used. ABAIR offers a number of voices representing the three main dialects of Irish, and is an ongoing initiative of the Phonetics and Speech Laboratory, Trinity College Dublin [25]. Thus, CARAMILLA can provide the learner with access to models of the dialects of Irish, which differ considerably. In CARAMILLA, for both English and Irish, we send TTS queries to web servers in Trinity College Dublin to get the synthesis results. Since we have very long texts for the dictogloss game module, lags can be introduced by sending the complete TTS query to the web server. In addition, if we send the whole text content for synthesis, there will be no significant pause between sentences, which sounds unnatural and makes comprehension difficult. Therefore, we split the long text into single sentences, which are sent for synthesis sequentially, resulting in less lag. This method has the serendipitous advantage of creating natural sounding pauses in the synthesised reading of the texts.

In this implementation of dictogloss, the game is browser-based and is accessed from CARAMILLA’s main menu. On the menu page, the user selects dictogloss and chooses whether they want to listen to or read the text and then chooses a topic from a currently predefined list. In the future the list could just be a list of subjects that may be of interest to the user and when they choose the subject current articles could be scraped from news websites or journals and returned. Once the user has chosen an action (read or listen) and a predefined topic, the text is scraped from the English or Irish Wikipedia page and a predefined number of sentences are formed for the user. If the user chooses the read option, text will be displayed for them, but if they chose to listen, the TTS system will be called. Once the user has finished reading or listening, they are taken to a page where a grid for the text is presented, as a series of underscores representing words separated by spaces, with the punctuation from the original text already filled in, indicating where sentences end. On this page the user has a textbox in which they can enter words that they think is in the text, modelling the scribble page used in the classroom scenario. If the word that the user types, regardless of case, is in the text, the word replaces the underscores in all the places it appears in the original text. When the user guesses the correct word they get points for the word. Currently each word is only worth a single point, but in the future words could have different points based on difficulty. While the user is going through and inferring which words come next, they may get stuck and if this happens the user has the chance to ask for help just as they would if they were going through the process with a teacher. By entering in the text ‘hint’ the user is currently given the missing word, but this is being extended to allow the system to give the user the definition of the word or a synonym of the word. In real life, added motivation can be provided by grouping learners into teams, playing against a clock, or the use of more detailed scoring systems to reward good guesses - when a learner proposes a word which is the correct part of speech for example. It is clear that several of these additions can be easily incorporated into later versions of dictogloss, such as including virtual players or scoreboards.

4.2. Implementation of Pronunciation Module

The pronunciation training module is based on the use of automatic speech recognition to compare a learner’s production of an utterance with a model. The intention is to provide pronunciation training for learners of different first languages (L1), as different L1 speakers are known to make characteristic errors.

4.2.1. Corpus of typical errors for pronunciation module

In order to feed our GOP module with some data, we created a pilot corpus of 50 practice sentences with common pronunciation mistakes in English by language background. Pairs of frequently confused English language phonemes for learners from 10 first languages (Arabic, Chinese, Croatian, Dutch, Finnish, French, German, Korean, Spanish, and Turkish) were collected from an expert knowledge website [26]. A carrier sentence for each phoneme pair was then created for use in the application. For each sentence, we created tips for how to
pronounce the commonly mistaken phoneme pair, including a visualisation of the vocal tract, and a phoneme transcription with lexical stress from the CMU pronouncing dictionary [27]. For example, for Spanish or French speakers, the phrase ‘These shoes fit my feet’ containing the I vs i sounds in the minimal pair ‘fit’ and ‘feet’ was used to test these commonly confused vowel sounds.

4.2.2. GOP

Pronunciation scoring is generally utilised in language learning applications to obtain global scores which tells overall goodness of proficiency on an utterance. However, global scores do not give specific information of where students make mistakes, which results in less useful applications [28]. Hence, pronunciation testing should give not only global scores but also local scores at phoneme levels, and learner should pay attention to which phonemes they cannot pronounce correctly.

One method to detect phoneme level errors is the Confidence-Measure (CM) based error detection that often uses Hidden Markov Model (HMM) [29]. A practical advantage of this approach is to utilise an Automatic Speech Recognition (ASR) system using speaker adaptation techniques such as Maximum Likelihood Linear Regression (MLLR). However, the CM-based approach uses the same feature set for all the phones, which might not be an optimal approach to detect explicit errors on a specific phoneme. Moreover, Speaker adaptation can bring over-fitting to a target user, which makes confidence scores even more unreliable. In contrast, Linear Discriminant Analysis (LDA)-based classifiers optimise acoustic feature sets for each phoneme, which results in significant improvement of error detection on particular phones [28]. However, their evaluations are quite limited to small Dutch phoneme sets, which cannot be generalised.

Our approach is based on Witt’s goodness of pronunciation (GoP) [29]. Figure 1 describes the basic algorithm. This algorithm calculates a distance between an answer (log likelihood of forced-alignment results) and a target (that of a phoneme recogniser) as follows:

$$ GOP_i(q_i) = \frac{\log(p(O|q_i))}{NF(O)} - \frac{\max_j^{N} \log(p(O|q_j))}{NF(O)} \tag{1} $$

where $q_i$ is i’th phoneme in an utterance and $NF(O)$ is the sum of log-likelihood of all frames in the observation.

Hence, we modelled explicit error networks for mother languages (e.g. Chinese, Korean, Arabic, and etc.). For this, we collected utterances where non-native speakers make common mistakes from a user group during the eNTERFACE workshop. Figure 2 depicts an example of the error networks.

$$ GOP_j(q_j) = GOP_i(q_j) + KGOP_i(q_j) \tag{2} $$

where $K$ is a scaling factor. The details to calculate $GOP_i(q_j)$ can be found in [29]. In addition, our system explicitly displays which type of errors a specific user makes. Errors are categorised into deleted (missing), substitute, inserted as the same way used in the evaluation of ASR system. Hence, the user could realise both their proficiency and explicit errors.

For reproduction, we implemented our system using HTK [30] and Sphinx 4[31] toolkits. For the robust recognition in the wild, we employed sub-band OSF-based voice activity detection [32].

5. Conclusion and Future Work

The dictogloss module is currently being extended to incorporate texts of different levels and on different subjects in both English and Irish, and is being tested with real users. The English language version is being piloted with adult refugee learners in a centre for language and integration courses in Dublin, while the Irish language version is being piloted with Irish language students in Trinity College Dublin. The pronunciation system is currently available in English only and is being tested with the refugee learners.

For future work on the two modules, we plan to integrate an animated avatar into the system. For pronunciation tutoring, apart from getting a score from GOP, it will be very beneficial for the learner to learn how to pronounce by seeing the lip movement from a virtual agent, provided that a suitably accurate avatar can be used. Moreover, when students are playing the dictogloss game, having a virtual agent acting as tutor to read the story and even to talk with them may provide motivation and add to engagement - this is a question we plan to explore, as is that of adding a second agent as a competing or collaborating ‘classmate’. We are exploring the open source animation toolkit Smartbody [5] to build a virtual avatar for our CARAMILLA system.

We are currently porting the existing functionality of the MILLA system to CARAMILLA. MILLA already contains two chatbots – a male character at beginner level and a more advanced female character providing free conversation practice.
MILLÀ's existing user record system, conversation, grammar and pronunciation modules will, combined with the two new modules, will result in a comprehensive learning environment. On the curriculum side, the plan is to eventually provide enough content to allow learners to complete an online portfolio of communicative tasks, modelled on the European Language Portfolios [33]. This portfolio, with learning objectives matching the requirements of the CEFRL, will provide a dynamic record of the learner's competence, both motivating reflection and self-assessment leading to further learning and linking progress to an internationally recognised assessment paradigm.

We hope that the system can eventually be used as a free web-based resource for language learners wishing to learn the English language needed to live and work in Ireland, and by learners of Irish worldwide.

6. Acknowledgments

The authors would like to thank the organizers of Enterface '16 and the University of Twente for making this work possible. This work is supported by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-NET (CHISTERA). JOKER project, JOKE and Empathy of a Robot/ECA: Towards social and affective relations with a robot, and by the Speech Communication Lab, Trinity College Dublin.

7. References


An attention based model for off-topic spontaneous spoken response detection: An Initial Study

Andrey Malinin, Kate Knill, Anton Ragni, Yu Wang and Mark J. F. Gales

University of Cambridge, Department of Engineering, Trumpington St, Cambridge CB2 1PZ, UK
{am969, mjfg}@eng.cam.ac.uk

Abstract

Automatic spoken language assessment systems are gaining popularity due to the rising demand for English second language learning. Current systems primarily assess fluency and pronunciation, rather than semantic content and relevance of a candidate’s response to a prompt. However, to increase reliability and robustness, relevance assessment and off-topic response detection are desirable, particularly for spontaneous spoken responses to open-ended prompts. Previously proposed approaches usually require prompt-response pairs for all prompts. This limits flexibility as example responses are required whenever a new test prompt is introduced.

This paper presents a initial study of an attention based neural model which assesses the relevance of prompt-response pairs without the need to see them in training. This model uses a bidirectional Recurrent Neural Network (BiRNN) embedding of the prompt to compute attention over the hidden states of a BiRNN embedding of the response. The resulting fixed-length embedding is fed into a binary classifier to predict relevance of the response. Due to a lack of off-topic responses, negative examples for both training and evaluation are created by randomly shuffling prompts and responses. On spontaneous spoken data this system is able to assess relevance to both seen and unseen prompts.

Index Terms: Spoken Language Assessment, Relevance Assessment, Deep Learning

1. Introduction

Automatic assessment systems are becoming attractive with a growing demand for assessment of English as an additional language [1]. They allow language assessment programmes to economically scale their operations whilst decreasing throughput time and provide testing on demand. Spoken language proficiency is assessed based on a candidate’s responses to a series of question prompts, such as ‘describe a difficult situation at work, why was it difficult?’ . These assessment systems operate on features derived from recordings of the candidate’s responses. Automatic speech recognition (ASR) is used to transcribe the responses to provide structured features, in addition to features derived directly from the audio. Modern systems, such as ETS’ SpeechRater [2] and Pearson’s AZELLA [3], typically assess pronunciation and fluency, and it is not clear to what extent content is assessed. Although pronunciation and fluency are highly correlated with spoken language proficiency, reliable and robust high-stakes assessment requires the assessment of the semantic content, construction and relevance of the response to the question prompt. Such a system should assess whether the candidate has given an off-topic response, either due to misunderstanding the question and/or memorizing a response. This is the problem addressed in this paper.

A standard approach to assessing topic relevance and off-topic response detection, both for essays and speech, is based on measuring the similarity between a response and the test question or prompt. Commonly, this is done by measuring the similarity between vector representations of responses and prompts, such as TF-IDF, Latent Semantic Analysis (LSA) [4, 5] or Latent Dirichlet Allocation (LDA) [6, 7]. There are two major deficiencies with this approach. Firstly, it is based on bag-of-words vector representations which lose sequential information important to evaluating the semantic content of responses. Secondly, such systems require having prompt-response pairs for all prompts in the test and can only assess relevance to prompts which they have seen in the training data. The approach proposed in [8] overcomes the first limitation. It uses a topic adapted Recurrent Neural Network Language Model (RNNLM) to rank the topic-conditional probabilities of a response sentence. However, this approach still requires having prompt-response pairs for all prompts and cannot assess relevance to new and previously unseen test prompts. Furthermore, re-training the system may be computationally costly. This limits the flexibility and increases the cost of deployment of such systems, as it is necessary to collect example responses to newly introduced prompts in order to have a system which is able to detect off-topic responses to these prompts. This work aims to overcome this limitation.

Recent work in the fields of Neural Machine Translation and Question Answering [9, 10] has come up with a number of attention-based deep learning architectures. Their key advantage is their ability to use an attention mechanism to extract relevant information from a variable-length sequence model in the form of a fixed-length embedding, conditioned on another embedding. This approach was used by [9] to achieve breakthrough results in English-to-French machine translation. Such approaches have also been successfully applied to assessment of multiple-choice questions [11]. In a related piece of work, a Recurrent Neural Network was used to extract optimal sequence features from spoken assessment in [12].

This paper presents an initial investigation of a novel neural attention-based model for assessing the relevance of spontaneous spoken responses to open ended prompts without the need to see them in training. This model uses a Bidirectional Recurrent Neural Network (BiRNN) embedding of a prompt to attend over a BiRNN embedding of a response. The resulting fixed-length prompt-conditional response embedding is fed into a binary classifier to predict the relevance of the response to the prompt. The model is trained on ASR transcriptions of spoken responses. Due to a lack of off-topic responses, negative examples for both training and evaluation are created by randomly shuffling prompts and responses. The model is evaluated using the area under a binary classification receiver-operator curve for measuring the similarity between a response and the test question or prompt.
(ROC AUC) metric. The ability of this model to assess the relevance and detect off-topic responses to prompts which were both seen, and crucially, not seen in the training data is demonstrated on spoken data from the Cambridge Business Language (BULATS) exam.

The rest of this paper is structured as follows: section 2 introduces and describes the proposed model, section 3 describes the data and experimental setup, section 4 contains the results and analysis, and section 5 is the conclusion.

2. Model

This section describes the proposed neural attention-based model for assessing the relevance of responses to prompts. The model is illustrated in Figure 1. It consists of four components: a prompt encoder, a response encoder, an attention mechanism and a binary classifier.

The proposed model assesses the relevance of responses to prompts by using the prompt to extract information from the response which is used to assign a relevance score. This is accomplished by learning to dynamically compute a representation (embeddings) of the prompt using the prompt encoder. This prompt embedding is used to attend over a representation (embedding) of the response via an attention mechanism, which should highlight the parts of the response most relevant to the prompt. Based on this information, a binary classifier assigns the probability of the response being relevant to the prompt.

The prompt (eq. 1) and response (eq. 2) encoders are Bidirectional Recurrent Neural Networks (BiRNN) [13] with LSTM recurrent units [14, 15] which process the words of the prompt and response, respectively. The prompt and response are represented by the word sequences \( \{w_p^1, \ldots, w_p^L\} \) and \( \{w_r^1, \ldots, w_r^L\} \). The prompt embedding \( \tilde{h}_p^r \) is computed by concatenating the final forward in time \( \tilde{h}_p^r \) and backward in time \( \tilde{h}_p^r \) hidden states of the prompt encoder (eq. 3). The forward in time \( \tilde{h}_r^i \) and backward in time \( \tilde{h}_r^i \) hidden states of the response encoder are concatenated at every time step to produce a hidden state \( \tilde{h}_r^i \) (eq. 3), which contains information about how the complete surrounding context relates to the current word.

\[
\tilde{h}_p^r = \text{LSTM}_p(w_p; \theta_p) \quad (1)
\]

\[
\tilde{h}_r^i = \text{LSTM}_r(w_r; \theta_r) \quad (2)
\]

\[
\tilde{h}_r^i = \left[ \begin{array}{c} \tilde{h}_r^i_L \\ \tilde{h}_r^i_r \end{array} \right] \quad (3)
\]

A fixed-length prompt-conditional embedding \( \tilde{f} \) of the response is computed as a weighted sum of the hidden states \( \tilde{h}_r^i \) of the response encoder given a set of attention weights \( \alpha_t \) via an attention mechanism (eq. 5). The attention weights for each hidden state are computed as a softmax (eq. 6), where the logits are given by a similarity function between the prompt embedding and the response hidden state. The similarity function (eq. 7) computes how strongly a hidden state of the response encoder relates to the embedding of the prompt. The parameters of the attention mechanism are \( \Theta = \{v_r, \Lambda_1, \Lambda_2, b_r\} \). This similarity function was used in [9] for neural machine translation. Alternative attention mechanisms, with different similarity functions, [16] and attention sharpening [10] could potentially be used.

\[
c = \sum_{t=1}^{T} \alpha_t \tilde{h}_r^i \quad (4)
\]

\[
\alpha_t = \frac{\exp(s_t(\tilde{h}_p^r, \tilde{h}_r^i))}{\sum_{t=1}^{T} \exp(s_t(\tilde{h}_p^r, \tilde{h}_r^i))} \quad (5)
\]

Figure 1: Neural attention-based response-prompt relevance model.

The fixed-length response embedding \( \tilde{f} \) is fed into a binary classifier \( f \) (eq. 7) which outputs the probability \( P(\text{rel}|w_r^i, w_p^i) \) of the response relating to the question. In this work \( f \) is a deep neural network (DNN) with parameters \( \Theta_f \).

\[
P(\text{rel}|w_r^i, w_p^i) = f(c; \Theta_f) \quad (7)
\]

This model is trained using minibatch stochastic gradient descent with a logistic loss error function (eq. 8) over all parameters \( \Theta = \{\Theta_p, \Theta_r, \Theta_f\} \). The model is trained on a balanced data set of prompt-response pairs containing both positive and negative examples of relevance.

\[
\mathcal{L}(\Theta) = \frac{1}{N} \sum_{i=1}^{N} t_i \log(P(\text{rel}|w_r^i, w_p^i)) \\
+ (1 - t_i) \log(1 - P(\text{rel}|w_r^i, w_p^i)) \quad (8)
\]

2.1. Relation to Previous Work

Previously proposed methods, such as [4, 17, 18, 19, 20, 8] require an active set of question or prompt representations to be maintained. Typically, these are vector representations of topic-based on TF-IDF, LSA or LDA [6, 7]. These are commonly constructed from example responses to the questions or prompts. Thus if a new prompt is introduced, there is a need to collect example responses and to re-train the model, both of which could be expensive and time-consuming, limiting the flexibility of deployment of such models.

The primary advantage of the proposed attention-based relevance model is that, unlike previous methods, it does not need to maintain an active set of topic (prompt) embeddings, but can automatically embed any prompt into the appropriate space.
via the prompt encoder. This also eliminates the need to precompute a set of topic representations from examples responses. All components of the model are trained jointly, which allows them to learn the necessary representations and transformations which make this possible. This allows the model to assess the relevance of responses to newly introduced prompts without the need to collect example responses to the new prompt or for the model to be re-trained. However, the model needs to be trained to generalize well in order to effectively handle unseen prompts, especially if they are quite different to the prompts seen in the training data.

Previous Deep Learning based approaches to off-topic response detection [8], which were also evaluated on the BULATS data used in this work, constructed a topic-adapted RNNLM which is conditioned on an active, fixed set of LSA topic embeddings trained separately on example responses. This is a discriminative sentence model conditioned on the topic P(w | w'). Topic relevance is assessed via the approximation in eq. 9. By using a uniform prior P(w | w') over topics, it is possible to induce an implicit generative model over topics via Bayes rule. Since a single response may be related to multiple topics to different degrees, relevance is assessed by taking the top-N highest ranking probabilities.

\[
P(\text{rel}|w', w) \approx P(w | w') \approx \frac{P(w'|w)}{\sum_{w''} P(w'' | w)} \quad (9)
\]

In contrast, the proposed model calculates \(P(\text{rel}|w', w)\) directly, and there is no need to use ranking to assess topic relevance. Furthermore, the proposed model does not have an explicit model of topic \(P(w' | w')\). Since a response can be relevant to varying degrees to several different topics, a potential disadvantage of the proposed model is that a certain amount of confusion can be introduced by having negative examples which are very similar to the positive example.

3. Data and Experimental Setup

A series of experiments were run to assess the ability of the proposed automatic systems to rate the relevance of responses to prompts. Data from the Business Language Testing Service (BULATS) English tests was used for training and test. The text for each response was generated using an ASR system. The 1-best recognition hypothesis was then passed to a relevance assessment system, which decided whether the candidate had spoken off topic by assigning a probability of whether the response was relevant to the prompt. To avoid a data mismatch, the recognition hypothesis was used both in training and test.

3.1. BULATS Test Format and Data

The BULATS Online Speaking Test has five sections [21]. This work focuses on the 3 sections where open ended prompts (which appear on screen) elicit spontaneously constructed responses:

- C Candidates talk about a work-related topic (e.g. the perfect office).
- D Candidates must describe a graph such as a pie or a bar chart related to a business situation (e.g. company sales).
- E Candidates are asked to respond to 5 open-ended prompts related to a single context prompt (e.g. a set of 5 questions about organizing a stall at a trade fair).

The 3 sections consist of 7 prompts in total.

3.2. Training Data Construction

As the data is taken from tests run with human examiners the responses are virtually all on topic. To produce negative, off topic training examples, the responses and prompts for both training and evaluation were shuffled. As was shown in [8], responses to prompts from the same section tend to be more similar so are more confusable. Thus, two topic shuffling strategies are considered: Naive, where prompts are shuffled across all sections; and Directed, where prompts are shuffled only within the same section [8]. Naive topic shuffling represents a more likely scenario, as real off-topic responses are unlikely to come from the same section. The data sets were balanced, so that for every response there are as many matched positive examples as there are mismatched negative examples in the training data. Thus, if more than one negative example is shown for a particular response, the positive example would be over-sampled the corresponding number of times. For multi-part prompts, which contain a main prompt that describes the overall question, and several (5 in this case) sub-prompts, all sub-prompts were prepended with the main prompt. These sub-prompts are considered distinct topics. During training, sub-prompts to the same overall prompt are considered competing negative examples to each other during shuffling.

3.3. ASR System

A speaker independent hybrid deep neural network - hidden Markov model (DNN-HMM) system is used for ASR [23]. The acoustic models are trained on 108.6 hours of BULATS test data (Gujarati L1 speakers) using the HTK v3.5 toolkit [24, 25]. A Kneser-Ney trigram language model is trained on this data and is then interpolated with a general English language model
trained on a large broadcast news corpus, using the SRILM toolkit [26]. This ASR system has a word error rate of 32\% on a Gujarati L1 ASR development set taken from the BULATs data. Performance on other L1s varies from 42-53%.

3.4. Model and Training Hyper-parameters

The proposed relevance assessment model was implemented in Tensorflow [27]. It consists of 2 BiRNN encoders with 400 LSTM recurrent units with hyperbolic tangent (TanH) non-linearities, 200 for the forward states and 200 for the backward states. The model was trained for 5 epochs with the Adam optimizer [28], with an initial learning rate of 1e-3, and an exponentially decaying learning rate with decay factor 0.96 per epoch. Dropout regularization [29] was used with a keep probability of 0.8, dropout was applied to all layers except for the LSTM hidden-to-hidden connections and the word embeddings. The binary classifier was a DNN with 2 hidden layers of 200 rectified linear (ReLU) units and with a 1-dimensional logistic output. The word embeddings, shared by both the response and prompt BiRNNs were initialized from an RNNLM language model trained on the TRN responses and were kept fixed during training. Four main models are examined in this work: models N1 and D1, with Naive and Directed topic shuffling of training data, respectively, and 1 negative example per response, and models N5 and D5, with Naive and Directed topic shuffling of training data, respectively, and 5 negative examples per response. N1 and D1 take roughly 2.5 hours to train while N5 and D5 take 12 hours to train in an nVidia GTX 980M graphics card.

3.5. Assessment Criteria

The models are evaluated using the area under a Receiver-Operator Characteristic (AUC), which plots the True Positive vs. the False positive rate at different decision thresholds. In order to be able to do this, negative examples (true negatives) need to be introduced into the evaluation data sets. This is done using the same method as for training, with one positive and one negative example for every response, both with Naive and Directed shuffling. It must be noted, that results are based on a particular instance of shuffling the prompts for evaluation.

4. Experiments

This section presents the results of investigations into the properties of the proposed model. Subsection 4.1 investigates several key properties of the model when all the prompts are seen. First, the baseline performance of a model trained on data with Naive topic shuffling and 1 negative example per response. Secondly, the effect of CEFR grade level [22] on relevance assessment performance is investigated. Thirdly, the effect of using training data with 5 negative examples per response is assessed. Finally, the effect of using training data with Directed topic shuffling is investigated. Subsection 4.2 investigates the performance of the model on unseen topics (prompts).

4.1. Baseline Performance

Table 2 and Figure 2 show the AUC scores for the baseline N1 model for all evaluation data sets. There are several notable trends in the data. Firstly, overall, the model achieves a high AUC of 0.95 on ALL evaluation data with Naive topic shuffling, and a lower AUC of 0.90 with Directed topic shuffling. This supports the findings in [8] which state that it is more difficult to distinguish prompts from the same section than from across sections. However, the AUC score of 0.95 reflects the more likely operating scenario, as Naive topic shuffling is more representative of real off-topic responses. This trend holds for all evaluation subsets. The performance on subset EVAL1 was highest, which reflects both the dominance of Gujarati L1 candidates in the training data as well as the better quality of the ASR transcriptions of responses of Gujarati candidates.

<table>
<thead>
<tr>
<th>Topic Shuffling</th>
<th>EVAL1</th>
<th>EVAL2</th>
<th>EVAL3</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.97</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>Directed</td>
<td>0.94</td>
<td>0.89</td>
<td>0.88</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 2: Baseline AUC scores for model trained on data with Naive topic shuffling and 1 negative example per response (N1).

Figure 2: ROC curve for model trained on data with Naive topic shuffling and 1 negative example per response (N1).

Table 3 shows how the AUC performance of the baseline N1 model varies with the CEFR level of the candidates. Clearly, AUC increases with increasing proficiency level from the lowest, A1, to the highest, C. This reflects both the increasing complexity of the response, allowing it to be more easily distinguished from a response to a different prompt, and the rising quality of the transcription - it is easier to correctly transcribe the response of a good candidate using ASR. This trend holds for all subsets EVAL1-3.

<table>
<thead>
<tr>
<th>Topic Shuffling</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.88</td>
<td>0.94</td>
<td>0.94</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Directed</td>
<td>0.82</td>
<td>0.88</td>
<td>0.91</td>
<td>0.93</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Table 3: Per grade level breakdown of performance on ALL for model trained on data with Naive topic shuffling and 1 negative example per response (N1).

The results of the investigation of the effect of using Naive vs Directed shuffling of training data, as well as the effect of using more negative-examples per response are presented in Table 4. Using 5 negative examples with Naive shuffling (N5) gives very high performance on both the Naive, and especially, Directed evaluation data. Clearly, as the model is exposed to a greater variety of negative examples it learns to generalize better. This performance boost relative to the model trained with 1
negative sample is illustrated in Figure 3. The same trend can be seen for models trained with Directed shuffling of the training data (D1 and D5). Interestingly, model D1 has similar performance on both evaluation sets, while model N1 has clearly better performance on the Naive evaluation set. This distinction is blurred for the N5 and D5 models, both of which have comparable performance on all evaluation datasets.

### 4.2. Performance on Unseen Prompts

In the above experiments all the prompts have been seen in the training data. This section considers the scenario where some prompts are not seen in training, investigating the proposed model’s ability to generalize to new prompts. Since real unseen prompt-response pairs are unavailable, 10-fold cross validation over prompts (topics) was used on the training and evaluation data. A fixed block of data, TRN-fixed (Table 5), is never removed from the training data, as it contains topics which dominate the training data and topics which do not appear in the evaluation set ALL. The TRN-xVal data was used in cross validation. A subset of ALL, called ALL-sub, without the dominant topics of TRN, was used for cross validation evaluation. All parts of related multi-part prompts are held out together.

The training data uses Naive response shuffling with 1 negative example per prompt, described in section 3.3. However, evaluation data responses are shuffled differently to the previous section, for these experiments. The prompts presented to the model are always either from the subsets which are seen or unseen in the training data. Evaluation responses are always new (not reused from the training data, same as in section 4.1), but can be related to prompts either seen or unseen in training. Three strategies for shuffling evaluation responses for negative examples are considered: seen, unseen and balanced. The first uses responses to seen prompts as negative examples, the second uses responses to unseen prompts as negative examples, and the last is an equal mix of the two. This produces six experiments: seen prompts with seen, unseen and balanced response shuffling; unseen prompts with seen, unseen and balanced response shuffling. The first three illustrate how well the model understands what relates to seen prompts and how well it generalizes to increasingly differing responses. The latter three experiments illustrate how well the model generalizes to new, unseen prompts. Generalization performance is increasingly stressed with seen, balanced and unseen evaluation response topic shuffling, since the responses become increasingly unfamiliar. Relevance probabilities are combined across all 10 folds to produce one ROC curve and AUC score for each experiment. These curves, and the associated AUC scores, represent the ‘average’ AUC on the data.

<table>
<thead>
<tr>
<th>Topic shuffling</th>
<th>N1 Naive</th>
<th>N5 Naive</th>
<th>D1 Directed</th>
<th>D5 Directed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.95</td>
<td>0.97</td>
<td>0.90</td>
<td>0.95</td>
</tr>
<tr>
<td>Directed</td>
<td>0.90</td>
<td>0.95</td>
<td>0.91</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 4: Comparison of using Naive and Directed training data and using more negative-examples on ALL evaluation data.

### 4.2. Performance on Unseen Prompts

In the above experiments all the prompts have been seen in the training data. This section considers the scenario where some prompts are not seen in training, investigating the proposed model’s ability to generalize to new prompts. Since real unseen prompt-response pairs are unavailable, 10-fold cross validation over prompts (topics) was used on the training and evaluation data. A fixed block of data, TRN-fixed (Table 5), is never removed from the training data, as it contains topics which dominate the training data and topics which do not appear in the evaluation set ALL. The TRN-xVal data was used in cross validation. A subset of ALL, called ALL-sub, without the dominant topics of TRN, was used for cross validation evaluation. All parts of related multi-part prompts are held out together.

The training data uses Naive response shuffling with 1 negative example per prompt, described in section 3.3. However, evaluation data responses are shuffled differently to the previous section, for these experiments. The prompts presented to the model are always either from the subsets which are seen or unseen in the training data. Evaluation responses are always new (not reused from the training data, same as in section 4.1), but can be related to prompts either seen or unseen in training. Three strategies for shuffling evaluation responses for negative examples are considered: seen, unseen and balanced. The first uses responses to seen prompts as negative examples, the sec-

<table>
<thead>
<tr>
<th>Data</th>
<th>#Topics</th>
<th>#Resp.</th>
<th>#Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRN-fixed</td>
<td>178</td>
<td>142.8K</td>
<td>6.8M</td>
</tr>
<tr>
<td>TRN-xVal</td>
<td>201</td>
<td>150.1K</td>
<td>6.6M</td>
</tr>
<tr>
<td>ALL-sub</td>
<td>201</td>
<td>2955</td>
<td>127.7K</td>
</tr>
</tbody>
</table>

Table 5: Topic, response and word statistics of the prompt-response data sets used for 10-fold cross validation.

The results presented in Table 6 show that once prompts have been seen in training data, the model has a clear understanding of what is relevant to them and is not sensitive to the nature of the negative-example responses. However, on unseen prompts there is a degradation of performance, which ranges from 0.78 to 0.72 as evaluation response topics shuffling changes from seen to unseen. Clearly, the model is able to generalize well to unfamiliar responses, and to a lesser degree, to new prompts, even in the extreme scenario (0.72 AUC). This is expected, as the model is exposed to a greater variety of responses than prompts. ROC curves for performance on seen
and unseen prompts with balanced response topic shuffling are shown in Figure 4.

5. Conclusions and Future Work

This paper presented an initial study of a novel neural attention-based model for assessing the relevance of spontaneous spoken responses to open ended prompts. This model uses a bidirectional recurrent neural network (BiRNN) embedding of a prompt to attend over a BiRNN embedding of a response. The resulting fixed-length prompt-conditional response embedding is fed into a binary classifier to predict the relevance of the response to the prompt. Due to a lack of off-topic responses, negative examples for both training and evaluation are created by randomly shuffling prompts and responses. The primary advantage of this model is that it is able to assess the relevance and detect off-topic responses to prompts which were both seen, and crucially, not seen in the training data.

Improvements could be added to the model in future work. For example, the model could be trained with dynamic sampling of negative examples during training in order to expose the model to a greater number of competing examples at lower computation cost. Furthermore, it is interesting to investigate what the attention mechanism learns, and how its focus over particular words in a response varies across prompts. Correlation of response relevance with grade level should be investigated. Due to time constraints, it was not possible to run 10-fold cross-validation on unseen topics using a model trained on more than 1 negative example per response. A more robust method for evaluation, such as using a greater number of samples, should be considered. The proposed method should be compared to previously proposed approaches, such as [8].

6. Acknowledgements

This research was funded under the ALTA Institute, University of Cambridge as well as the Engineering and Physical Sciences Research Council. Thanks to Cambridge English, University of Cambridge, for support and access to the BULATS data.

7. References

Empirical Evaluation of the Communicative Effectiveness of an Automatic Speech-to-Speech Translation System

Xinhao Wang†, Keelan Evanini‡

Educational Testing Service R&D
†90 New Montgomery St., Suite 1500, San Francisco, CA, USA
‡660 Rosedale Rd., Princeton, NJ, USA
{xwang002,kevanini}@ets.org

Abstract

In this study we evaluate the ability of a state-of-the-art speech-to-speech machine translation system, Skype Translator, to preserve the communicative effectiveness of its input speech. Specifically, we elicited Spanish spoken responses from 24 native speakers of Spanish in the following two scenarios: academic speaking (in the context of 6 questions from a standardized assessment of academic speaking proficiency) and daily communication (in the context of 6 map-based directions tasks). Skype Translator was then used to translate these Spanish responses into English, and expert human raters subsequently provided speaking proficiency scores using both holistic and analytic scoring rubrics for both the original Spanish responses and the corresponding automatically translated English versions. The results indicate that the average holistic scores for the academic speaking task decreased from 3.5 (for the original Spanish responses) to 1.9 (for the English translations) on a 4-point scale, and the average task completion scores for the directions tasks decreased from 3.4 to 2.0. While these results indicate that automated speech-to-speech machine translation does not yet fully meet the communicative demands of these speaking tasks, they also indicate how far the technology has improved in recent years and can serve as a benchmark for tracking further improvements in the future.

**Index Terms:** speech-to-speech machine translation, human scoring, communicative effectiveness

1. Introduction

Automatic speech-to-speech translation is a complex task that brings together multiple NLP and speech processing systems, including automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS). Early speech-to-speech translation systems used speech recorded under controlled conditions with restricted vocabularies and were therefore limited to specific topic domains and speaking styles. During the past two decades there have been several large-scale speech-to-speech translation research initiatives involving more authentic and complex spoken language, including TRANSTAC and BOLT (both funded by DARPA in the U.S.) and TC-STAR (funded by the European Commission), and these have helped to spur on dramatic improvements in the quality of automatic speech-to-speech translation. Recently, a few companies have publicly released free speech-to-speech translation systems that have received widespread media attention, including Google Translate¹ and Skype Translator². Given the continually increasing accuracy of these systems across an increasingly diverse set of conversational topics and situations, one may wonder whether they will someday be sufficient for enabling successful communication between speakers of two different languages in all situations, thus potentially rendering foreign language learning unnecessary. In this study, we explore this question by examining how well a state-of-the-art automatic speech-to-speech translation system performs in two different communicative situations.

Specifically, we employed the Skype Translator system, which is a state-of-the-art speech-to-speech translation system developed by Microsoft. Skype Translator currently enables near-real-time translation to and from the following 8 languages: English, French, German, Chinese, Italian, Spanish, Portuguese, Arabic, and Russian. English and Spanish were selected as the target language pair for this study, due to the fact that this was the initial language pair available in the first release of Skype Translator and that the performance for this language pair appears to be best. We empirically examined how effectively Skype Translator translates spoken Spanish into spoken English in two different communicative scenarios. One set of tasks simulated simple language that would be used with tourists; in these tasks, speakers provided directions in Spanish to locations on a map. The second set of tasks elicited more complex speech similar in nature to speech that would be used in an academic environment, such as at a university. The main research question targeted by this study is to determine to what extent the English speech-to-speech translation output based on the Spanish input is intelligible and meets the communicative demands of the two types of speaking tasks. This question was addressed by obtaining expert human scores across a range of speaking proficiency dimensions for both the spoken Spanish responses and corresponding English output produced by the speech-to-speech translation system and then comparing the two sets of scores to determine the strengths and weaknesses of the system.

This paper is organized as follows: first, Section 2 reviews previous studies concerning the use of speech-to-speech MT systems in the context of a language learning classroom environment; next, Section 3 presents the details of the data used in this study, including the speaking tasks, the spoken responses, and the human scores; Section 4 presents the empirical results; finally, Section 5 summarizes the main findings and provides suggestions for future research.

2. Prior Work

Several studies have investigated the impact of MT technology on foreign language learning, in particular, how it can be used as...
a pedagogical tool. For example, one potentially beneficial approach would be to use an MT system to translate a text in the target language into a learner’s native language and then have the learner revise the translated text by referring back to the original text [1]. This process of correcting the MT errors leads learners to focus on linguistic differences between the original text and the MT output. A similar approach is to use MT systems in a bi-directional manner, i.e., by translating texts into both the learner’s native language and target language and having the learner review and correct both sets of MT output either by using native-speaker judgments (for native language translations) or reference materials (for target language translations) [2]. [3] discusses these approaches along with others in a comprehensive overview of different potential classroom uses for MT tools and their associated advantages and disadvantages. While most of the studies using MT technology in a foreign language learning environment have focused on text MT systems (since high-quality speech-to-speech machine translation systems have only been made freely available for widespread use in recent years), a few have also looked at speech-to-speech translation systems. For example, [4] applied a speech-to-speech translation system in the context of a multilingual dialog system for queries about the weather; in this application, language learners were able to speak a question in their native language (English) and obtain an automated translation in the target language (Mandarin Chinese) that they could then use as a model for their spoken response.

Empirical evaluations of the quality of speech-to-speech translation systems typically employ metrics that can be computed automatically based on references (such as Word Error Rate for the ASR component and BLEU for the MT component) as well as metrics based on human annotation of important characteristics of the spoken utterance, such as fluency and adequacy [5]. Some studies have emphasized the importance of evaluating the quality of the speech-to-speech translation output to meet the communicative demands of the speaking task; for example, [6] demonstrates how conventional measures of speech-to-speech translation accuracy underestimate the functional accuracy of a commercial system that produced translations in the tourism domain. Similarly, the current study focuses on the ability of a speech-to-speech translation system to meet the communicative demands of a range of speaking tasks, in particular, by investigating to what extent the quality of the automated translations deteriorate from the original responses provided by native speakers.

3. Data

This section provides details about the speaking tasks that were used in this study, the methodology for recruiting participants and collecting spoken responses from them, the procedure for processing the responses using Skype Translator, and the human scores that were obtained for the Spanish spoken responses and English Skype Translator output.

3.1. Speaking Prompts

Speaking tasks from two different domains, academic speaking and daily communication, were included in this study in order to evaluate the performance of automatic speech-to-speech translation systems with different types of spoken language. For the academic speaking domain, six speaking tasks from a standardized assessment of English proficiency for academic purposes, the TOEFL iBT®, were included. All six tasks were drawn from a single TOEFL iBT test, and thus included two Independent tasks (in which test takers are asked to speak about a familiar topic or provide a personal opinion) and four Integrated tasks (in which test takers are provided with reading and/or listening passages on university-related and academic topics and are asked to answer a question about the content of these passages). The six TOEFL tasks (including the reading and listening passages) were translated into Spanish for this study; Spanish versions of the listening passages were recorded by native speakers of Spanish. For the daily communication domain, six questions eliciting directions based on two different maps that depict landmarks and streets in fictitious cities were included; an example of one of these map-based tasks is as follows: ¿Cómo puedo llegar al hospital desde el estadio?

3.2. Data Collection

24 participants with Spanish as their native language were recruited for this study and were compensated $50 for their participation. The speaking tasks (including stimulus materials such as reading passages, maps, etc.) were presented to them in Microsoft PowerPoint slides, and the built-in Windows sound recorded was used to capture their spoken responses via a headset microphone. The participants were given fixed response times for the academic speaking tasks corresponding to the response times in the TOEFL iBT assessment (45 seconds for the Independent speaking tasks and 60 seconds for the Integrated tasks). For the map-based tasks, there was no fixed response window; participants were instructed to advance to the next task when they had completed their response to the previous one; typical responses to the map-based tasks consisted of approximately 2-4 short sentences. The participants responded to the twelve questions in Spanish, and the audio file from each participant’s recording session was manually segmented into 12 separate audio files, one for each Spanish spoken response; a total of 288 spoken responses were collected.

3.3. Skype Translator Processing

Each Spanish response was input manually into Skype Translator on one computer by setting Stereo Mix as the recording input and playing the audio file. The following three types of output from Skype Translator were then captured on another computer that was connected to the first one via Skype: automatic speech recognition, automatic machine translation, and text-to-speech synthesis.

3.4. Human Scoring

Human raters were recruited from the authors’ organization to provide scores for the Spanish spoken responses and English Skype Translator output; separate rater pools were recruited for scoring the two different sets of responses. A total of 13 raters, all with experience scoring spoken constructed responses, were recruited to score the English responses; 9 of them scored the academic responses exclusively, and the other 4 scored the map-based responses. 12 raters, all native speakers of Spanish and 7 with prior experience scoring spoken constructed responses, were recruited to score the Spanish responses; 7 of them scored the academic responses exclusively, 3 of them scored the map-based responses exclusively, and the remaining 2 scored both.

---

3See the following document for sample TOEFL iBT speaking tasks: https://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf.
academic and map-based responses.

The academic responses (both Spanish and English versions) were scored using both holistic and analytic scoring rubrics. The standard TOEFL iBT scoring rubrics\(^4\) were used to obtain the holistic scores; analytic scores were obtained using rubrics focusing on the following 6 specific components of speaking proficiency: pronunciation, rhythm and intonation, lexical range and accuracy, grammatical range and accuracy, content appropriateness and accuracy, and cohesion. The map-based tasks (both Spanish and English versions) were only scored using analytic scoring rubrics, since pre-existing scoring rubrics did not exist for this task and the analytic rubrics provide more detailed information that can help highlight the strengths and weaknesses of the speech-to-speech translation system; the analytic scoring rubrics for the map-based tasks focused on the following 5 components of speaking proficiency: pronunciation, rhythm and intonation, lexical and accuracy, grammatical range and accuracy, and task completion. The holistic and analytic scoring rubrics both employed 4-point rating scales (with 1 representing the lowest proficiency and 4 representing the highest).

As an example of the analytic scoring rubrics that were used, Table 1 presents the scoring rubrics for the lexical range and accuracy dimension of speaking proficiency. These rubrics for lexical range and accuracy were part of the Language Use section of the scoring rubrics, which also included grammatical range and accuracy, and was prefaced by the following general description: This domain relates to the range and precision of the vocabulary and grammatical structures used. Control of a range of vocabulary as well as comfort with a variety of grammatical forms and structures result in efficient and effective expression of ideas. The analytic rubrics were shared across the academic and map-based tasks.

Benchmark responses were selected for both the holistic and analytic scoring rubrics to exemplify how the rubrics should be applied. Raters first reviewed the rubrics and benchmarks and then proceeded to rate responses that were randomly assigned to them based on the following constraints:

- All responses were double scored
- The same rater could not provide both ratings for a response

- Raters who scored the academic tasks first scored responses using the holistic rubrics and then scored responses using the analytic rubrics
- Raters who scored the academic tasks could not provide both the holistic score and the analytic score for the same response

The raters who scored the English responses produced by Skype Translator were informed of the nature of the study and were aware that the spoken responses were produced by an automatic speech to speech translation system based on original recordings of Spanish responses. They were instructed to score the responses in an identical manner as they would typically score English spoken responses, and to not modify their interpretation of the rubrics based on the fact that they were produced by a speech-to-speech translation system; this point was discussed during the training phase and the raters reviewed the benchmark responses they agreed about the interpretation of the rubrics for the English responses.

### 3.5. Sample Responses

Table 2 provides transcriptions of two sample Spanish responses that were collected in this study, one for an academic task and one for a map-based task, along with the resulting Skype Translator automatic speech recognition and machine translation output and the human scores that were obtained for each response (the scores shown are the average of the two human scores for each of the rubrics). Note that the raters provided scores in all cases based solely on the audio files (the original Spanish response or the audio file produced by Skype Translator’s TTS engine based on the automatic translation); the text transcriptions are presented in Table 2 to exemplify the procedure, but the raters did not have access to them.

### 4. Results

#### 4.1. Speech Recognition Accuracy

In order to evaluate the accuracy of the Skype Translator’s Spanish automatic speech recognition system, manual orthographic transcriptions were obtained from a native Spanish speaker for the Spanish spoken responses that were used as input to Skype Translator. To compute the WER between the transcriptions and the Skype Translator Spanish ASR output, punctuation and filled pauses (um, eh, etc.) were removed from both the transcriptions and the Skype Translator output. The overall WER across all Spanish responses was 0.310, and the WERs for the 12 different speaking tasks ranged from 0.268 to 0.340.

#### 4.2. Comparison of Proficiency Scores

In order to evaluate to what extent the Skype Translator system could meet the communicative demands of the speaking tasks, we compared the scores provided for the original Spanish responses with the scores provided for the English Skype Translator output. Since each response was double scored, the average of the two expert scores was used as the final score for each response; for this analysis, the benchmark responses with their associated gold standard scores were also included. Table 3 presents descriptive statistics (mean, standard deviation, minimum, and maximum) for the average of the two human scores for both the Spanish and English responses across all of the scoring rubrics. In addition, the table also presents descriptive statistics for the difference between the the average of the two human scores. The difference was calculated

---

\(^4\)https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf
Table 2: Two sample responses along with their associated Skype Translator output (automated speech recognition and machine translation) and human scores based on the average of two human scores for each of the rubrics

<table>
<thead>
<tr>
<th>Rubric</th>
<th>Spanish transcription</th>
<th>Skype Translator Spanish ASR</th>
<th>Skype Translator English MT</th>
<th>Scores (Spanish, English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic</td>
<td>Creo que el lugar adecuado para estudiar es el lugar que te da la tranquilidad, que te genere un ambiente propicio para, para tener la concentración y los recursos que necesitas para tu estudio, entonces tienes que tener una buena iluminación en un buenos recursos, si necesitas libros, computadores, eh un ambiente apropiado con poco ruido, eh una silla cómoda para que te sienten bien y y bueno sobre todo el entorno que te genere tranquilidad y y no distracciones para tener la concentración suficiente para estudiar.</td>
<td>Creo que el lugar adecuado para estudiar es el lugar que te da la tranquilidad que te genera un ambiente propicio para. Para tenerla con su bueno, me voy. No bueno, iluminación en los buenos recursos. Si necesitas libros computadores. La un ambiente apropiado con poco ruido en una silla cómoda para que te sienten bien, y y bueno, sobre todo el entorno que te genere tranquilidad, y, y no distracciones para tener la concentración suficiente para estudiar.</td>
<td>I think the right place to study the place you the peace of mind that generates an environment conducive to. To have her with her well, I'm off. Not good, lighting in the good resources. If you need computer books. The appropriate atmosphere with little noise in a comfortable chair so that you feel good, and and well, especially the environment that you generate peace of mind, and, and no distractions to have enough concentration to study.</td>
<td>holistic: 4.0, 2.0</td>
</tr>
<tr>
<td></td>
<td>Desde la tienda, sales hacia el restaurante a mano derecha, llegas al centro comercial, volteas a la izquierda y ahí encuentras la piscina después del centro comercial.</td>
<td>Desde la tienda salida hacia el restaurante mano derecha llegado centro comercial volteado a la izquierda. Y ahí encuentra la piscina del centro comercial.</td>
<td>From the store exit to the restaurant right hand arrived Mall flipped over to the left. And there is the pool of the Mall.</td>
<td>pronunciation: 4.0, 2.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>prosody: 4.0, 2.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>lexicon: 4.0, 2.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>grammar: 4.0, 2.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>content: 3.5, 1.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cohesion: 3.0, 2.0</td>
</tr>
<tr>
<td>Map-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

by subtracting the English score (for the Skype Translator output) from the Spanish score (for the original response). Since the scores for the Spanish responses were nearly always higher than the corresponding scores for the English responses, the differences are nearly always positive; however, there were a few cases in which the English response received a higher score than the Spanish response, as exemplified by the few negative values in the column displaying the minimum score difference in Table 3.

As shown in Table 3, the average scores for the Spanish responses are close to the maximum value of 4.0; however, not all speakers received perfect scores for all scoring rubrics, as might be expected given the fact that they are native speakers of Spanish. Since the scoring rubrics are based on how the different aspects of speaking proficiency contribute to how effectively the spoken response meets the communicative demands of each specific speaking task (as shown in Table 1), native speakers are not expected to always receive perfect scores. In particular, Table 3 shows that the analytic scores for the Spanish responses that were based on rubrics that focused primarily on the content of the response (content and cohesion for the Academic tasks and task completion for the Map-based tasks) were slightly lower than the analytic scores based on rubrics that focused primarily on delivery and language use (lexicon and grammar).

The main result shown in Table 3 is that the proficiency scores for the translated English responses are consistently lower than the scores for the original Spanish responses. All of the average differences between the two sets of scores except one are 1.2 or greater, with the largest average difference being 1.9. The mean scores for the English responses cluster mostly around 2.0 across the different rubrics for both the Academic and Map-based responses. These relatively low mean scores for the English responses and relatively large differences between the scores for the Spanish and English responses indicate that the automatically translated output produced by Skype Translator does not sufficiently meet the communicative demands of the speaking tasks. In particular, the large average differences between the Spanish and English holistic scores for the Academic task (1.7) and between the Spanish and English task completion scores for the Map-based task (1.4) indicate that the quality of the translated responses in meeting the communicative demands of the speaking task was substantially degraded.

Table 3 shows that the score differences between the Spanish and English responses were lowest for the pronunciation scoring rubric for both task types, with an average difference of 0.5 for the Academic tasks and 1.2 for the Map-based tasks. This result indicates that the text-to-speech component of Skype Translator generally produces translated English speech that is intelligible and natural sounding. However, the larger degradations in scores for the other analytic rubrics demonstrate how the English responses contain issues with other aspects of English speaking proficiency that obscure the communicative intent of the original Spanish responses.

Table 4 presents the Pearson correlations between the scores provided by the raters based on the different scoring rubrics for the Spanish spoken responses and the automatic spoken English translations produced by Skype Translator. As shown in the table, the correlations are highest for the holistic scores for the Academic tasks (0.419) and for the task completion.
scores for the Map-based tasks (0.456) and are quite low for the pronunciation and prosody scores for both tasks. This result is consistent with the distributions of scores as shown in Table 3: the Spanish holistic and task completion scores are somewhat lower and have relatively larger standard deviations whereas the Spanish pronunciation and prosody scores are nearly all 4.0 (which contributes to the low correlations).

4.3. Human Agreement

The human-human agreement statistics for the double ratings are presented in Table 5 in terms of quadratically weighted Kappa and percentage of exact match (the benchmark samples are excluded from this analysis). As shown in the table, the Kappa values are quite low for many of the analytic scoring rubrics, especially for the scores for the Map-based tasks. This is likely an indication of the difficulty of providing reliable scores for specific aspects of speaking proficiency for such short spoken responses, since the responses contain only minimal evidence for each of the proficiency aspects. Also, the low Kappa values are caused in some cases by the high levels of change agreement due to the fact that a large number of the Spanish responses received scores of 4; this is demonstrated, for example, by the exact match percentage of 0.943 for the pronunciation scores for the Map-based tasks.

5. Discussion and Conclusion

In this study, we recruited 24 native speakers of Spanish and elicited their Spanish spoken responses for 12 speaking tasks, 6 drawn from an assessment of academic speaking proficiency and 6 targeted to elicit map-based directions. Skype Translator, a state-of-the-art speech-to-speech machine translation system, was used to translate the Spanish spoken responses into spoken English. Expert human raters provided analytic scores covering a range of speaking proficiency dimensions for all responses as well as holistic scores for the responses to the Academic tasks. Analyses were then conducted to compare the differences between the scores for the original Spanish spoken responses and their automatically translated English counterparts. The results indicate that the quality of the translated English spoken responses is substantially degraded across all aspects of speaking proficiency and that the translated responses do not adequately meet the communicative demands of the speaking tasks.

Contrary to initial expectations, there was not a noticeable difference in quality between the English translations for the Academic tasks and the Map-based tasks. Both resulted in similar degradations from the scores for the original Spanish responses despite the fact that the language used in the Map-based tasks was less complex and therefore expected to be more amenable to automated translation.

While the results of this study indicate that automatic speech-to-speech translation systems are not imminently poised to completely meet the communicative demands in speaking tasks that involve open-domain, complex, spontaneous speech,
follow-up studies should be conducted on a regular basis, since the technology is improving rapidly. It is expected that the ASR performance will continue to improve in the near-term with the application of more sophisticated Deep Neural Network-based acoustic models and the availability of greater amounts of training data, thus leading to substantially lower word error rates. In addition, developments in Neural Machine Translation models continue to improve the quality of MT systems. With sufficiently accurate ASR and MT components, it is conceivable that an automatic speech-to-speech translation system could successfully translate spoken language to meet the communicative needs of a wide range of speaking tasks in the future, thus potentially obviating the need to learn foreign languages. Future research should therefore continue to investigate the performance of these systems and their impact on foreign language learning and pedagogy.

6. Acknowledgments

The authors would like to thank several colleagues who contributed to this project, including Melissa Lopez, Florencia Tolentino, and Ayana Stevenson for the data collection; Florencia Tolentino, Hillary Molloy, and Ben Leong for the data processing; Pamela Mollaun and Molly Palmer for leading the scoring effort; and all of the raters who scored the spoken responses.

7. References

Detecting listening difficulty for second language learners using Automatic Speech Recognition errors

Maryam Sadat Mirzaei¹, Kourosh Meshgi¹, Tatsuya Kawahara¹

¹Graduate School of Informatics, Kyoto University, Japan
mirzaei@iap.i.kyoto-u.ac.jp, meshgi-k@sys.i.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp

Abstract

This paper introduces a new approach to detect difficulties in speech for second language (L2) listeners using automatic speech recognition (ASR) systems. In this study, the ASR systems are viewed as a model to predict L2 learners’ listening difficulties and the ASR erroneous cases are analyzed to find useful categories of errors that can epitomize language learners’ transcription mistakes. Annotation of the ASR errors revealed the usefulness of several categories in predicting learners’ listening difficulties when watching TED videos delivered by American native speakers. Experiments with L2 learners of English confirmed that these categories lead to listening problems for the majority of the learners. One application to make use of these errors can be found in partial and synchronized captioning (PSC), in which only difficult words are selected and shown to facilitate listening, while easy words are hidden. Findings of the experiments attested that embedding the useful categories of the ASR errors into PSC improves learners’ comprehension.

Index Terms: automatic speech recognition, second language listening, error analysis, partial and synchronized caption

1. Introduction

For majority of the L2 learners, listening to authentic contents, which are created by native speakers (not specially designed for language learning purposes) is a very demanding task. There are many different sources of listening difficulties involved varying from lexical, speech-related factors to linguistic-related complications [1]. Among these, some speech related factors, such as the speech rate and perceptual difficulties are known as the prominent sources of problem for many language learners [2, 3]. When it comes to the automatic recognition of speech, ASR systems are also subjected to some errors, some of which stems from similar factors [4]. While human listeners have little difficulties in dealing with recognition of spoken language in acoustically challenging situations, ASR systems often lack the same robustness that is achieved by the humans [5]. This observation has been the source of motivation for the studies that investigated the ASR errors and HSR (human speech recognition) difficulties with the purpose of bridging the gap between the two and incorporating HSR findings to improve ASR performance [6, 7, 8, 9, 10]. The subjects of these studies are either a native speaker of the target language or non-native speakers with no knowledge of the target language (e.g., Japanese with no knowledge of French tested with French audio, which includes words with the maximum phonetic similarity between the two languages). Through such studies, the researchers attempt to improve the ASR performance and eliminate the ASR errors [11]. In this paper, however, we assume that some of the speech related difficulties, which prevent L2 learners from recognizing a speech can lead to the emergence of the ASR errors.

There are numerous factors accounted for L2 listening difficulties, some of which are also observed when investigating ASR systems’ errors. For instance, fast or too slow speech rate increases ASR error rate [12], similarly too fast or too slow speech rate can hinder L2 listening comprehension [2]. In frequent words are likely to be misrecognized by ASR systems [13] and also cause perceptual complexity for L2 learners [14]. The length of the word serves as a useful predictor of ASR errors [13], while strongly affects L2 listening recognition [15]. Automatic recognition of male speakers are more challenging for ASR systems [16], and L2 learners also find it more difficult to recognize male utterances [17]. Finally, perceptual difficulties in speech impede the recognition of both ASR systems and L2 listeners [3, 4].

The effect of speech rate, word frequency and word length are investigated in [18], therefore these factors are not considered in this study. The gender of the speaker and class of the words are very broad predictors of the ASR errors, thus, excluded from this study. Therefore, we investigate the perceptual difficulties in the speech focusing on ASR error categories that signals L2 learners’ listening difficulties.

2. ASR Predicts L2 Listening Difficulties

2.1. ASR Error Analysis

In this study, 52 TED Talks (~15 hours) were annotated by Julius ASR 4.3.1 [19], which is pre-trained on 780 TED talks using a lightly-supervised approach [20]. The ASR transcripts were compared with human annotations (available from TED website) using word-level alignment. The reason to select a trained ASR is to obtain a reasonable amount of ASR errors to analyze. For the same reason, only 1-best ASR hypothesis is used for error detection. Figure 1 shows that in total we had 16.7% ASR errors while the majority of the errors belongs to the substitution category. Since we are interested in misrecognition of words, our main focus will be on the substitution category.

2.2. Root-Cause Analysis

We performed a root-cause analysis on the ASR error substitution cases and found the following clusters: (i) homophones, (ii)
minimal pairs, (iii) negatives, (iv) breached boundaries, (v) verb inflection, (vi) noun inflections, (vii) determiners, (viii) interjections, (ix) derivational suffixes, (x) stop list, and (xi) unknown sources. The distribution of these errors is shown in Figure 2.

We labeled the ASR substitution error cases, as useful (i) if a similar misrecognition could be expected by L2 listeners and (ii) if providing the learners with these cases in the form of a caption can facilitate their recognition. Around 10% of the words are annotated by another annotator based on the same criteria to obtain the annotation agreement. We found a very high level of agreement (Cohen’s $\kappa = 0.81$), which refutes the subjectivity of the annotation. Figure 3 presents the usefulness ratio of each category based on the annotation results. The figure suggests that minimal pairs, homophones, negatives, and breached boundaries are the most useful categories of the ASR errors. This is in line with the findings of the studies on L2 listening difficulties [3, 21] and makes these categories of ASR errors the potential predictors of L2 listening difficulties.

2.3. Automatic Categorization of ASR Errors

We developed an ASR error analysis unit, which uses syntactical analyzers, large-scale corpora, and phonetic dictionaries to determine the categories of ASR errors. For each ASR error case, the ASR transcript and the original transcript are aligned and checked for the word pairs that can be confused with each other. Word lemmatizers, language-specific grammar rules, and COCA corpus [22] were used to detect verb/noun inflections, determiners, interjections, and derivational suffixes.

Homophones and minimal pairs are detected by comparing the phone transcription of the utterance (by ASR) and the transcript (using CMU phonetic dictionary). Homophones are words with different writings, but identical phone sequence (e.g., feet and feat /ฟิท ฟีต/). An exception to this rule is cases such as American and British spelling mismatches, that are handled in our implementation. Minimal pairs are the words whose phone sequences differ only in one phonological element (e.g., fund /ฟัน/ and fun /ฟัน/). To detect minimal pairs the words whose phone sequences have a Levenshtein distance of one are considered.

Breached boundaries are cases in which the boundaries of the perceived utterance are converged or diverged from the correct location when compared to the transcription, thus creating new word sequences (e.g., in close instead of enclose, thick atmosphere instead of to keep this fear). Many language learners cannot set the right boundaries between the words [23], and there is no comprehensive rule to detect such cases. In the rare case, the two phrases have identical phone sequences while the boundaries and the resultant words are different. The following four cases are derived from the linguistic studies that focused on language learners’ boundary misrecognition by investigating many cases that were misrecognized by the language learners. We found these categories very useful in detecting the major breached boundary cases:

- **Higher Frequency:** when the speaker uses less-frequent or out-of-vocabulary words, the listeners tend to associate the uttered words to high-frequency words, which are generally more familiar to them [24], similar to what happens in ASR systems when facing such words [25], e.g., achieve her way is heard as a cheaper way. To detect such instances of breached boundaries, the average of the frequency of the words in both the ASR and the original transcript are calculated and compared. In this calculation, function words –that have excessively high frequency– are excluded [24].

- **Stress Syllables:** strong syllables typically appear at the beginning of the words, so L2 learners tend to believe that the words begin with strong syllables. Therefore in most of the cases, they insert a boundary before a strong syllable, e.g., the skies instead of disguise. On the other hand, learners tend to merge the word starting with a weak syllable to the previous or the next word, e.g., twenty two instead of ten to two. Based on these findings detailed in [24], potential breached boundaries can be detected between the sequence of words in the ASR and the original transcripts.

- **Resyllabification:** learners tend to attach the final consonant of the word to the beginning of the next word [3] and create false boundaries, e.g., made out instead of may doubt.

- **Assimilation:** in this phenomena a sound morphs into a similar/neighbor sound in special patterns [26]. In some languages such as English these patterns are regular, and can be easily encoded into the system, e.g., Sam which instead of sandwich.

Among other situations where acoustic and speech artifacts impede word recognition for ASR and L2 learners, negative forms have the most influence over the comprehension of the speech. Negative form of modals (e.g., can instead of can’t) and negative prefixes (e.g., legal instead of illegal) are detected as negative cases to address the language learners’ difficulties.
Figure 4: Feature statistics of ASR Correct–PSC Shown. FR, SR and SP denotes word frequency, speech rate, and specificity, respectively.

3. Addressing L2 Listening Difficulties

In the following section, an overview of the Partial and Synchronized Caption (PSC) is provided, its advantages in addressing L2 listening difficulties are discussed, and the target ASR categories are included in this system to provide an Enhanced PSC that better assists L2 learners in the listening task.

3.1. Partial and Synchronized Caption (PSC)

To facilitate training L2 listening skill, PSC was developed\(^1\) to present the difficult words in the caption and hide easy ones [27]. In this caption, an ASR system is employed to align the transcripts with their respective speech segments (synchronization) and difficult words are selected from the transcript based on speech rate, word frequency, and specificity (partialization). This framework strives to find the most problematic factors for L2 listening by drawing upon studies on L2 listening difficulties. By evaluating individual learner’s proficiency level, this system adjusts feature parameters to realize a personalized caption for the individual learners. In addition, a stop list (including marginal words, propositions, etc.) and a repetition counter is embedded into the PSC to improve the word selection process.

The synchronization feature of PSC aids word boundary detection and promotes speech-to-text mapping. On the other hand, partialization prevents the learners from over-reliance on reading the captions and encourages them to listen more and read less. By using proper features, PSC is capable of providing the right amount of scaffold for different learners. This characteristic of the system is further enhanced by its adaptation to the learners’ proficiency. Another great advantage of the PSC system is that it is fully automated.

While word frequency and specificity accounts for lexical difficulties, speech rate is the only feature in the Baseline PSC system that represents acoustic and speech aspects of the listening material [2, 28]. However, there are a number of acoustic and speech factors that may cause difficulties for L2 listening such as hesitations [29], noise [30], speaker’s variations [17], and perceptual difficulties in speech [3, 31]. Among them we focus on the four target categories of ASR errors capable of predicting L2 listening difficulties: minimal pairs, homophones, negatives, and breached boundaries.

3.2. Enhancing the PSC

To improve the word selection in the Baseline PSC, we extended this framework with the ASR error analysis unit. Instead of discarding ASR errors, this unit compares the original and ASR transcripts to identify the source of the errors. If the detected ASR error source falls into one of the target categories, the word is decided to be shown in the Enhanced PSC.

To maintain the desirable textual density of the final caption, some of the most trivial words of the ASR Correct–PSC Shown category should be removed. Figure 4 demonstrates that most of these words are included in the PSC by speech rate feature, therefore a switching mechanism is designed to set a more strict threshold for showing such words when ASR recognized them correctly. In addition, some specific (academic) words (e.g., research, positive) are frequent in the contemporary language, hence, they could be less challenging for the learners. Therefore, a threshold based on ASR correct or erroneous cases is introduced to this category. Upon correct recognition of the specific word by the ASR, the frequency of the word is checked and if the frequency exceeds the threshold, the word will be hidden in the caption. Figure 5 compares the distribution of the word categories in the Enhanced PSC against the Baseline.

4. Experiments

To evaluate the performance of designated ASR error categories as predictors of L2 learners’ listening difficulties, three different tests were conducted. The participants were 38 Japanese and Chinese undergraduate students, with TOEIC ITP scores ranging from 450 to 560 implying that their proficiency level was pre-intermediate.

The test material is taken from annotated videos, filtered for native American speakers. From these videos, the “difficult” segments involving ASR error–PSC hidden cases (using Baseline PSC) were selected, which contained one of the four target categories (minimal pairs, homophones, negatives, and breached boundaries). The video segments were not repeated in the experiments.

4.1. Transcription Tasks

In this experiment, a short video clip (25 to 35 seconds long) was given to the participants. The video was suddenly paused, and the participants were supposed to transcribe the last 4 to 6 words (including the target word), that included the target words. This test was timed to prevent the participants from rethinking and reformulating and no clue was given about the tar-

---

\(^1\)http://sap.ist.i.kyoto-u.ac.jp/psc/#DEMO
get word(s) or time of the pause. Through this process, we tried to check the participants’ listening recognition when the video segment included an ASR erroneous case. As a control measure, some “easy” segments of the videos from ASR correct—PSC hidden is selected from the same video.

Table 1 demonstrates that participants’ scores on the easy segments are significantly higher than the difficult segments, which included four target categories of ASR errors. As a result, the findings indicate that these four categories are challenging for the participants compared to the easy segments of the same video. It can be concluded that the participants share the difficulty with the ASR systems in transcribing homophones, minimal pairs, negatives, and breached boundaries.

Table 1: Transcription Test: Transcription scores on difficult segments (ASR errors) vs. easy ones (ASR correct).

<table>
<thead>
<tr>
<th>Average Score in Transcription</th>
<th>Easy</th>
<th>Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homophone</td>
<td>81.9%</td>
<td>12.4%</td>
</tr>
<tr>
<td>Minimal Pairs</td>
<td>89.3%</td>
<td>14.3%</td>
</tr>
<tr>
<td>Negatives</td>
<td>83.3%</td>
<td>11.2%</td>
</tr>
<tr>
<td>Breached Boundaries</td>
<td>87.6%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Total</td>
<td>85.2%</td>
<td>16.2%</td>
</tr>
</tbody>
</table>

4.2. Caption Selection

In this experiment, similar to the previous experiment, the participants were supposed to transcribe the last 4 to 6 heard words when the video paused. PSC was provided during the video playback, except for the last sentence. After the transcription, the participants receive both Baseline or Enhanced PSC for the final sentence, and they were asked to choose the caption which provided better clues to overcome their listening difficulties. The idea is that after transcription, the participants become aware of their difficulties and misrecognition, hence they can select the most informative choice between the Baseline and the Enhanced version. To conduct a fair comparison we ensured that both captions have a similar number of shown words.

Table 2 shows that upon encountering a problem in transcription, the participants preferred the Enhanced PSC, which includes the ASR errors. This again shows that the participants shared the difficulty in recognizing the target word(s) with the ASR. In addition, the participants have selected the Enhanced PSC 61% of the times, indicating that the Enhanced version could better assist them with recognition difficulties and provided them with better choices of words.

Table 2: Caption selection test: the preferred caption of participants with respect to their transcription correctness.

<table>
<thead>
<tr>
<th>Transcription</th>
<th>Baseline PSC</th>
<th>Enhanced PSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Transcription</td>
<td>10.2%</td>
<td>3.2%</td>
</tr>
<tr>
<td>Incorrect Transcription</td>
<td>28.7%</td>
<td>57.9%</td>
</tr>
</tbody>
</table>

4.3. Paraphrasing

Paraphrase tests emphasize the recognition of a specific part of the listening material. To perform this test we randomly divided the participants into two groups, one group received the Baseline PSC along with the video and the other received Enhanced PSC. They watched a short video clip (10 to 15 seconds long) and tried to paraphrase the last sentence they heard, once the video was paused (they were given two paraphrase options to choose from). The paraphrases pivoted on the target word(s), therefore selecting the wrong paraphrase choice conveys the misrecognition of the target word(s) that were chosen from the four categories of the ASR error.

Table 3 shows that given the Baseline PSC, the participants could not resolve the listening difficulty and they chose the correct and incorrect choices chance-like. On the other hand, given Enhanced PSC that included the target word(s), the participants performed significantly better. This emphasizes the role of the word selection in PSC and demonstrates that Enhanced PSC (including the ASR errors) realizes a better word selection to foster listening comprehension.

Table 3: Paraphrasing test: the average scores of two groups of participants given Baseline vs. Enhanced PSC

<table>
<thead>
<tr>
<th>Group</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline PSC (G1)</td>
<td>50.9%</td>
<td>49.1%</td>
</tr>
<tr>
<td>Enhanced PSC (G2)</td>
<td>76.6%</td>
<td>23.4%</td>
</tr>
</tbody>
</table>

5. Conclusions

In this study, we introduced some categories of ASR errors as good predictors of problematic speech segments for L2 learners. An extensive analysis of the literature on the L2 listening and ASR errors indicated some similarities between the two. Furthermore, a careful investigation of the ASR substitution errors revealed that homophones, minimal pairs, negatives, and breached boundaries are among the most important categories to predict L2 learners’ listening difficulties. Experimentally, we showed that the designated categories of ASR errors are able to predict some of the L2 listening difficulties. Additionally, our findings revealed that incorporating these categories into the PSC framework can lead to a significant improvement in the word selection of PSC.

The current study considers ASR as a simplified and general model of L2 learners, however, the next step would be to make ASR systems similar to L2 learners in term of listening proficiency and make them adaptive to the different levels of the learners. This can be done through degrading the ASR system so that its errors can provide more useful instances for PSC on language learners with different proficiency levels. Moreover, this framework can be extended to other languages to be used as a universal training tool for L2 listening development. This can be realized by substituting the word-frequency corpora, ASR models, and syntactic analyzers. Furthermore, with regards to the breached boundary category, the current framework focuses on the detection of most dominant cases, however, detecting all possible cases of breached boundaries in ASR errors (which are also misrecognized by L2 learners) requires more investigation.

6. References


Predicting Clinical Evaluations of Children’s Speech with Limited Data
Using Exemplar Word Template References

Gary Yeung\textsuperscript{1}, Amber Afshan\textsuperscript{1}, Kaan Ege Ozgun\textsuperscript{1}, Kantapon Kaewtip\textsuperscript{1},
Steven M. Lulich\textsuperscript{2}, Abeer Alwan\textsuperscript{1}

\textsuperscript{1}Department of Electrical Engineering, University of California, Los Angeles, USA
\textsuperscript{2}Department of Speech and Hearing Sciences, Indiana University, Bloomington, USA

garyyeung@g.ucla.edu, amberafshan@g.ucla.edu, kaanege@g.ucla.edu, kkaewtip@ucla.edu,
slulich@indiana.edu, alwan@ee.ucla.edu

Abstract

The need for automated speech pathology diagnostic tools for children has increased in recent years. Such tools can help speech pathologists identify speech disorders in children at an early age. This paper introduces an approach to automated clinical evaluations of children’s speech using limited data. A database of ten normally developing first-grade children administered the Goldman-Fristoe Test of Articulation, 3rd Edition (GFTA-3) was recorded. Graduate clinicians evaluated the pronunciation of the rhotic sounds by evaluating words in the GFTA-3 containing the letter ‘r’. The rhotic sounds were specifically chosen due to their late acquisition in children. Experiments were performed attempting to predict the results of the clinical evaluations. Five children, judged to have proper rhotic pronunciations, were chosen as exemplar templates for the experiment. The remaining children, used for evaluation, were aligned in time to match the five templates using dynamic time warping, and the difference between a test child’s ‘r’ and a template child’s ‘r’ was measured using the cosine distance. Multiple linear regression on the difference scores was shown to be effective at producing predictions that were well-correlated with human clinical evaluations. Several sublists of words with rhotic sounds were used to evaluate the regression, and the sublist containing words with the most mispronunciations performed best. Further discussion includes how much each individual template contributed to the regression and how consistent the clinicians were at scoring children’s speech production.

Index Terms: children’s speech assessment, pronunciation evaluation, template-based

1. Introduction

In a number of clinical settings, automated assessments of children’s speech production can provide clear benefits and advantages. In particular, speech-language pathologists are currently required to both administer speech evaluations for children and design treatment plans for those children diagnosed with a speech delay or disorder. Automated speech assessments for children can allow clinicians to focus on therapy instead of time-consuming examinations. Additionally, human speech pathologists do not always agree in their evaluations of children’s speech. Machine assessments can maintain objectivity in judgments and assist clinicians in their evaluations.

Many systems for speech evaluations have been proposed since the 1980s. Several studies were performed by Kewley-Port and her colleagues in the 1980-90s using templates to evaluate speech [1, 2, 3, 4]. The Indiana Speech Training Aid (ISTA) used the best recordings from a subject as templates, and new templates were recorded to replace old ones as the subject’s pronunciation improved [2]. While this system was shown to be effective, it also required heavy clinician involvement since templates constantly needed to be updated.

More recently, the advancement of machine learning in automatic speech recognition (ASR) has led to a number of Hidden Markov Model (HMM) ASR systems for children’s speech evaluations [5, 6, 7, 8, 9, 10]. The Speech Training, Assessment, and Remediation (STAR) system achieved an $r^2 = 0.6$ when using phoneme likelihoods in a linear regression to assess the pronunciation of the phoneme /r/ [6]. Another system achieved a 76% agreement between ASR and human listeners when measuring children’s speech intelligibility [9]. A more recent study used HMMs for forced-alignment and the Mahalanobis distance to explore trade-offs caused by thresholding scores [10].

While ASR systems have improved dramatically in recent years, children’s ASR is still not as well-understood as adult ASR [11, 12]. Children’s HMM-ASR systems, as well as deep neural network ASR systems, generally require a sizeable amount of data to train and are highly dependent on the data used [13, 14]. However, clinical speech data (especially for child speech) are much more difficult to acquire than normal speech data, and it is impractical for clinicians to do the work of gathering enough data for such systems. More research is needed to enable the development of clinical evaluation systems that can be used with small amounts of training data.

This study proposes a method that uses a limited amount of children’s speech data to train a clinical evaluation system for children. We return to a template-based approach to tackle this low-resource problem. In this paper, we examined children’s pronunciation of rhotic sounds due to the late acquisition of these sounds in children [15]. The children were reported (by their parents) to have no history of speech, language, or hearing impairment in a prior screening interview. From this study, we hope to understand whether clinician perception of children’s pronunciations can be modeled using a small number of exemplar pronunciations.

The rest of the paper is organized as follows. Section 2 describes the data collection and clinical evaluation process. Section 3 describes the pronunciation evaluation system. Section 4 discusses the experiments and results. Finally, Section 5 concludes the paper with a brief summary and description of future work.
Table 1: Words containing rhotics in the GFTA-3: Sounds in Words divided into ‘onset’, ‘coda’, ‘medial’, and ‘cluster’ categories.

<table>
<thead>
<tr>
<th>Onset</th>
<th>Coda</th>
<th>Medial</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>Brother</td>
<td>Giraffe</td>
<td>Brother</td>
</tr>
<tr>
<td>Ring</td>
<td>Chair</td>
<td>Crown</td>
<td>Brushing</td>
</tr>
<tr>
<td></td>
<td>Door</td>
<td>Drum</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Finger</td>
<td>Frog</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Guitar</td>
<td>Green</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hammer</td>
<td>Princess</td>
<td>Truck</td>
</tr>
<tr>
<td></td>
<td>Star</td>
<td>Zebra</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teacher</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tiger</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spider</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Data Collection

2.1. Children’s Speech Data

This paper is part of a larger study by UCLA and Indiana University which aims to improve children’s speech-language pathology tools through a longitudinal analysis of children’s speech. The database currently being collected consists of recordings of elementary school children. All children were screened beforehand to ensure that they did not have any speech disorders. Each child was recorded taking the Goldman-Fristoe Test of Articulation, 3rd Edition (GFTA-3) [16]. GFTA-3 is a common standardized speech test which evaluates children’s pronunciation using clinically relevant utterances. Student clinicians tracked the quality of phoneme pronunciation of each child using the GFTA-3 assessment format. Each child was seated inside a double-walled sound booth with the student clinician who administered the GFTA-3. A SHURE KSM32 microphone was placed approximately 1 m in front of the child and 0.5 m to the side of the student clinician, who was facing the child. Audio was recorded at 48 kHz with 16-bit quantization.

Ten first grade children, aged between 6-7 years, were recorded for this study. All of the children made some pronunciation mistakes, but these mistakes were typical and developmentally appropriate for children of such an age. The GFTA-3 has two sections: “Sounds in Words” and “Sounds in Sentences”. All tests were administered by graduate speech-language pathology clinicians. For this paper, only the “Sounds in Words” data were used, in which each child was prompted by the clinician to say specific words in a picture-naming task. For example, to cue the child to say ‘table’, the clinician would show a picture of a table and ask, “What is this?” There were 21 words from the GFTA-3 containing 22 ‘r’ sounds in a variety of phonetic contexts. These words are listed in Table 1. The word ‘brother’ is listed twice as it has an ‘r’ in both onset cluster and coda positions. We will refer to the onset cluster ‘r’ in this word as ‘brother1’ and the coda ‘r’ as ‘brother2’. These words compose the master word list for this paper. As each of the ten children said the 22 words once (counting ‘brother’ as two words), there were 220 total word utterances. The ten children were separated into two groups, five in a “template” group and five in a “trial” group. The children chosen for the trial group included two children who were suspected of having more errors in ‘r’ pronunciation than the rest of the children while still being typically developing. The remaining children were divided in such a way to ensure the template group had minimal ‘r’ mispronunciations. As these children had few pronunciation errors in general, the division was randomly assigned.

2.2. Clinician Scoring

Ten graduate clinicians from Indiana University rated the quality of production of the ‘r’ sounds from the five trial children. Utterances were played in a random order to each clinician, and each utterance was played a total of 3 times. As such, each utterance was judged a total of 30 times, and each clinician made 330 judgments (3 judgments of 5 children saying 22 words each). Graduate clinicians rated the quality of a child’s ‘r’ production by clicking (with a computer mouse) within a circular bullseye, displayed on a graphical user interface, with three levels: ‘no impairment’ as level 1 (inside), ‘mild impairment’ as level 2, and ‘severe impairment’ as level 3 (outside). Such a rating system was familiar to the graduate clinicians from their prior clinical experiences and is commonly used to explain quality of pronunciation to children.

The distance between the center of the bullseye and the clinician’s selected point was chosen as the clinician’s score where the radius of the bullseye was normalized to 1. A histogram of one clinician’s scores on the evaluation of 330 rhotic phonemes. The scores are clearly separated into three groups. Clinician scores were in the range of 0 to 1 where 0 represented a perfect pronunciation and 1 represented a severe mispronunciation.

2.3. Word Lists

Various word lists were used in these experiments. Six different word lists were chosen as follows:

1. All words
2. brother2, chair, door, finger, guitar, hammer, spider, star, teacher, tiger
3. brother2, finger, hammer, spider, teacher, tiger

Figure 1: Histogram of one clinician’s scores on the evaluation of 330 rhotic phonemes. The scores are clearly separated into three groups. Clinician scores were in the range of 0 to 1 where 0 represented a perfect pronunciation and 1 represented a severe mispronunciation.
3. Pronunciation Evaluation System

3.1. Feature Extraction

Feature sets investigated included Mel frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) coefficients, and linear predictive cepstral coefficients (LPCC). For all feature sets, a window size of 25 ms, a window shift of 10 ms, a pre-emphasis filter with coefficient 0.97, and a sinusoidal lifter with coefficient 22 were used. A filter bank with 23 filters was used for the MFCC features. A 12th order linear predictive coding (LPC) polynomial was used for both the PLP and LPCC features. For all feature sets, 13 coefficients were extracted, and the 0th (energy) coefficient was removed for a total of 12 dimensions per frame. Utterances were downsampled to 8 kHz before feature extraction.

3.2. Template Setup

Of the ten first-grade children recorded, utterances from 5 children who were judged to have no rhotic pronunciation errors from the original GFTA-3 assessment, and few errors in general, were chosen to serve as templates. These children are referred to as “template children” for the remainder of the paper. All 22 words containing ‘r’ were used for each of the 5 template children for a total of 110 templates, 5 templates per word. For each template, 3 consecutive frames from the corresponding feature sets were chosen manually at the center of the rhotic sound as a region of interest (ROI). The ROI for each word utterance can be thought of as an exemplar pronunciation of ‘r’.

3.3. Evaluation Procedure

The 5 children not chosen to be templates were used to model clinician scores. We will refer to these children as “trial children” for the remainder of the paper.

The feature set of each word utterance from a trial child was aligned to the corresponding word from a template child using dynamic time warping (DTW) with Euclidean distance as a metric. For example, the utterance ‘door’ spoken by a trial child was aligned in time to match the utterance ‘door’ spoken by a template child. After time alignment, the cosine distance between the ROI of the template and the corresponding frames in the aligned trial utterance was calculated, averaged over 3 frames. The resulting distance served as a similarity measure of the trial child’s ‘r’ and the template’s ‘r’.

For each word from a trial child, the above procedure was repeated 5 times, once for each template child. As a result, each trial child’s word utterance had 5 different scores representing the similarity of the trial child’s ‘r’ and each template child’s ‘r’. Figure 2 illustrates this procedure. For subsets of the word list, these 5 similarity ratings were used as inputs to a multiple linear regression with the mean clinician score as the prediction.

An alternative procedure considered using all words from a subset of the word list to create a single template to represent each template child. In this case, all ROIs from the chosen word list spoken by a single template child were averaged to create a single exemplar ‘r’, one for each template child. The DTW alignment and choice of ROI for the trial utterance was identical to the first procedure. However, the cosine distance was computed between the ROI in the aligned trial utterance and the mean exemplar ‘r’ of the corresponding template child. This procedure can be illustrated with Figure 2 by simply replacing the single template ROI with the mean ROI as the input of the cosine distance. As with the first procedure, the 5 resulting scores were used in a multiple linear regression to predict the mean clinician score for each trial child’s word utterance. However, the procedure using mean templates did not perform as well and will not be reported.
Table 2: Results of the multiple linear regressions using similarity scores between trials and templates to predict clinical evaluations of rhotic phonemes for all six word lists. Both $r^2$ and adjusted $r^2$ are shown.

<table>
<thead>
<tr>
<th>Word List</th>
<th>MFCC $r^2$</th>
<th>adjusted $r^2$</th>
<th>PLP $r^2$</th>
<th>adjusted $r^2$</th>
<th>LPCC $r^2$</th>
<th>adjusted $r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.267</td>
<td>0.231</td>
<td>0.220</td>
<td>0.183</td>
<td>0.103</td>
<td>0.060</td>
</tr>
<tr>
<td>2</td>
<td>0.473</td>
<td>0.413</td>
<td>0.420</td>
<td>0.354</td>
<td>0.370</td>
<td>0.299</td>
</tr>
<tr>
<td>3</td>
<td>0.691</td>
<td>0.627</td>
<td>0.537</td>
<td>0.440</td>
<td>0.615</td>
<td>0.535</td>
</tr>
<tr>
<td>4</td>
<td>0.107</td>
<td>0.047</td>
<td>0.119</td>
<td>0.059</td>
<td>0.108</td>
<td>0.048</td>
</tr>
<tr>
<td>5</td>
<td>0.574</td>
<td>0.511</td>
<td>0.431</td>
<td>0.347</td>
<td>0.310</td>
<td>0.208</td>
</tr>
<tr>
<td>6</td>
<td>0.762</td>
<td>0.699</td>
<td>0.586</td>
<td>0.478</td>
<td>0.570</td>
<td>0.457</td>
</tr>
</tbody>
</table>

4. Experiments and Results

4.1. Regression Results

Table 2 shows the $r^2$ and adjusted $r^2$ results of the clinical evaluation regression models for the various word lists. In general, MFCCs performed the best out of all the feature sets. Figures 3a and 3b show the results of the regression scores plotted against clinician scores using MFCCs on Word Lists 3 and 6, respectively, which gave the two best regression results. Word List 6 with MFCCs gave the best regression results, modeling over 76% of the variance of the clinician scores. This is likely due to the fact that Word List 6 better represented mispronounced ‘r’ phonemes (higher clinician scores) while the other word lists may have overrepresented words with proper pronunciation (lower clinician scores). Additionally, Word List 3 gave decent regression results when used with MFCCs, modeling almost 70% of the variance of the clinician scores. This suggests that using the specific subset of syllabic ‘r’ sounds can improve the evaluation procedure. One noticeable issue is that Word List 3 did not have many words that were judged as severely impaired by the clinicians. As seen in Figure 3a, only a small number of points represented higher scores in the linear regression. As such, the results from Word List 3 may be questionable.

Interestingly, the regression results in most cases indicated that some of the template children contributed to the model significantly more than others. Table 3 shows the significance of contribution for the five template children from the regression using MFCCs and Word List 6. Only template child 1 and 4 contributed significantly in predicting the clinician scores. As such, the remaining templates were likely not reliable exemplars of properly pronounced ‘r’ phonemes as judged by clinicians. Recomputing the regression using only the two significant template children resulted in an $r^2 = 0.721$ and adjusted $r^2 = 0.695$, representing only a small decrement in performance.

Table 3: Significance of contribution of the five individual template children for the clinical evaluation regression model using Word List 6 with MFCCs.

<table>
<thead>
<tr>
<th>Template Child ID</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.006</td>
</tr>
<tr>
<td>2</td>
<td>0.220</td>
</tr>
<tr>
<td>3</td>
<td>0.314</td>
</tr>
<tr>
<td>4</td>
<td>0.002</td>
</tr>
<tr>
<td>5</td>
<td>0.789</td>
</tr>
</tbody>
</table>

In an attempt to improve the results, vocal tract length normalization (VTLN) was tested at the feature extraction step to improve alignment and template scoring. Various numbers of filters for MFCCs, LPC orders for PLP and LPCC, and window sizes were tested as well. Additionally, the Euclidean and Mahalanobis distances were also tested for similarity scoring between templates and trials. However, these approaches did not reveal any notable improvements in most cases and were thus discarded.

Figure 3: Clinician evaluation scores vs. scores predicted from the regression model using Word List 3 (top) and Word List 6 (bottom) with MFCCs. The line represents an ideal regression.
4.2. Discussion

One major point of interest is how well DTW was able to align trial utterances to templates to ensure that the ROIs corresponded to ‘r’ sounds. A manual inspection of the trial-to-template alignments revealed that only 2 of the trial-to-template alignments were misaligned out of 550 (5 trial children aligned to 5 template children for 22 words) when using MFCCs. In general, the alignment was successful. One of the cases of misalignment had an obvious mispronunciation in the trial word ‘ring’ as [wiŋ]. As a misalignment would likely cause the system to score the phoneme as mispronounced, we believe that these alignment mistakes are acceptable for identifying difficult words for children. The other case was in the word ‘brother2’ with no obvious pronunciation error. In this worst case scenario, the system would mistakenly classify this phoneme as mispronounced, which may be preferable to mistakenly classifying a phoneme as well-pronounced in some applications (e.g. a screening tool with high sensitivity).

Another point of interest is how difficult clinicians found the task of scoring the ‘r’ quality to be. Inter-clinician score standard deviations, defined as the standard deviation across the mean judgments for each of the ten clinicians, and intra-clinician score standard deviations, defined as the standard deviation across the three judgments from a single clinician, were computed for each word utterance. The average inter-clinician score standard deviation across words was 0.103, or 10.3% of the total range of possible scores. The maximum inter-clinician score standard deviation was 0.237 (23.7% of the range of possible scores) for a questionable pronunciation of the word ‘door’ by one particular child, indicating that some words and pronunciations were much less agreed upon across clinicians than others. The average intra-clinician score standard deviation across words and clinicians was 0.084 suggesting that clinicians were more consistent with themselves than with other clinicians. However, the maximum intra-clinician score standard deviation was 0.450 (45% of the range of possible scores) for a questionable pronunciation of the word ‘truck’ by one particular child, indicating that some pronunciations presented consistency issues for clinicians. Figure 4 shows histograms of inter-clinician and intra-clinician score standard deviations, which have most of their density in the lower standard deviations. The bimodal nature of these histograms indicates that most scores for a particular word either were within a single bullseye level or spanned two adjacent levels. The low peak between 0.3 and 0.5 in the intra-clinician score standard deviation histogram indicates that a small number of words was scored across all three levels. Table 4 shows the mean inter-clinician and intra-clinician score standard deviations for each word in the master word list. It is clear that some words, such as ‘princess’, caused more difficulty in scoring than others, indicated by a high mean score standard deviation in Table 4. In general, a large standard deviation in scoring an utterance was not correlated with the ability of the regression model to predict the mean clinician score.

Finally, we note that one fundamental difficulty in this study was the usage of only five trial children to predict clinician scores. Most of the poorly pronounced ‘r’ sounds were due to only two of the trial children, although the remaining three chil-

<table>
<thead>
<tr>
<th>Word</th>
<th>Inter-Clinician</th>
<th>Intra-Clinician</th>
</tr>
</thead>
<tbody>
<tr>
<td>brother1</td>
<td>0.130</td>
<td>0.100</td>
</tr>
<tr>
<td>brother2</td>
<td>0.135</td>
<td>0.104</td>
</tr>
<tr>
<td>brushing</td>
<td>0.132</td>
<td>0.100</td>
</tr>
<tr>
<td>chair</td>
<td>0.079</td>
<td>0.085</td>
</tr>
<tr>
<td>crown</td>
<td>0.083</td>
<td>0.071</td>
</tr>
<tr>
<td>door</td>
<td>0.085</td>
<td>0.085</td>
</tr>
<tr>
<td>drum</td>
<td>0.135</td>
<td>0.099</td>
</tr>
<tr>
<td>finger</td>
<td>0.090</td>
<td>0.075</td>
</tr>
<tr>
<td>frog</td>
<td>0.080</td>
<td>0.087</td>
</tr>
<tr>
<td>giraffe</td>
<td>0.088</td>
<td>0.098</td>
</tr>
<tr>
<td>green</td>
<td>0.090</td>
<td>0.082</td>
</tr>
<tr>
<td>guitar</td>
<td>0.100</td>
<td>0.057</td>
</tr>
<tr>
<td>hammer</td>
<td>0.103</td>
<td>0.073</td>
</tr>
<tr>
<td>princess</td>
<td>0.141</td>
<td>0.114</td>
</tr>
<tr>
<td>red</td>
<td>0.087</td>
<td>0.073</td>
</tr>
<tr>
<td>ring</td>
<td>0.143</td>
<td>0.078</td>
</tr>
<tr>
<td>spider</td>
<td>0.097</td>
<td>0.073</td>
</tr>
<tr>
<td>star</td>
<td>0.097</td>
<td>0.082</td>
</tr>
<tr>
<td>teacher</td>
<td>0.092</td>
<td>0.088</td>
</tr>
<tr>
<td>tiger</td>
<td>0.083</td>
<td>0.071</td>
</tr>
<tr>
<td>truck</td>
<td>0.067</td>
<td>0.069</td>
</tr>
<tr>
<td>zebra</td>
<td>0.123</td>
<td>0.078</td>
</tr>
</tbody>
</table>
5. Conclusion

This study proposed a framework to predict clinician scores of ‘r’ sounds produced by children. A database of ten first grade children was used. Speech utterances from five children were chosen as exemplar templates of ‘r’ production. The remaining five children were scored by clinicians and used to model clinician responses. The template ‘r’ and trial ‘r’ sounds were aligned with DTW, and the cosine distance between the phonemes was used as an automated scoring metric. The five scores from each trial word were used in a linear regression to predict mean clinician scores. Various word lists were used for regression, and it was found that the regression performed best when poorly pronounced phonemes were well-represented.

Future work will include expanding clinical scoring of children’s speech with more data, as well as taking into account various ages, dialects, and speech disorders.

6. Acknowledgements

The research was supported in part by the NSF.

7. References


Abstract
In this paper, we propose a deep context model based on recurrent neural networks (RNN) for grammatical error correction. For a specific error type, we treat the error correction task as a classification problem where the grammatical context representation is learnt from native text data that are largely available. Compared with traditional classifier methods, our model does not require sophisticated feature engineering which usually requires linguistic knowledge and may not cover all context patterns. Experiments on CoNLL-2014 shared task show that our approach significantly outperforms the state-of-the-art classifier and machine translation approaches for grammatical error correction.

Index Terms: Grammar error correction, deep context model, recurrent neural network

1. Introduction
Automated grammatical error correction (GEC) is an essential and useful tool for millions of people who learn English as a second language. In recent years, much work has been done including several shared tasks: HOO [1, 2] and CoNLL [3, 4]. The methods used in HOO and CoNLL are generally based on three types of methods: pre-defined rules, classification and machine translation (MT). Rule-based methods cannot cover all grammar error patterns and are usually used in combination with other methods. In the classifier approach [5, 6], GEC is cast as a multi-class classification problem, where a confusion set is specified for a given error type, and features typically consist of surface forms of text as well as linguistic abstractions (e.g., part-of-speech tags, and parse information). In the classifier approach, error types shall be defined clearly before they can be corrected. For example, the article classifier [5] corrects errors using maximum entropy classifier, where features are combinations of words and part of speech tags. Other classifiers including averaged perceptron and naive Bayes algorithm are also used for GEC [6]. In these methods, features must be designed manually and it is difficult to cover all situations, and as a result, manually-designed features may not be sufficient for GEC due to the complexity of language.

Another mainstream method is based on statistical machine translation (MT)[7] and aims to translate incorrect text into correct text. One advantage of the machine translation approach is that it can take advantages of both large-scale linguistic resource (web-scale language models) and error-corrected texts. However, phrase-based MT methods suffer from limitations of discrete word representation, linear mapping and lack of global context. Recently, the neural machine translation (NMT) method has been applied to the GEC problem using encoder-decoder framework [8]. The NMT approaches can cope with redundancy and non-idiomatic phrasing errors which the classification method cannot handle. Other neural network models like bidirectional LSTMs [9] are also used in grammatical error correction tasks.

While the MT approaches cover a larger variety of error types and are better at dealing with complex mistakes such as those where multiple errors interact, classifier approaches enjoy at least two advantages. Firstly, it does not rely on annotated learner data which are expensive but required by most MT approaches. Secondly, classifier approaches are easy to incorporate higher level context information that goes beyond the surface form. Many grammatical errors may benefit from generalizations based on POS or parse information, and indeed it has been shown that classifiers perform better on errors that require linguistic abstractions [10].

In this paper, we propose a novel classifier approach for GEC based on a deep context model. Instead of using surface and shallow features (POS, parse information, etc), we use deep features directly. In particular, we use bidirectional Gated Recurrent Units (GRUs) to represent context. Compared with traditional classifier approach for GEC, our new method does not require elaborated feature engineering for each error type. Deep context representations are learnt from large plain text corpora in an end-to-end fashion.

Note that learning context representations with task-specific optimization using labelled data has been applied to various NLP tasks, including word sense disambiguation [11], coreference resolution [12] and paraphrase detection [13]. Generic word embeddings, such as word2vec [14] and Glove [15], learned from the large scale corpus, also capture the semantic and syntactic information about each individual word. In those methods, there are effective neural network architectures modelling the context [16]. Indeed, context is essential for the word choice and can help us correct the grammatical errors. On the other hand, unlike in those tasks, where large amount of supervised data is usually required but only available in limited size, our approach for GEC leverages the abundant native plain text corpora and learns context representation and classification jointly to correct grammatical errors effectively. Experiment results on CoNLL-2014 dataset show that our approach significantly outperforms state-of-the-art classifier approaches as well as MT approaches for GEC.

2. Model
2.1. Model overview
For a certain error type, the corresponding model learns an embedding function of variable-length contexts around the target word and then predicts the target word with the context embedding. If the predicted word is different from the original target word, the original word is flagged as a mistake and the prediction is then used as correction. Our deep context model uses a bidirectional Gated Recurrent Units and is based on the text2vec’s [16] architecture. The context representation can be trained either from the beginning or the end of the sentence to the target word, or from the context words within a fixed-size
Figure 1: Deep context model for grammatical error correction

<table>
<thead>
<tr>
<th>Error Types</th>
<th>Values of y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>0 = a/an, 1 = the, 2 = None</td>
</tr>
<tr>
<td>Preposition</td>
<td>label = preposition index</td>
</tr>
<tr>
<td>Verb form</td>
<td>0 = base form, 1 = gerund or present participle, 2 = past participle</td>
</tr>
<tr>
<td>Subject agreement</td>
<td>0 = non-3rd person singular present, 1 = 3rd person singular present</td>
</tr>
<tr>
<td>Noun number</td>
<td>0 = singular, 1 = plural</td>
</tr>
</tbody>
</table>

Table 1: Values of y and their corresponding meanings for different error types.

Figure 1 illustrates how deep context vector represents the context and corrects text. We use two GRU recurrent neural networks. For the target word "go", we feed one GRU network with the context words ("to school everyday") from right to left. Given a context \( w_{1:n} \), our context vector for the target \( w_i \) is defined as the following equation:

\[
b_{tGRU}(w_{1:n}, i) = IGRU(w_{1:i-1}) \oplus rGRU(w_{i+1:n})
\]

where the IGRU is a GRU reading the words from left to right in a given context and rGRU is a reverse one reading the words from right to left. \( l/r \) represent distinct left-to-right/right-to-left word embeddings of the context words. After that, we feed the concatenated vector to the multi-layer perceptron (MLP) to capture the inter-dependencies of the two sides, at the second layer of MLP, we use a softmax layer to predict the target word or its status (e.g., singular or plural) of the target word:

\[
MLP(x) = \text{softmax}(ReLU(L(x)))
\]

where MLP stands for Multi Layer Perceptron, ReLU is the Rectified Linear Unit activation function, \( ReLU(x) = \max(0, x) \), \( L(x) = Wx + b \) is a fully connected linear operation. The final output of our model is

\[
y = MLP(b_{tGRU}(w_{1:n}, i))
\]

where \( y \) could be either the predicted word or the predicted status of the target word. If the prediction is different from the original word or its status, a grammatical error is detected and the prediction is used as correction. For different error type, \( y \) is defined in different ways as shown in Table 1. In the article model, if \( y \) equals 0, 1 or 2, it means the article should chosen “a/an”, “the”, or non-article respectively; In the preposition model, \( y \) represents the index of each preposition; In the verb form model, \( y \) denotes the form of the verb (0 is for the base form, 1 is for the gerund or present participle, and 2 is for the past participle); In the subject agreement model, \( y \) represent the noun-3rd person singular present, and \( 1 \) represents the 3rd person singular present. In the noun number model, \( 0 \) represents the singular noun, while \( 1 \) represents the plural noun.

We denote the labels of the classification as \( \hat{y} \), and the objective function of training is then

\[
loss = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i \log(y_i)
\]

where \( n \) is the number of training samples.

Note that the former deep learning method [8] considers all the errors unified and attempt to translate incorrect text into correct text, while our deep context approach learns a model for every specific grammatical error type.

3. Experiment

3.1. Dataset and metric

We evaluate the deep context model on CoNLL-2014 test dataset, which contains 1312 sentences. We use \( F_{0.5} \) as the main evaluation measure for error correction. \( F_{0.5} \) combines both precision (P) and recall (R), while assigning twice as much weight to precision, since accurate feedback is often more important than coverage in error detection.

The Precision, recall and \( F_{0.5} \) are defined as follows:

\[
P = \frac{e \cap g}{e} \quad R = \frac{e \cap g}{g} \quad F_{0.5} = \frac{(1 + 0.5^2) \times P \times R}{P + 0.5^2 \times R}
\]

where \( g \) is the gold standards of two human annotators for specific error type and \( e \) is the corresponding system edits. There are overlaps between many other error types and verb form error type, so \( g \) is based on the annotations of all error types when calculating verb form model performance.

We extract training samples from the wiki dump. In the experiments, we use the Glove word embedding [15] to initialize the word embedding which are later updated during the training process. We set word embedding size to 300. The input text is lowercased and all tokens which are not in the vocabulary are represented as a single \texttt{unk} token. The vocabulary is made up of the most 40000 common used words in the wiki dump.

3.2. Error types

We build deep context models for five common types of grammatical errors: article, preposition, verb form, noun number, and subjective agreement. For each error type, classifiers are trained independently.

NLP tools like Stanford corenlp tools [17] are used to locate the target words that need to be checked. If the prediction is different form the original label and the probability is larger than the predefined threshold, the grammatical error is deemed

\[\text{https://dumps.wikimedia.org/enwiki/}\]
to be found. For example, in the subject agreement task, we use NLP tools to extract the non-3rd person singular present words and 3rd person singular present word map relationships in advance. During the test, the tool can locate the verbs which should be checked by our model. If an error is detected, we can then use the extracted word mapping to correct the sentence.

English learners often have problems with when to use a (or an), the, or no article at the beginning of a noun phrase. We treat article error correction as a three-category classification problem: a/an, the and zero article. The position where the article can appear should be in front of noun phrases and we regard noun phrase as a combination of noun words and adjectival words in our grammatical error correction system.

Similar to article error correction, the subjective agreement task can be converted to a two-category problem: whether the verb should be third person singular present or not. We check every verb which is base form or the 3rd person singular present form.

We model the verb form error correction as a three-category problem: verb base form, gerund or present participle and past participle.

As for the preposition correction, we choose 11 most often used prepositions ("about", "at", "by", "for", "from", "in", "of", "on", "to", "until", "with", "against") as the classification labels.

Noun number correction can also be considered as a two-category problem: whether the noun should be plural or not. And we check all the noun words.

### 3.3. Window size

Correction of different types of grammatical errors might require dependencies from different distances. For instance, in subject agreement task, the status of verb can be affected by the subject which might be far away from the verb. In “frequently, the intention of the carriers does not want to tell their families is to continue their own ...” for example, the predicate “is” is far away from the subject “the intention”. On the contrary, which preposition can be used is determined by the words near the target word. In “to prevent the bigger problem from happen...” , “prevent from” is a collocation which is usually close to each other. Therefore, we use different context window sizes for different grammatical error types. For subject agreement and verb form, we use the whole sentence as context since these two error types typically require dependencies from context words that are far away from the target words. As for article, preposition and noun number errors, we introduce a window and only context words that are within that window are considered. The window size is chosen based on its performance on the CoNLL-2013 testset, as shown in Table 4.

<table>
<thead>
<tr>
<th>error type</th>
<th>window size</th>
<th>P</th>
<th>R</th>
<th>F_0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>preposition</td>
<td>3</td>
<td>4.84</td>
<td>1.93</td>
<td>3.72</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>16.7</td>
<td>3.86</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>24.3</td>
<td>2.89</td>
<td>9.8</td>
</tr>
<tr>
<td>article</td>
<td>3</td>
<td>35.3</td>
<td>24.2</td>
<td>32.5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>45</td>
<td>30.9</td>
<td>41.3</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>6.63</td>
<td>16.7</td>
<td>7.54</td>
</tr>
<tr>
<td>noun num.</td>
<td>10</td>
<td>50.0</td>
<td>32.8</td>
<td>45.3</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>51.3</td>
<td>33.8</td>
<td>46.5</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>52.1</td>
<td>30.8</td>
<td>45.8</td>
</tr>
</tbody>
</table>

Table 4: Performance of models trained using context of different window sizes
3.4. Lemma

Lemma is the base form of a word. For instance, words “walk”, “walks”, “walked”, “walking” all have the same lemma “walk”. In the noun number model, in addition to the existing context words around the target noun word, we also introduce the lemma form of the target noun word as extra “context” information, because whether the target should be singular or plural is closely related to itself. For example, for target word “apples” in sentence “many apples are...”, the left context is now “many apple”.

Table 5 shows that the noun number model fed with lemma can achieve better performance than the one without the lemma. So we choose to feed the lemma of the noun into the noun number model.

<table>
<thead>
<tr>
<th>contain lemma</th>
<th>P</th>
<th>R</th>
<th>F0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>26.0</td>
<td>13.3</td>
<td>21.8</td>
</tr>
<tr>
<td>Yes</td>
<td>50.6</td>
<td>30.3</td>
<td>44.6</td>
</tr>
</tbody>
</table>

Table 5: Performance of noun number model which is fed with and without the lemma

3.5. Result

Some examples of deep context model corrections are shown in Table 2. The grammatical error patterns can be captured with their context representation. The first and second sentences show that the verb form (from base form to gerund or present participle form and the reverse) errors can be corrected. The preposition error is corrected in 3rd and 4th sentences. Even though the subject is not near the verb, the error is still corrected in the 5th and 6th sentence. Even though the surrounding words are similar in 7th and 8th sentences, the deep model still successfully corrects these noun number errors. Article errors are demonstrated in the 9th and 10th sentences.

Table 3 shows type-specific performances of deep context model, and the best classifier methods in CoNLL-2014. Compared with the best classifier approach in CoNLL-2014 (CUUI), deep context model performs better on article, preposition noun number and subjective agreement error types. For all five error types, the deep context models have higher precision than the CUUI method. Our system is a more effective grammar error correction system since precision is more important than recall in GEC tasks. In fact, the precision numbers shown in Table 3 could be much larger. Since some error types interact with each other, the correct corrections are sometimes under-counted. For example, in sentence 11 in Table 2, the word “encourage” is corrected into “encourages”, but for this sentence “government” is annotated as “governments” in the gold-standard edits, and therefore the correct correction is then counted as a false correction. In other words, the errors which are left out in the gold-standard edits is one of the reasons for the under-estimated precision.

Lastly, we fix the mechanical errors (punctuation, spelling and capitalization errors) using existing resources and rule-based methods [10]. We measure the model performance and compare our system to several state-of-the-art systems on CoNLL-2014 shared task test dataset. The results are shown in Table 6. The top-1 system in CoNLL-2014 is a hybrid system combining rules and machine translation methods, while the top-2 system is a classifier based system (CUUI in Table 3). Our system outperforms these two systems significantly. We also compare our system with two more recent systems who have reported results on CoNLL-2014 testset. System in [8] considers all the errors in a unified way and attempts to translate incorrect text into correct text using an encoder-decoder recurrent neural network with an attention mechanism. Our system, although only addressing five common error types, achieves better results. In [10], the authors explore key strengths of both classifier approach and MT approach for GEC, and show that classifier approach actually does better in many aspects including the overall performance. We compare our system to the best classifier system in [10] without tailored training. As can be seen, our system also outperforms this system significantly.

4. Discussion and Future Work

We propose a new neural network architecture to learn context representation and then use it to correct grammatical errors. It outperforms state-of-the-art classifier and MT methods for GEC. Compared with traditional classifier method, our approach does not need complex feature engineering since the context feature representation can be learned jointly with classification in an end-to-end fashion, and the learning can be quite effective by utilizing enormous and easy-to-get native data. We find that different error types might require different amount of context information as shown in Section 3.3.

In the future we plan to introduce attention mechanism into our deep context model such that the model could focus only on those context words that affect grammatical usage. Also, as shown in [10], using a pipeline architecture where the MT is applied to the output of classifier can greatly improves the GEC performance. We believe that combining our deep context model with a state-of-the-art MT system will lead to further gain in the performance of GEC.

5. References


For some error types, only native data are used, while for some other error types, both native and learner data are used.

Table 6: Overall performance of deep context model compared with state-of-the-art

<table>
<thead>
<tr>
<th>system</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
</tr>
<tr>
<td>CoNLL-2014 top-2 system</td>
<td>41.8</td>
</tr>
<tr>
<td>CoNLL-2014 top-1 system</td>
<td>39.7</td>
</tr>
<tr>
<td>Xie et al.[8]</td>
<td>49.2</td>
</tr>
<tr>
<td>Rozovskaya et al.[10]</td>
<td>42.7</td>
</tr>
<tr>
<td>Deep Context Model</td>
<td>54.5</td>
</tr>
</tbody>
</table>


Studies of a Self-Administered Oral Reading Assessment

Jared Bernstein, Jian Cheng, Jennifer Balogh, Elizabeth Rosenfeld

Analytic Measures Inc., Palo Alto, California, U.S.A.
Jared @ AnalyticMeasures.com, Jian.Cheng @ AnalyticMeasures.com

Abstract

Reading assessments are most useful when they inform instruction. A prototype mobile-device app called Moby.Read presents a short self-administered test of oral reading fluency. Children read text passages aloud. Spoken responses are scored automatically on the device and displayed for teacher review. First, we report a usability study and a preliminary validation study with 99 children. Usability results are positive and score accuracy is high. Second, we discuss how accurate automatic reading assessment enables new analyses of reading performance (e.g., rate trends over time and within passage) that have not previously been available to guide individual instruction. We present early results of sub-passage and cross-passage results that hold diagnostic promise and discuss their potential to guide reading instruction.

Index Terms: oral reading fluency, comprehension, speech recognition, resilience, decoding, diagnosis, rate trend.

1. Introduction

Reading aloud is a convenient method for tracking progress in early reading. Informal reading inventories and oral reading fluency (ORF) tests yield observable evidence of a person’s reading processes [1]. ORF procedures are used to benchmark children’s reading in the early grades (ages 5-11), and ORF is often used to track progress in response to an instructional intervention. Although accurate automatic scoring of oral reading has been reported [2,3,4], teachers still spend about 4 million hours each year in the U.S. on administering and scoring basic reading assessments. Although ORF assessment is widespread, teachers and reading researchers have identified several problems including:

(a) inefficient use of teacher time [5];
(b) human scoring error from training deficits [5]
(c) emphasis on speed instead of expression of meaning [6];
(d) teacher uncertainty about how to use ORF scores [7]; and
(e) score instability from passage & individual factors [8],[9].

To address these problems, we developed an automated ORF assessment called Moby.Read, which is designed to improve the efficiency, specificity, and consistency of ORF measurement. Students take Moby.Read tests on a tablet computer, and it automatically scores oral readings and stores the recordings, which can be reviewed on the device.

Students self-administer the test and scores are reported automatically online. The app uses on-device speech recognition and scoring algorithms to return accurate reading rate (or words correct per minute: WCPM). Moby.Read’s automatic scoring will reduce the need for teacher training and help ensure consistency. Moby.Read’s scoring of comprehension and expression will emphasize reading for meaning instead of reading for speed, and the reporting will help teachers link assessment results to instruction. Finally, Moby.Read’s psychometric design will ensure better tracking of progress due to improved reading ability instead of spurious variability across passages.

2. Prototype App: Moby.Read

The prototype Moby.Read used in this research runs on an iPad as a stand-alone app. To get to the assessment, a simple, local ID is entered in the home screen. Entering a student ID causes Moby.Read to present one test session. When a Teacher ID is entered, Moby.Read will take the user to the Teacher Page where scores, trends, and recordings are reviewed.

The first time the student is taking a test, built-in video instructions are presented, explaining what to do and showing examples of a student performing the tasks. In each test session, students are asked to read a word list, read an easy practice passage, and read three grade-level passages. After reading the practice passage aloud, students are asked to retell the passage and answer two questions aloud. Figure 1 shows screenshots of the Moby.Read function flow for a practice passage.

After the second comprehension question, students are asked if they would like to listen to a reading of the passage and try reading it again. If the student hits “yes,” then the app presents the passage text again and plays a clear, expressive recorded reading of it. The text changes color in sync with the recorded reading, so children can easily follow along. Then the student is presented with the text again for a second reading (although this reading is not included in the overall score). Struggling readers liked this function because they gained confidence in their performance. Only the first, unpracticed reading is used in the overall score.

![Figure 1: Student flow for practice and test passages.](image)

After the practice passage, the student is presented with three grade-level passages (each followed by a request for a retelling, two comprehension questions, and the option to listen to the passage and read it again).
Figure 2 shows the teacher functions in Moby.Read. The teacher enters a Teacher ID to get to these pages on the device. The first page displays student names and IDs, and basic Oral Reading Fluency measures (rate, expression, comprehension). Selecting “View” activates the Graph screen, which shows a chart of the student’s performance over time.

Selecting “Hear” on the class page activates the Recording screen where teachers can listen to all oral responses (readings, retells, and Q-answers), view texts, and see scores.

The Moby.Read design has had several iterations, based on expert and user feedback. First, existing reading fluency assessments were surveyed and reviewed for flow, content and reporting structure to see what teachers and administrators would expect. A storyboard served as a catalyst for discussion of the product and feedback from several reading experts and teachers. This feedback led to several changes in the design.

3. Preliminary Usability and Validation

To understand the usability and accuracy of the Moby.Read prototype, a pilot validation study was conducted and in-situ usability tests were run in several classrooms. Two studies addressed five research questions:

A. Can students in grades 2, 3, 4 (age 6–10) use Moby independently?
B. Do students have a good experience with the app?
C. Are rate & accuracy scores similar to scores by human listeners?
D. Are scores similar to human+paper administered ORF scores?
E. Do teachers find the app useful, intuitive and/or convenient.

Study 1 was a usability study in which facilitators provided iPads with the Moby.Read app to classrooms and students ran the app. Usability data were collected from both students and teachers. Study 2 reports on 20 of the 94 students, who each had an adult administer the DIBELS Oral Reading Fluency (DORF) section of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) NEXT assessment [10], a widely used human-administered oral reading fluency assessment. Assessment order was balanced across students.

Participants in Study 1 were 99 school-aged children from four different elementary schools: two in New Jersey and two in California. The female to male ratio was 47:52. Ages ranged from 7 to 10 with an average age of 8. Students were enrolled in 2nd Grade (29%), 3rd Grade (40%) and 4th Grade (31%). Of the 99, 51% of the students were European-American, 19% were African American, 4% were Asian, and 25% were Hispanic or Latino. Twenty of these students also participated in Study 2. Four teachers gave usability feedback via a teacher questionnaire.

For Study 1, the 99 experimental sessions occurred during the normal course of a school day at the participant’s elementary school. Of the 99 student participants, 94 were able to run the App successfully. Several usability measures were gathered during the session: task completion, accurate reading rate in WCPM from the Moby.Read app, and a student judgment of how easy the app was to use (by selecting one of four emoji faces). After the sessions, teachers viewed score reports and could listen to student readings and retells. Then the teacher filled out the Teacher Questionnaire, which was a printed Likert-style survey.

For Study 2, a subset of 20 students took both a Moby.Read assessment and a DIBELS assessment. We used the fall benchmark form of DIBELS, which consists of three grade-leveled passages of about 250 words each. DIBELS administration and scoring followed procedures described in the test’s official documentation [10]. We collected a Moby.Read WCPM score and a DIBELS Next WCPM score for each of the 20 students. Section 3.1 reports Study 2 results.

3.1. Usability and Validation Results

A. Can students in grades 2, 3, and 4 use the app independently? Yes. Of the 99 students who attempted an assessment with the prototype Moby.Read, 94 were able to go through the app and provide responses that were scored and were usefully made available to the teacher. Of the five students who were not able to provide data, two had technical problems with the ear phones (not plugged in) and did not hear the audio instructions; and three spoke too softly or read silently.

B. Do students have a good experience with the app? Yes. The usability scale ranged over 1–4, as: totally confused = 1, not sure = 2, I knew what to do most of the time = 3, and easy = 4. Responses ranged from 2 to 4 (mean: 3.4), indicating most of the 94 students found Moby.Read easy to use and almost all knew what to do as the assessment progressed through the tasks.

In Study 2, each of the 20 students was presented with a forced choice question: Which one did you like better: the version on the iPad or on paper? Of the 20, 18 said iPad and 2 said both.

C. Are Moby’s rate & accuracy scores similar to scores judged by human listeners? Yes. Each response from the 94 student sessions was human scored by two expert raters with high inter-rater correlation. The correlation between human scores and automatic on-device Moby.Read scores was r = 0.96, which suggests that Moby’s recognition and scoring closely match human scores. Moby.Read responses are also uploaded to AMI’s servers. Server-based scoring featured more elaborate acoustic and language models. Server-based machine scores correlated with human scores with r = 0.987.

Figure 2: Teacher pages: class table, progress graph, text/audio access.

Figure 3: Session-level scatter of median WCPM; server-based scores vs. Human scores (r=0.987).
The human WCPM in Figure 3 is a session-level value that is derived by averaging same-passage WCPM values from two raters and then using the median value per session.

D. Are Moby.Read scores similar to human+paper ORF scores? Yes. \( r(0.88) \) The correlation found in Study 2 between Moby.Read scores and DIBELS NEXT scores was 0.88. Published studies of DIBELS report a test-retest reliability of 0.82 and an inter-rater reliability of 0.85 [11]. The reliability of an instrument limits the strength of the correlation between that instrument and other measurements. So, the correlation with Moby.Read is at the ceiling of what would be expected, given the reliability of DIBELS.

E. Do teachers find the app useful, intuitive and/or convenient? Yes. The teacher questionnaire assessed opinions as to how useful, intuitive, or convenient Moby.Read was, and responses were quantified on a seven-point scale. Responses: The average rating was 6.1, which suggests that teachers found the app useful, intuitive and/or convenient.

3.2. Acoustic Models

The acoustic model AMI used for speech recognition on device is a DNN-HMM one with four hidden layers, trained using Librispeech [12] 961 hours of clean native (L1) reading data. Moby.Read’s sample rate is 8,000. The DNN acoustic features are from a 40 Mel-scale filter-bank (FBANK). We concatenated a context of six feature frames left and right of the current frame (13 frames total). The number of senones (tied states) is 2064. There are several channel mismatch issues that may degrade the ASR performance: 1) an adult acoustic model was used to recognize children data; 2) we used narrowband; 3) the acoustic model used here was trained using very clean/quiet recordings, so Moby.Read scoring accuracy may diminish with very noisy data. Although we note these potential issues, the overall on-device acoustic model performance seems good.

4. New Analyses of Oral Reading

Automated scoring based on speech recognition opens new windows on reading processes. Sub-passage measurement enables us to see student reading performance in word-by-word detail. Reading rate can be calculated automatically for each word, or phrase, or window of N words in a passage. If we can automatically track reading rate fluctuations over the words and phrases in a passage, we can automatically diagnose which structures are most challenging for a student. If we collect and analyze a suitable sample of readings of a given passage, we can normalize that diagnosis with reference to passage and peer group.

The usual procedure in measuring oral reading fluency is to present three passages to read aloud and then report the median WCPM value from the three passages. Some readers are more familiar and fluent readers on one topic than on another. There are significant interactions in reading performance between individuals and passages, so taking a median WCPM value makes sense. However, this practice ignores the possibility that there may be an overall trend in WCPM within and between passages, and that there may be other valuable diagnostic information within each passage. For example, a reader may give a fluent reading of a passage with prosody that reflects and expresses meaning, but gets stuck on one word.

Especially for “struggling” readers who are reading below grade level, their accurate reading rate may decrease as a passage continues, or as they go from one passage to the next. Such a decrement may be caused by difficulty with particular words or topics, or by reading-fatigue that may cause some struggling readers to “give up” on the task and continue with a lowered level of active attention.

We report here on just one diagnostic measure of reading performance: Resilience (or perseverance). In particular, we seek evidence to support the hypothesis that some struggling readers may show a negative trend in accurate reading rate as Moby.Read’s passages continue or as the passage-after-passage administration continues. Such a negative trend is observed directly in reading rate (WCPM), but will be best understood when normalized against the rate trends of a cohort of readers.

5. Rate Data & Analysis

Each of 94 students with reportable data (24 in grade 2, 38 in grade 3, 32 in Grade 4) read three passages aloud in a Moby.Read assessment. For each passage reading from each student, we aligned the words in the oral reading response transcript with the original passage text by using a standard minimum edit distance. Then we moved a 10-word window over the passage source text, progressing through the text in 5-word shifts. So, for example, in the Zack-Backpack passage analyzed in Figures 4 and 5, there are 16 text spans. At each text span, when eight or more of the 10 source-text words were read aloud by that student in a correct order, we calculated that window’s read-aloud duration to compute WCPM in that window. If fewer than eight words (of the 10 in the source-text span) were read aloud, then that window was skipped and had no WCPM value. The window shifted to the right by five words and computation was repeated until we reached the end of the passage, producing a sequence of 16 WCPM values centered at every fifth word in the 89-word Zack-Backpack passage.

This running WCPM sequence has potential value for the struggling readers, as it yields information that locates the points in the text where readers are having trouble relative to their peers. This is of use in editing and preparing texts, as well as in providing formative feedback that teachers can use.

![Figure 4: Sequences of difference values of accurate reading rate (WCPM) for 10-word text-spans in one third grade story. At each point on the X axis, zero on the Y axis is the WCPM value averaged over all 38 3rd graders at that place in the passage.](image)

In Figures 4 and 5, the indices 1 to 16 are text spans. Span 1 is centered on word 5 and includes words 1-10; span 8 is centered on word 40 and includes words 36-45; span 16 is centered on word 80 and includes words 76-85. The upper line in Figure 4 shows average WCPM for the 11 fastest readers minus the
average value for that 10-word span across all 38 of the 3rd grade students. The lower line is the average value for the 11 slowest readers minus the overall averages. The source text for Figures 4 and 5 is one of 9 texts used in this study. It reads:

Zack sat down in his chair. It was time for class. The teacher asked for last night's homework. Zack opened his backpack to take out his homework folder. His homework folder wasn't in it. His recorder was gone too.

In the backpack there was a science book and a math book. They were both too hard for Zack. There was a piece of homework inside a green folder. The name on it said 'Jason'. Zack must have taken his brother's identical backpack while hurrying out the door this morning.

Data displayed in Figure 4 suggests that for relatively slow readers, the two hardest spans of text in this passage are #9 near word 45: …the backpack there was a science book and a math... and #12 around word 60: …hard for Zack. There was a piece of homework inside...

6. Results

There are many hypotheses that can be addressed with the within-passage data from Moby.Read, but for now we focus on questions about differences between faster and slower early readers. In particular:

Do slow readers show evidence of fatigue as a test progresses? Is rate variation different for slow and fast readers?

This data indicates that early readers may slow down over three passages that total about 270 words. Across all grades and readers, the average reading rate is 2% faster in the second passage, then 5% slower in the third passage.

Do slow readers slow down during a passage? Do they show differential slowing compared to faster readers? Analyzing data only from the 38 third graders who read Zach’s Backpack, data suggest that all readers generally slow down a bit over the course of a passage, and that the slower readers slow down much more. Using best-fit linear slopes, the 11 fastest readers (Fig. 4 upper line) lose about 4% of their reading rate over the length of that passage. On average, the 11 slowest readers lose about 54% of their reading rate over that same passage. The WCPM trends are very different, although Figure 4’s averaging and aspect ratio obscures this difference.

To examine the variation in reading rate, we calculated the standard deviation of the reading rates for each of the fastest and slowest readers, expressed as a percentage of that reader’s average rate. For the slow readers, the standard deviation averaged about 32% of their rate; for the fast readers, it averaged about 21% of their rate. An example of differential reading rate for two slow readers against the average reading rate of the slow group is shown in Figure 5. Note first that the rates of these two slow readers varies considerably from span to span, even though adjacent spans overlap. Second, note that the points of relative difficulty in the text are quite distinct for these two readers, which suggests that their sub-skill profiles are also likely to differ.

7. Discussion of New Analyses

This partial analysis of reading rate trends is consistent with the hypothesis that many struggling readers get discouraged and perform below their potential later in a given passage and in the later passages among a set. The finding of greater variability in rate for slower readers was expected and may offer the main key to providing useful formative information to teachers.

Note first that the passages were specified and written to increase in difficulty from beginning to end, so some part of the within-passage rate decrement is probably due to the passage structure. Second, note that the readings were presented in a fixed order to the 94 students whose data was analyzed, so the passage-to-passage rate decrement may partly be an artifact of differences in items difficulty.

8. Conclusions

Evidence from this small sample suggests that about 95% of U.S. school children (as young as 6 or 7) can successfully self-administer an oral reading fluency assessment as implemented in Moby.Read. Automatic scoring of these self-administered assessments correlated very highly with double human WCPM scorings of the same reading performances. The automatically scored Moby.Read tests predict and align with concurrent scores from standard, commonly used, adult administered ORF tests.

Preliminary results from sub-passage and inter-passage analyses suggests more (and different) instruction-relevant data can be automatically extracted from oral readings than has traditionally been inferred from measures of accurate reading rate (WCPM).

9. Acknowledgements

The research described here was supported by the Institute of Education Sciences, U.S. Department of Education, through the Small Business Innovation Research (SBIR) program contract ED-IES-16-C-0004 to Analytic Measures Inc. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

10. References


Speech perception training as a serious game in the EFL classroom

M.P. Hommel Utrecht Institute of Linguistics OTS, Utrecht

m.p.hommel@uu.nl

Abstract
Speech perception training improves listeners’ perception, but it could also significantly improve production. The benefits of perception training are interesting for the English as a foreign language (EFL) classroom, where this tool is currently not being used. This study looks at students’ perception and production performance on a pretest and, following perception training sessions, on a posttest and retention test. Results show that perception training improved students’ perception and to a lesser extent students’ production.

1. Introduction
1.1. Perception training and pronunciation
In the EFL classroom in the Netherlands, little attention is paid to production training and many Dutch students still struggle with the pronunciation of certain English speech sounds [1]. A tool that potentially helps improve students’ pronunciation is speech perception training. The intention of this paper is to find out whether this type of training significantly improves students’ perception as well as students’ pronunciation. Contrary to previous studies, this study will be done in a classroom setting and with many phonemes. Numerous perception training studies have shown that perception training significantly improves perception of the trained stimuli (e.g. [2], [3], [4], [5], [6]). Moreover, many studies have shown that perception training automatically improves speech production (e.g. [2], [3], [5], [6], [7]). Akahane-Yamada et al. [2] claim that "training in speech perception transfered to improvements in the production domain, suggesting a close link between speech perception and production". Neurological studies involving fMRI have shown an "overlap between the cortical areas active during speech production and those active during passive listening to speech" [8, pp. 371]. These scans demonstrate a relationship between perception and production. This evidence suggests that perception training could be a very beneficial tool for language learning in the classroom.

1.2. Gamification in the classroom
The perception training task in this paper will be a type of serious game. Adding game elements to classroom tasks, or gamification, suits the current generation of students. Game-based learning can turn “disconnected, bored learners into engaged participants” [9]. Games involve for example providing feedback, a high level of engagement and competition. Students get constant feedback, and “the more frequent and targeted the feedback, the more effective the learning” [9]. Games can also create a realistic context, rather than abstract information. Understanding the use and purpose of a task makes students more involved. When students are actively involved or engaged, they learn more than simply absorbing passive information. Yet another advantage of a game is that it provides leaners with the freedom to fail. Students can learn a lot from their own mistakes; it promotes trial-and-error learning [9]. Finally, games have a competitive element, either between players or within the player himself.

1.3. A large phoneme inventory for perception training
Nishi and Kewley-Port [4] reviewed earlier training studies and concluded that most of them focused on difficult L2 contrasts (e.g. English /æ/ for Dutch L1 listeners) and that none of these studies included more than five vowel pairs. Nishi and Kewley-Port trained one group of listeners with nine vowels (fullset) and one group with only three vowels (subset). The results showed that the fullset vowel training was more effective than the subset vowel training, which suggests that “efficient learning of nonnative vowels requires exposure to a full set of vowel categories” [4, pp. 1506]. Positive results in studies that used a small set of vowel pairs may be skewed as listeners could improve their performance by using vowel-specific cues or strategies “that may fail to generalize to the genuine perceptual learning that is needed to classify the complete set of vowels in the target language” [4, pp. 1507]. Kewley-Port et al. [10] looked at the effects of large training sets using vowels as well as consonants and found that speech perception training significantly improved listeners’ vowel and consonant perception. These findings suggest that a large phoneme inventory is most beneficial for language learners, which is why this study uses many phonemes for perception training. The main research question is: does perception training as a serious game improve students’ perception and pronunciation?

2. The experiment
2.1. Outline of the experiment
In order to find out whether perception training improves students’ perception and production in the classroom, an experiment was conducted. Dutch students who received perception training sessions were compared with a control group regarding their perception and production skills. First, a pretest was conducted testing students’ perception. In addition, students’ pronunciation was recorded. Next, the test group received perception training sessions. After the training sessions were finished, a posttest was conducted and students’ pronunciation was again recorded. Two months after the posttest, a retention test was conducted. The expectation was that the test group would outperform the control group on both perception and production.

2.2. Method of the pretest, training session, posttest and retention test
2.2.1. Design of the tests
SPATS-ESL (Speech Perception Assessment and Training System - English as a Second Language) contains a large inventory of 109 position-dependent phonemes (i.e., a phoneme could be in onset, nucleus or coda position) [11]. This large amount of English phonemes makes it unpractical to use all of them, as it would simply take up too much time in the English classroom; it is recommended to practice with SPATS-ESL for around 25 hours [11]. Another practical issue is that Dutch listeners might already achieve ceiling
performance in certain L2 vowels or consonants and therefore have little room for improvement. A pilot, the SPATS-ESL quick test, was conducted to solve the issue of manageability by removing phonemes with ceiling performance (i.e. >90% correct) and selecting the more difficult phonemes (290% correct). Fifty Dutch students (27 male, 23 female and M = 19 years, SD = 2.1) took the SPATS-ESL quick test. The average score was 83% correctly perceived phonemes (SD = 5.9%). Overall, 38 difficult phonemes (out of 109) remained as items for the pretest and training sessions. Both tasks consist of real words as opposed to non-words to create a realistic context, promoting student engagement.

2.2.2. Procedure of the tests
Eight native speakers with a Southern British English (BrE) accent were recorded in sound proof cabins in the phonetics lab at Utrecht University and the Max Planck Institute for Psycholinguistics in Nijmegen (both in the Netherlands). Audacity [12] software was used for the recordings (sampling frequency of 48 kHz). Because certain speakers had a louder voice and others a softer voice, the recorded data were set to an average intensity of 60 dB. Zep [13] was used to implement the four-interval forced-identification task where students would hear a stimulus and had to choose one of four answers (1 target, 3 foils). The listener’s task was to select the answer containing the speech sound that matched the one heard in the stimulus at the indicated position. For example, students would first hear ‘push’ and could then see four answers: cook, egge, ewg and coup. The right answer is ‘cook’ because it has the /ʊ/-sound in the underlined position as ‘push’.

It took a student around 10 minutes to complete the pretest and around 20 minutes to complete one training session. The posttest and retention test were the same as the pretest. The pretest and a training session contained the same target sounds, but the sounds were put in different words and there were different foils. The pretest and a training session also consisted of different speakers to see if listeners would generalise what they learned during the training sessions to novel items and voices. During the pretest, listeners received no feedback. During the training sessions, listeners received feedback on the correctness of each answer. This feedback gave students insight in their mistakes. At the end of either the pretest or a training session, students got to see their average score (in percentage correct). The competitive element was that students could try and improve their own test score in each training session, and compare their score to classmates’ scores. Students had five similar training sessions which took place once a week. The pretest took place in September 2016 and the posttest took place two months later. Two months after the posttest, the retention test was conducted to see if any positive results remained.

2.2.3. Stimuli of the pretest and training sessions
The pretest consisted of 38 items (19 fixed items spoken by a male and the other 19 fixed items spoken by a female). Each training session was the same and contained 2 × 38 items (38 items spoken by a male and the same 38 items spoken by a female). The items contained the phoneme (clusters) shown in Table 1.

2.2.4. Participants
The participants (initially test group: N = 92, of which 36 male and 56 female; control group: N = 65, of which 28 male and 37 female) were Dutch students attending secondary education (M = 19 years, SD = 1.3). Their average level of English was between upper-intermediate and advanced (CEFR level B2 - C1). Students in the test group received training sessions and students in the control group received regular English classes without training sessions. Of the initial 92 participants in the test group, 16 were dyslexic, 5 did not have Dutch as their native language and 23 students did not attend all 5 training sessions, leaving 49 students (17 male, 32 female) for data analysis of the training sessions.

Of the 49 students who attended all training sessions, 2 missed the posttest, resulting in N = 47 (17 male, 30 female) for data analysis of the posttest. Of the 65 students in the control group, 5 were dyslexic, 5 did not have Dutch as their native language, 1 missed the pretest, and 6 missed the posttest, resulting in N = 48 (21 male, 27 female) for data analysis of the posttest.

Quite some students missed the retention test, resulting in N = 29 (6 male, 23 female) for the training group and N = 38 (15 male, 23 female) for the control group.

2.3. Method of rating students’ production
2.3.1. Design of the production rating
Three native English speakers (BrE) rated the subjects’ production. As it was impractical for the native speakers to rate the pre-, post- and retention test production of all 38 items of at least 67 students, only 6 phonemes were chosen for native speakers to listen to, two in each phonological position (see Table 2). These 6 phonemes did not have an equivalent in Dutch, but three (one in each position) showed the most progress in perception for students who received training sessions, and the other three showed no improvement.

Table 1: The targets sounds by phonological position

<table>
<thead>
<tr>
<th>Onset</th>
<th>Nucleus</th>
<th>Coda</th>
</tr>
</thead>
<tbody>
<tr>
<td>/f/</td>
<td>/ʊ/</td>
<td>/ʊ/</td>
</tr>
<tr>
<td>/v/</td>
<td>/ɒ/</td>
<td>/ɒ/</td>
</tr>
<tr>
<td>/z/</td>
<td>/θ/</td>
<td>/θ/</td>
</tr>
<tr>
<td>/d/</td>
<td>/dʒ/</td>
<td>/dʒ/</td>
</tr>
<tr>
<td>/l/</td>
<td>/l/</td>
<td>/l/</td>
</tr>
<tr>
<td>/s/</td>
<td>/s/</td>
<td>/s/</td>
</tr>
<tr>
<td>/θ/</td>
<td>/θ/</td>
<td>/θ/</td>
</tr>
<tr>
<td>/z/</td>
<td>/z/</td>
<td>/z/</td>
</tr>
</tbody>
</table>

Table 2: The targets sounds for the production rating

<table>
<thead>
<tr>
<th>Onset</th>
<th>Nucleus</th>
<th>Coda</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ʊ/</td>
<td>/θ/</td>
<td>/θ/</td>
</tr>
<tr>
<td>/ɒ/</td>
<td>/l/</td>
<td>/l/</td>
</tr>
<tr>
<td>/dʒ/</td>
<td>/dʒ/</td>
<td>/dʒ/</td>
</tr>
</tbody>
</table>

2.3.2. Stimuli of the production rating
The pronunciation of twenty speakers in the test group and five speakers in the control group were rated. The retention test was included in the rating of the test group but excluded in the control group (6 stimuli × 3 tests × 20 speakers = 360 and 6 stimuli × 2 tests × 5 speakers = 60 items, which is 420 items
in total. Then, recordings with poor quality such as background noise were excluded. There was a total amount of 372 items for native listeners to rate.

2.3.3. Procedure of the production rating
The method used was a two-interval forced choice task. The spoken target sounds in the pretest were presented in pairs with those from the posttest, pretest with retention test, and posttest with retention test. The forced choice task consisted of two parts, each consisting of 186 items, and listeners could have a break after the first part. Excluding the break, it took them around 60-70 minutes to complete the task.

2.4. Results
2.4.1. Results of the pretest
Nine phonemes already showed ceiling effects (more than 90% correct) on the pretest: onsets /ʃ/, /z/, /g/, /s/, /r/, /b/, nucleus /ɪ/, and codas /ð/ and /θ/. The two phonemes with the lowest scores were nucleus /s/ (13% correct) and coda /ð/ (6% correct). These two phonemes also scored the worst of the nuclei and codas in a study by Cutler et al. [14]: 22% and 8% correct respectively.

There is a low to medium correlation between pretest score and a student’s performance on the Dutch national listening test ($r = .38$, $p = .01$). Also, there is a medium correlation between pretest score and a student’s performance on the Cambridge English test for schools ($r = .55$, $p = .01$).

2.4.2. Results of the training sessions
There was a significant improvement in test score due to test time ($F(4,180) = 36.22$, $p < .01$). The Bonferroni post hoc test indicated that the test group significantly improved after the first and second training sessions ($p < .01$), but that the third training session did not differ from the fourth and that the fourth did not differ from the fifth training session. These results suggest that students’ perception did not improve significantly after the third training session. The average score on perceiving the 38 target sounds after the first training session was already high ($M = 77\%$, $SD = 9.5\%$), and the score after the fifth training session was significantly higher ($M = 86\%$, $SD = 9.4\%$).

2.4.3. Results of the retention test
A multinomial logistic regression was carried out to determine the effect of perception training on test score. Time, group, and interaction of time and group as well as level were fixed effects in the model. These effects were all significant: time ($F(2) = 13.14$, $p < .001$), group ($F(1) = 5.41$, $p = .02$), time × group ($F(2) = 5.93$, $p < .001$) and level ($F(1) = 4.68$, $p = .03$). Sex was excluded as fixed effect because it was not significant. Intercepts for subjects and items were random effects in the model. The regression model explained 80.6% of the variation in students’ response ($F(7) = 8.96$, $p < .001$). The model showed that the test group outperformed the control group on the posttest and retention test. The odds of the test group having a higher test score than the control group was 1.8 times, so the effect size was relatively small. Moreover, the interaction between test group and pretest was significant, indicating that the test group performed worse on the pretest than on the posttest and retention test. The odds of the test group scoring better on the posttest and retention test than on the pretest was 1.6 times, also indicating a small effect size. In addition, two outperformed havo. The odds of two having a higher test score than havo was 1.4; this was again a small effect size. The effect sizes were small, but these results suggest that the training sessions improved students’ speech perception and that this improvement was retained (see Figure 1).

2.4.4. Results of the production rating
Reliability analysis showed a moderate reliability ($\alpha = .46$) between the native raters, but reliability would be lower if any native speaker was excluded. The inter-item correlation was low (between .20 and .24) for all listeners, which could mean that listeners were judging different things, but it is more likely that listeners had a hard time hearing a difference between many productions. As the task forced listeners to choose one production over the other, they sometimes had to ‘guess’ which production was better, hence rendering a relatively low Cronbach’s Alpha.

For each production, the raters’ response was categorised (production before training sessions was better or production after training sessions was better) and counted. The scores for the control group on the pre- and posttest did not differ, $\chi^2(5) = 10.53$, $p = \text{n.s}$, but the scores for the trained group on the pre- and posttest did differ, $\chi^2(5) = 27.25$, $p < .001$. Post-hoc analyses showed that students’ production in the test group significantly improved for three out of six phonemes (onsets /θ/, $p < .01$ and /ð/, $p < .001$ and coda /θ/, /ð/).
The scores for the trained group also differed between pretest and retention test, $\chi^2(5) = 15.45, p < .001$. Students’ production in the test group significantly improved for two out of six phonemes (onset /θ/, $p = .01$ and nucleus /ɔ/, $p = .04$). As expected, the scores for the trained group on the posttest did not differ from those on the retention test, $\chi^2(5) = 8.62, p = \text{n.s}$, indicating that the improvement in production remained over time. Raters considered some phonemes in students’ pronunciation after the training sessions as better than before.

3. Discussion

As mentioned in the introduction, fMRI scans only show a partial overlap between perception and pronunciation. Results here paints a similar picture. Three phonemes that showed progress in perception were the onset /θ/, the nucleus /ɔ/ and the coda /θ/. Three phonemes that showed progress in production were the onsets /θ/ and /θ/, nucleus /ɔ/ and coda /θ/. Only onset /θ/ improved in both perception and production as a result of perception training sessions. Nucleus /ɔ/, onset /θ/ and coda /θ/ only improved in pronunciation. Nucleus /ɔ/ seemed to deteriorate in production in the posttest, but this difference could not be found when production of the pretest and the retention rest were compared. One possible explanation for this finding is hypercorrection of the LOT vowel by Dutch students; the production of this vowel by Dutch speakers is already confused with the STRUT vowel by British English listeners [15], and perhaps hypercorrection made this confusion worse for these listeners.

The experiment conducted here shows that perception training with a large phoneme inventory and in a classroom setting improves perception and to some extent production. Data from this experiment will be scrutinised in order to see how training sessions affected each student individually. The next experiment that will be conducted looks at whether perception training also improves students’ word recognition and general listening skills. For future research it would be interesting to examine what the effectiveness is of perception training combined with explicit phonetic instruction on various language skills.

4. Conclusion

Similar to the findings of previous studies, perception training improves perception and to a certain extent production. Students improved their speech perception and generalised what they learned during the training sessions to novel items and voices, and retained their improved perception and production over time. This study shows that perception training also works in a classroom setting. It is plausible that perception training as a serious game improves students’ perception and production, which is why it is recommended to start using this tool in the foreign language classroom.

5. Acknowledgements

First and foremost, I would like to thank my supervisors René Kager and Willemijn Heeren. This work is part of the Doctoral Grant for Teachers with project number 023.007.008, which is financed by the Netherlands Organisation for Scientific Research (NWO).

6. References

Menusigne: A Serious Game for Learning Sign Language Grammar

Manny Rayner, Irene Strasly, Nikos Tsourakis, Johanna Gerlach, Pierrette Bouillon

FTI/TIM, University of Geneva, Switzerland


Abstract

We present Menusigne, a serious game designed to help beginner students learn basic sign language grammar. At the first level, the game uses a generation grammar and a signing avatar to let the student create signed utterances from menu-based patterns; at higher levels, the game presents avatar-generated or human-produced signed utterances, and the student uses the menus to indicate the meaning. The intention is to introduce the students to the principles of sign language grammar, and the game in particular emphasises the crucial role played by non-manual (non-hand) movements. We describe an initial course for teaching basic Langue des Signes Française (French sign language) to French students.

Index Terms: CALL, sign language, serious games, grammar

1. Motivation and background

To a greater and greater degree, sign languages are becoming accepted as having the same status as conventional (aural/oral) languages, a trend that is paralleled by changes in their legal status. There are now many hearing people, not just relatives of Deaf people1, who are interested in learning a sign language for reasons similar to those which might have led them to learn an auroral/oral language; because they wish to understand Deaf culture (there is a rich body of recorded Deaf literature), because they want to broaden their mental horizons, or simply because they are curious. This is particularly the case in the US and Australia, where ASL and Auslan courses have been popular for some time; other countries are catching up.

The idea of providing CALL tools to support students is as natural for sign languages as it is for auroral/oral languages and can be justified in similar terms. There are never enough language teachers. People do not want to pay for expensive tuition, or they cannot find time to attend classes. Conversation exchange (tandem learning) is unattractive for beginners, who feel that asking anyone to talk with them is an unwarranted imposition until they have acquired some basic fluency. In addition, some issues specific to sign language make CALL even more relevant than it is normally. A student can read an elementary Spanish textbook on the bus, and may well pick up useful grammar and vocabulary; reading a book on sign language is not nearly as useful, since sign languages have no accepted written form. Books for beginners tend to make heavy use of diagrams showing hand movements, but the problems with this approach are obvious. While signs are dynamic, paper is unfortunately static.

It is easy to see why CALL technology might be useful for students of sign language, but systems developed to date are rather simple, and basically amount to environments for showing and recording video clips [2], connecting teachers to students through the web [3], or in the best case performing word-for-word translation of an auroral/oral language into a sign language [4]. In this paper, we describe an initial attempt to build a more ambitious type of application, where we have combined grammar-based language generation and signing avatar technology to construct a simple web-deployed game. “Menusigne”, that helps beginner students practice understanding sign language2. Menusigne was built by reconfiguring resources from speech2sign [5], a platform wrapping software resources including the JASigning avatar [6, 7], which we have developed to support rapid creation of web-based speech to sign language translation applications. The game is freely available online. Instructions and a link to the live demo can be found at http://www.issco.unige.ch/en/research/projects/MenusigneDoc/build/html/index.html.

The rest of the paper is organised as follows. §2 presents an overview of the game from the user perspective. §3 briefly outlines speech2sign, and §4 describes how it was used to build the sign language game. §5 describes current content, and §6 initial reactions from users. The final section concludes and suggests further directions.

2. Overview of Menusigne

Menusigne is designed to help beginner students of Langue des Signes Française (LSF; French sign language) learn basic properties of the language. One obvious aspect is to teach a hundred or so signs, giving the student some initial vocabulary. More interestingly, it also aims to give the student a grounding in the elements of LSF grammar.

A common misconception concerning sign languages is that they are signed forms of spoken languages, with spoken words replaced by hand signs; so, in the present case, LSF would be in one-to-one correspondence with French. This view is entirely wrong. Sign languages have in virtually all cases a completely different syntax from the oral/aural language of the surrounding hearing community, and LSF syntax, in particular, is completely different from French syntax. It is also not true that sign languages are only rendered by hand signs. Movements of other parts of the body (“non-manual components”) have integral importance for nearly all sign languages. Important examples include shaking and nodding the head, eye widening and narrowing, direction of eye gaze, raising and furrowing of the eyebrows, and shrugging of the shoulders. Thus, for example, the sentence “I am Swiss” would normally be rendered in French as Je suis suisse; in LSF, it is rendered as a sequence of the two signs usually glossed as MOI (pointing to oneself) and SUISSE (making a cross over the left chest). This is already rather different from French; the differences become

1 We follow the widely recognized convention of using the uppercased word Deaf to describe members of the linguistic community of sign language users and, in contrast, the lower-cased word deaf to describe the audiological state of a hearing loss [1].

2 “Menusigne” is pronounced in the French style, to rhyme with “limousine”.

181
even clearer when we transform the simple declarative statement into a question or a negation. “Are you Swiss?” will in French be Es-tu suisse? or Est-ce que tu es suisse? In LSF, the sentence consists of a sequence of two signs glossed as TOI (pointing to the other partner) and SUISSE; the question marking is rendered using nonmanual elements, so the sign SUISSE is accompanied by lowering the head and slightly moving forward the right shoulder. Similarly, “I am not Swiss” is in French Je ne suis pas suisse, but in LSF is the three sign sequence glossed as MOI SUISSE PAS. Here, the sign PAS is a left-to-right movement of the raised forefinger, but the sentence is almost incomprehensible if the hand signs are not accompanied by a head-shake on the PAS.

The idea of the game is to introduce students to LSF vocabulary and grammar in a series of levels. Content is divided up into lessons, each of which contains a number of patterns presented in the L1, here French. Continuing the example from the previous paragraph, a simple pattern might be

\[ \text{je suis} \; \text{<NATIONALITY>} \]

This is presented to the student as two words of fixed text followed by a menu. In the first level of the game, the student uses the menu to produce signed utterances. For example (Figure 1), they can select SUISSE as the value of \text{<NATIONALITY>} and press “Submit” to see signed LSF for MOI SUISSE produced by the JASigning avatar. Other lines in the figure show a related pattern for \text{je suis} \; \text{<OCCUPATION>} and patterns for single signs.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{First level of game: the student selects items from the menus, and presses “Submit” to create avatar-animated signing.}
\end{figure}

When the student has experimented enough with the first level that they feel confident they know the vocabulary and grammar introduced, they can proceed to the second level, which turns the game around. This time, the app randomly creates avatar-based signing from the patterns, and the student responds by choosing from the menus to show that they have understood. For example, if the app signs MOI ETUDIANT (“I am a student”; ETUDIANT is signed by a gesture miming the student raising their hand in class), the student needs to use the fourth pattern, selecting ETUDIANT as the value of \text{<OCCUPATION>} (Figure 2).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Second level of game: the student presses “Random animation” to get avatar-animated signing, and then selects appropriate items from the menus to show they have understood.}
\end{figure}

Finally, the student may progress to the third level. This is structurally like level 2, but with the important difference that the student does not watch the rather artificial avatar animations, but recorded videos of signing produced by human signers (Figure 3). Thus, in three moderately easy steps, the student has progressed to understanding simple but nonetheless real signed language.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Third level of game: the student presses “Random video” to get a video of real human signing, and then selects appropriate items from the menus to show they have understood.}
\end{figure}

In the next two sections, we explain how the functionality described above has been implemented using resources taken from the speech2sign platform.

### 3. The speech2sign platform

The basic purpose of the speech2sign platform is to wrap resources for avatar-based production of sign language and grammar-based speech recognition in a way that makes it possible for people whose expertise is in linguistics rather than software engineering to create grammar-based speech-to-sign translation apps. Sign language animation is handled by the JASigning avatar [6, 7] and speech recognition by the commercial Nuance Recognizer platform. To avoid the necessity of installing various kinds of complex software, the platform is fully web-enabled. The details of how to write, upload, compile and run apps are presented in the online documentation [8].

For the app described here, speech recognition is irrelevant and we will only discuss the part of the system which has to do with sign language. There are three levels of representation, which correspond to the traditional notions of phonology, syn-
tax and semantics. Partly due to the special requirements of sign language, each of these is interpreted in a somewhat non-standard way. Starting at the bottom, the phonological level represents the simulated bodily movements used to create the signed language produced by the JASigning avatar. This level is encoded in SiGML [9], an XML-based representation based on the well-established Hamburg Notation System for Sign Languages (HamNoSys) [10]. A SiGML form is a sequence of elements, each of which consists of two parts, which respectively represent the manual (hand) and non-manual components. There is thus an implicit assumption that non-manual activity is always synchronised with some manual activity, which is generally believed to be true for non-manual activities that serve linguistic functions. A problem is that non-manual activities cannot be extended across several manual activities in a straightforward way; however, workarounds are possible [11].

As one would expect from an essentially phonological formalism, SiGML is not convenient as a representation level for describing syntax. The level used for syntax, the “sign table”, is an abstracted version of the SiGML. A sign table is a matrix with one column for each sign and one row for each channel of non-manual activity, modelling the basic structure of the SiGML representation but removing the phonological detail. Figure 4 shows the sign table for the two sign utterance realising the question “Are you Swiss?”. The top row shows glosses for the signs, and the others the non-manual activity. The question marking is in the head and shoulders rows for SUISSE. Note also the mouthing line: (silent) mouthing is an important part of sign language. Sign tables are similar to the diagrams commonly used in books on sign language linguistics [12, 13].

gloss TOI SUISSE
head Neutral Down
gaze Neutral Neutral
shoulders Neutral HunchRightForward
mouthing twe sui

Figure 4: Sign table for “Are you Swiss?” The non-manual part of the second sign expresses the question marking.

The third level is for the semantic representation. In most formalisms, this will be some kind of structured expression like a lambda-calculus form. We have however found it convenient to make semantic representations simply another string; for example the semantic representation we assign to “Are you Swiss?” might be

yn-question you swiss

The framework perhaps looks somewhat nonstandard, but the design choices are not random: the payoff for doing things this way is that the transformation rules can be made very simple. The transformation between the syntax and SiGML levels is specified by two tables; one maps glosses to HamNoSys strings, the other non-manual constants to SiGML non-manual forms. More interestingly, the mapping between the semantic and syntactic levels can be written as a set of rules in a version of Synchronised Context Free Grammar (SCFG; [14, 15]). The online documentation and the earlier papers referred to give more details; Figure 5 shows a simple example of how SCFG rules can be used to link together the semantic and syntactic levels for the running example.

Utterance
Source yn-question you $$nationality
gloss TOI $$nationality
head Neutral Down
gaze Neutral Neutral
shoulders Neutral HunchRightForward
mouthing twe $$nationality
EndUtterance

TrPhrase
TrPhraseId $$nationality
Source swiss
gloss SUISSE
mouthing sui
EndTrPhrase

(... more nationalities ...)

Figure 5: Toy SCFG rules that map semantic expressions representing “Are you <Nationality>?” into sign table form.

The framework contains one more construction, a mechanism for abstracting over rules. The top-level rule in Figure 5 translates the pattern exemplified by “Are you Swiss?”. We could duplicate the rule and modify it slightly to create a similar rule for expressions like “Are you happy?”, where the first element, as before, is a personal pronoun and the second an adjective expressing an emotional state. A more compact solution is to define a rule template, which can be instantiated to create both rules. Figure 6 illustrates in schematic form how this works.

Template ynq_pron_property PROPERTY
Source yn-question you PROPERTY
gloss TOI PROPERTY
head Neutral Down
gaze Neutral Neutral
shoulders Neutral HunchRightForward
mouthing twe PROPERTY
EndTemplate

Apply ynq_pron_property "$nationality"
Apply ynq_pron_property "$emotional_state"
(... rules for $nationality...)
(... rules for $emotional_state...)

Figure 6: Toy rule template that generalises over semantic expressions representing both “Are you <Nationality>?” and “Are you <EmotionalState>?”

4. Implementing the game

Having introduced the speech2sign framework, we now describe how we used it to implement the sign language game. There are two separate sets of issues: defining the rules that map menu-based pattern content to sign language, and packaging the functionality as a usable web game.
4.1. Menu-based generation of sign language

In principle, the framework makes it straightforward to define the mapping from menu-based pattern to sign language. We let each menu choice create a piece of text at the semantic representation level, concatenate these pieces of text together to form a string, and provide flat rules like the ones shown in the preceding section, one per pattern, to map the strings produced by each pattern into the correct sign tables. For example, the top-level rule in Figure 5 can in a natural way be associated with a pattern which contains a menu for nationalities, mapping the sentences produced by the menu into a sign table in the way shown.

In practice, there are two problems which make the exercise less than trivial. First, we wish to have a large number of patterns, and if we implement a flat mapping rule for each one it will be difficult to keep the rules coherent with each other. This is in particular the case with the non-manual elements, which as we have seen are important to the syntax of sign language. Considerations of this kind dispose us towards the idea of writing a more “linguistic” grammar, where rules share common structure. This leads us to the second problem: in our group (and, we suspect, in many others working with sign language), the relevant expertise is divided between several people. People with a mainstream language technology background are comfortable writing the structured grammar, if they receive some input from a sign language expert. The sign language expert must however design the course, which will require frequent changes to the patterns and menus. The question is how to set things up so that responsibility can be cleanly divided.

The solution we have adopted is to define two separate levels of semantic representation, which call “interlingua” and “pattern”. The complex set of structured rules, which is maintained by the language technology expert, maps the interlingua level to the sign table; at the level above, a large collection of flat rules, maintained by the sign language expert, maps the menu patterns to the interlingua. The first set of rules encode the linguistic structure, namely the relationship between abstract semantic representation and sign table. The second defines the pedagogical structure, namely the relationship between the patterns shown to the student and the abstract semantic structures they are associated with.

In §5, we describe the initial content we have created at these two levels.

4.2. User interface and gamification

The game is implemented using a standard client/server architecture. Most of the processing, in particular grammar-based generation and avatar-based synthesis, is carried out on remote servers. A thin HTML5/jQuery client wraps the main functionality. The user interface is illustrated in the screenshots from §2. The avatar pane includes controls which make it possible to change the number of frames per second (in effect playing the animation faster or slower), and rotating the character. This makes it possible to replay a piece of avatar signing, viewing it from several angles.

Gamification [16] has in recent years become extremely common in CALL, and forms an important part of many widely-used apps. (A prominent example is Duolingo). Menusigne is gamified using a simple score-and-leaderboard model. The student “practices” at the first level and “plays” at the second and third levels. At the “play” levels, each correct response scores points which are added to the running total for the current game. The number of points awarded per response depends on the difficulty of the lesson, which is determined by the total number of prompts available and the modality. Video prompts and lessons with more prompts give higher scores, the intention being to tempt the student out of their comfort zone. Every incorrect response loses a life. When the student has lost all their lives, the game is over. If the final score is high enough, it is posted on the leaderboard, an HTML page which is displayed inside the game. It is also externally accessible in case players should wish to link to it.

In later versions, we may make the gamification more elaborate. An obvious idea is to add a timer, with the time available to respond to a prompt decreasing as the student’s score increases.

4.3. Videos

The example videos used at the third level of the game (cf. Figure 3) are created using an efficient workflow integrated into the grammar compilation process. At compile-time, two files are created. The first of these contains a list of declarations, one for each string that can be generated from the defined menus; a declaration line says that a video example should be created for the string in question. The intention is that the course designer will edit this file, retaining the examples they wish to use, and incorporate it into the grammar. The second file lists the declared video examples which have not yet been recorded. This is uploaded to a web tool which prompts the signer to create the actual videos [17].

5. Initial content

We have used our framework to build an initial course for teaching LSF to French-speaking hearing students. The content is designed to give the student a vocabulary of about 125 signs and a knowledge of basic LSF syntax. We describe the linguistic and pedagogical content.

5.1. Linguistic content

The linguistic content is encoded as a speech2sign SCFG grammar and lexicon which map interlingua-level text strings into LSF sign tables. The lexicon contains about 300 signs. Each sign is associated with a HamNoSys entry taken from a lexicology developed at LIMSI under the ViSiCAST project [18]; less than half of the signs are used in the initial course. The grammar provides basic coverage of letters (used for finger-spelling), numbers, pronouns, nouns, adjectives, verbs (transitive, intransitive, subcategorising for VP and null copula), YN-questions, WH-questions, negation and adverbs.

The grammar has a structure loosely based on PSG [19]. There are three main groups of rules, for S, VP and NP; the interesting ones are in the S and VP groups. To summarise very briefly, the basic LSF word order is SOV. WH-questions, as is common for sign-languages, are formed by right-movement of the WH+ phrase. Thus for example “You drink beer” is the three sign sequence

TOI BIERE BOIRE

i.e. “You beer drink”, while “What do you drink?” is the three sign sequence

TOI BOIRE QUOI-q

i.e. “You drink what-q”, where the “-q” indicates simultaneous non-manual question movements of the head and shoulders.
Following standard practice, the central rule-schema in the grammar is of the form

\[
VP \rightarrow \text{COMPS} V
\]

(Since the clausal word-order is verb-final, the COMPS naturally precede the V). S-level rules define declarative sentences, YN-questions and WH-questions; because of the unusual word-order, the rule for WH-questions with a wh+ question element is schematically of the form

\[
S \rightarrow \text{NP} \text{VP/NP} \text{NP: [wh+]}
\]

Here, the “slash category” \( \text{VP/NP} \) represents a VP with an NP gap. VPs and Vs are marked for negation and possession of a question-element.

In the present implementation, features and slash categories are rather unsatisfactorily implemented using the rule template mechanism described at the end of §3; the next version will use a proper feature system. We return to this point in the last section.

5.2. Pedagogical content

We have constructed a basic course in LSF grammar. The current version is divided into ten lessons, containing a total of 58 patterns and 169 example videos, and uses a vocabulary of 125 signs. The student is introduced in turn to politeness phrases, letters and finger-spelling, numbers up to 99, adjectives, simple declarative sentences, negation, yes-no questions, requesting expressions, possessives and WH-questions. Table 1 summarises the current lesson content.

Table 1: Lessons in current sign language course. For each lesson, we list the grammatical constructions covered, the number of new signs introduced, the number of patterns and example videos provided, and one or two examples of patterns. French pattern content has been translated into English. Uppercase items in square brackets are menus.

<table>
<thead>
<tr>
<th>Name</th>
<th>Grammar</th>
<th>Signs</th>
<th>Patterns</th>
<th>Videos</th>
<th>Example patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greetings</td>
<td>Politeness</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>[GREETING]</td>
</tr>
<tr>
<td>Letters</td>
<td>Finger-spelling</td>
<td>28</td>
<td>3</td>
<td>8</td>
<td>[LETTER] [FINGER-SPELLED-WORD]</td>
</tr>
<tr>
<td>Numbers</td>
<td>Numbers</td>
<td>27</td>
<td>6</td>
<td>18</td>
<td>I am [TENS-NUMBER] [UNITS-NUMBER] years old</td>
</tr>
<tr>
<td>I am</td>
<td>Adjectives</td>
<td>15</td>
<td>4</td>
<td>16</td>
<td>I am [NATIONALITY]</td>
</tr>
<tr>
<td>About me</td>
<td>Declarative sentences</td>
<td>27</td>
<td>10</td>
<td>21</td>
<td>I like [LOVABLE-THING]</td>
</tr>
<tr>
<td>Negation</td>
<td>Negated sentences</td>
<td>1</td>
<td>7</td>
<td>27</td>
<td>I don’t like [LOVABLE-THING]</td>
</tr>
<tr>
<td>Questions</td>
<td>Yes-no questions</td>
<td>0</td>
<td>7</td>
<td>20</td>
<td>Do you like [LOVABLE-THING]</td>
</tr>
<tr>
<td>Food and drink</td>
<td>Requesting</td>
<td>16</td>
<td>11</td>
<td>18</td>
<td>I want [FOOD] please</td>
</tr>
<tr>
<td>My family</td>
<td>Third-person subjects</td>
<td>7</td>
<td>7</td>
<td>39</td>
<td>My [FAMILY-MEMBER] is a [PROFESSION]</td>
</tr>
<tr>
<td>What?</td>
<td>WH-questions</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>What do you [TRANSITIVE-VERB]</td>
</tr>
</tbody>
</table>

Table 1: Lessons in current sign language course. For each lesson, we list the grammatical constructions covered, the number of new signs introduced, the number of patterns and example videos provided, and one or two examples of patterns. French pattern content has been translated into English. Uppercase items in square brackets are menus.

6. Initial user reactions

The game is still in an early phase of testing. During this period, the main focus has been on using it to transfer some practical knowledge of sign language from Irene Strasly, the one member of the team who is a fluent signer, to other members, none of whom previously knew more than a handful of signs.

Using the content shown in Table 1, the first and second levels of the game already seem to work well. It is easy to progress through the course, watching signed animations at the first level and then practising responding to them at the second level; the gamified structure does a good job of encouraging the student to repeat examples enough times to start remembering them. The subjective experience is that these levels of the game are quite engaging and addictive.

For the third level, we are still experimenting with different strategies for creating the videos. Our initial plan was to record them in a “natural” way, i.e. without particular thought for the requirements of the game. This reflects the fact that signing styles vary a great deal, with little standardization, and it conforms to the common pedagogical principle of forcing students to confront this problem head-on from an early stage. Unfortunately, however, the subjective experience was that this strategy made the third level of the game too demanding. Even after watching a video several times, knowing what it meant, it was often hard to understand the signing. After discussing the issues, we created a revised version where the videos were recorded in a more careful manner; the signer used an exaggeratedly slow and explicit style and tried to render most utterances in the same way as the avatar. This does not stretch the student as much, but it seems more appropriate for the beginners who will be the main users of the course; the videos are now readily comprehensible, and the game is fun to play at the third level as well. We anticipate making further adjustments as we get more user feedback.

7. Summary and further directions

We have presented a simple web-deployed game for learning sign language. In contrast to previous work, the game focuses from the beginning on learning real sign language grammar, and in particular stresses the central importance of non-manual
movements. The avatar-based levels of the game clearly work well. The video-based level is still under development and can certainly be improved, but even the version we have now is quite usable.

The current system is only a sketchy prototype, so we are encouraged by the fact that it already seems to be a useful tool for learning sign language. Looking ahead, there are several easy ways to improve it. The first priority is to move to a grammar formalism more expressive than SCFG; although it has so far been possible to write the kinds of grammars we need in this framework, it is not good practice to simulate features using a template mechanism. We do not think it will be necessary to go as far as Marshall and Sáfár [20] and move to a full HPSG grammar, which would pose many technical problems; based on previous experience, a typed feature grammar, compiled down into SCFG form, should be sufficient. We can do this straightforwardly by integrating processing modules from our earlier Regulus project [21] into the compilation sequence of the current platform.

Once we have the extended grammar framework in place, the next step will be to create more content. This will be developed in the context of the newly established Swiss Competence Centre for Barrier-Free Communication3, where our group plays a leading role. The first practical goal will be to give Deaf sign-language teachers the possibility of assigning more effective home exercises to students. Another potential target group is parents doing “baby-sign” (teaching sign language to very young children: http://www.babysignlanguage.com/). This will require development of suitably adapted content—baby-sign typically uses only a simplified form of sign language grammar—but should not be difficult to do.

Finally, an interesting idea is to crowdsource the recording of videos, which is the most labor-intensive part of the content-development process. Since our videos are already created using a web tool closely integrated with the platform (cf. §4.3), this is challenging to implement: all that is required is to add a control which allows the user to invoke the recording tool at the practice level, storing the resulting video together with metadata linking it to the sentence selected from the menus. The less obvious question is whether Deaf people will want to get involved. But it is an easy experiment to try, and if it works the payoff is very substantial.

8. Acknowledgements

We would like to thank John Glauert for invaluable help with the JASigning avatar and Michael Filhol and Annelies Braffort for making the LIMSI HamNoSys lexicon available to us.

9. References


Lärka: an online platform where language learning meets natural language processing

Ildikó Pilán¹, David Alfter¹, Elena Volodina¹

¹Språkbanken, University of Gothenburg, Sweden
(ildiko.pilan, david.alfter, elena.volodina)@gu.se

1. General overview

We present Lärka¹, an Intelligent Computer Assisted Language Learning (iCALL) platform developed at Språkbanken. Lärka is an openly available web-based tool that builds on a variety of existing language resources such as corpora, lexical resources and language technology tools. This makes the platform flexible and a valuable source of additional learning material (e.g. via corpus-based exercises) and a support tool for both teachers and learners of Swedish. Lärka has recently received a new user-interface that is more suitable for different screen sizes. Moreover, the system has also been augmented with new functionalities. These recent additions aim at improving the usability and the usefulness of the platform for pedagogical purposes.

Thanks to Lärka’s service-oriented architecture, most functionalities are also available as web services which can be easily re-used by other applications.

2. Automatic exercise generation

One of the main functionalities of Lärka is the automatic generation of exercises based on real-life language examples from corpora. Exercise generation is aimed at two groups of learners: students of (Swedish) linguistics and learners of Swedish as a second language (L2). The currently available exercises have a multiple-choice format. Each exercise item consists of a sentence containing either a highlighted word or a gap, as well as a list of five answer alternatives out of which one is the correct answer and four are distractors, i.e. incorrect options. Students of linguistics can train parts of speech, syntactic relations and semantic roles. Language learners can choose between vocabulary exercises and inflectional exercises. A recent addition to our platform is a simple word-level exercise, WordGuess, that takes a step towards gamified learning. WordGuess re-implements the well-known Hangman game mechanism: users are presented with a number of hidden characters and their task is to guess characters contained in the word, which eventually helps them guessing the word itself. Every time the guessed character is not in the word, users receive penalty points. In our learning-oriented version of the game, users can choose to receive clues such as the translation of the word, its definition in Swedish or information about its morphological form. This game is a simple example of reusing information from lexical resources for gamified language learning activities.

3. Corpus example selection

In Lärka, the automatically generated exercises for language learners rely on HitEx (Hitta Exempel¹ ‘find examples’), a tool for selecting and ranking corpus examples. The main purpose of HitEx is to identify sentences from generic corpora which are suitable as exercise items for L2 learners. The suitability of the sentences is determined based on a number of parameters that reflect different linguistic characteristics of the sentences. Through a graphical user interface, it is also possible to perform a sentence search based on parameters customized by the user. The selection criteria include a wide variety of linguistic aspects such as the desired difficulty level based on the CEFR (Common European Framework of Reference for Languages), typically based on word co-occurrence measures, as well as the absence of anaphoric expressions and sensitive vocabulary (e.g. profanities), just to name a few. Besides its applicability to the language learning domain, HitEx can be also useful for lexicographers for finding dictionary examples that illustrate the meaning and usage of lexical items.

4. Text complexity evaluation

Another functionality, TextEval, offers an interface to automatically assess Swedish texts for their degree of complexity according to the CEFR. Texts can be either learner productions (e.g. essays) or texts written by experts as reading material for learners. The machine learning based automatic analysis returns an overall CEFR level for the text, as well as a list of linguistic indicators relevant for measuring text complexity, such as the average length of sentences and tokens, LIX score and nominal ratio. In addition, it is possible to add a color-enhanced highlighting for words per CEFR levels which provides users with a straightforward visual feedback about the lexical complexity of a text. Behind this feature are two word lists, one based on expert-produced texts to reflect receptive vocabulary and another based on learner-produced texts representing productive vocabulary. In the case of both lists, frequency distributions of lemmas have been mapped to a single CEFR level. For each CEFR level, a darker and a lighter shade of the same color represent productive and receptive vocabulary respectively at the given level.

5. Ongoing work and planned extensions

Besides the activities described in Section 2, the migration of the previous version of our spelling exercises and the addition of new exercise formats are currently under development. In the near future we plan to add a login functionality as well as an infrastructure to log user data. This would enable us to create a valuable resource for modeling learners (e.g. L1-specific errors, learners development over time) and to offer adaptive exercises. We also plan on offering a diagnostic test to assess learners’ proficiency levels based on different exercises. Further extensions under development include an annotation interface for learner corpora which facilitates the process of entering metadata about learner essays. We are also investigating the possibility of annotation learner errors through this interface.
New Features and Effectiveness of Suzuki-kun, the First and Only Prosodic Reading Tutor of Tokyo Japanese

Nobuaki MINEMATSU, Daisuke SAITO
The University of Tokyo Project OJAD

\{mine,dsk_saito\}@gavo.t.u-tokyo.ac.jp

Abstract

We have updated Suzuki-kun, which is the prosodic reading tutor of Japanese that we developed as one feature of OJAD (Online Japanese Accent Dictionary) \([1, 2]\). The original Suzuki-kun can visualize the prosodic and hierarchical structure of any given sentences to read them aloud in Tokyo Japanese and provide a synthetic speech sample based on the visualized prosody. The added new features are 1) generation of multiple speakers’ voices, 2) speaking rate control, and 3) generation of dialogue speech among these speakers. Further, experimental results of examining the effects of Suzuki-kun on the naturalness of learners’ spoken Japanese are described. It was found that visualized prosody is significantly more effective than auditory prosody.

As of late August in 2017, 111 tutorial workshops of OJAD will have been given in 29 countries. In demonstration, teachers’ reports will also be shown as well as Suzuki-kun’s new features.

Index Terms: OJAD, Suzuki-kun, Tokyo Japanese, visualized prosody, naturalness assessment

1. Suzuki-kun, the first and only teaching material for Tokyo Japanese

Native Japanese generally speak Tokyo Japanese (TJ) in public even when they are from local cities. TJ has a unique prosodic control of word accent and phrase intonation, but it is rarely taught to learners because TJ’s prosodic control is complicated and teaching time is limited. A lexical accent dictionary published for professional broadcasters \([3]\) explains TJ to be the dress code for speaking Japanese in public, but it is not taught well to learners. Suzuki-kun was developed to change this problematic situation and it is based on speech synthesis technologies. For a given sentence, a speech synthesizer does text analysis at first partly for prosody prediction. In Japanese, accentual phrase boundaries and intonational phrase boundaries are predicted. Further, within a predicted accentual phrase, the position of its accent nucleus is detected automatically. Suzuki-kun uses the predicted prosody to visualize the pitch contour and the positions of accent nuclei to read the given sentence correctly in TJ. The predicted (and symbolic) prosody is also used to generate synthetic speech, which follows the visualized prosody. The schematic procedure of Suzuki-kun is shown in Figure 1.

2. New features added to Suzuki-kun

The original Suzuki-kun was demonstrated at SLaTE2015 \([4]\), after which, we added several new features to Suzuki-kun. They are 1) generation of multiple speakers’ voices (2 males and 2 females), 2) speaking rate control, and 3) generation of dialogue speech among the four speakers. For the third feature, input text has to include some command lines as shown in Figure 2. In a command line, which starts with double slashes, speaker identity and speaking rate are specified so that they are used for Suzuki-kun to read aloud subsequent sentences.

3. Effectiveness of Suzuki-kun

Effectiveness of Suzuki-kun was verified by asking eighty Chinese learners of Japanese to practice with Suzuki-kun \([5]\). They were divided to three groups. Group A practiced reading a given text only with that text and then practiced reading with synthetic speech from Suzuki-kun (auditory prosody). Group B practiced only with text and then practiced with visual output from Suzuki-kun (visualized prosody). Group C practiced with both auditory and visualized prosody after practicing only with text. By comparing the prosodic naturalness observed in the first and the second practices, group C showed the largest improvement. Comparison between the two groups of A and B showed that visualized prosody was significantly more effective than auditory prosody. Detailed results are described in \([5]\).

4. For demonstration

Suzuki-kun is the first and only pedagogical material to teach and learn the dress code for speaking Japanese in public. After releasing it, we received an enormous number of accesses especially from overseas. OJAD has been translated into 14 non-Japanese languages for beginners. As of late August in 2017, we will have given 111 tutorial workshops of OJAD in 29 countries. The latest one will be at University of Göteborg, Sweden before INTERESPEECH. In demonstration, not only Suzuki-kun’s new features but also teachers’ reports will be explained.

5. References

\[1\] I. Nakamura \textit{et al.}, Proc. INTERSPEECH, 2554–2558, 2013

\[2\] \texttt{https://www.gavo.t.u-tokyo.ac.jp/ojad}

\[3\] NHK accent dictionary, published by NHK publisher, 2016

\[4\] N. Minematsu \textit{et al.}, Proc. SLaTE, 189, 2015

Development and Maintenance of Practical and In-service Systems for Recording Shadowing Utterances and Their Assessment

Junwei YUE†, Daisuke SAITO†, Nobuaki MINEMATSU†, Yutaka YAMAUCHI‡

†The University of Tokyo, ‡Tokyo International University
{jwyue,dsk.saito,mine}@gavo.t.u-tokyo.ac.jp, yyama@tiu.ac.jp

Abstract

We demonstrate two systems, network-based and standalone, for collecting shadowing utterances and their automatic assessment. Since June 2016, these systems have been used in real English classes at several universities in Japan. The following features are highlighted: 1) To avoid pop noises from a speaker and reduce babble noises from other surrounding students, an ear-hook microphone is used with a USB audio device, 2) A network-based system and a standalone system were developed separately because network traffic not rarely causes technical errors when recording, 3) For beginners, easy-to-understand illustrations are prepared for them to get accustomed rapidly to shadowing practices, 4) DNN-based GOP calculation is run and its score is fed back to learners, and 5) To motivate them, the GOP score distribution over the learners are also fed back for them to compare their own scores with others'. In demonstration, each feature will be exhibited and explained in detail.

Index Terms: Shadowing, recording, assessment, ear-hook microphone, DNN, GOP, feedback instructions

1. Background

For good and valid assessment, an adequate recording environment and an encouraging learning environment are essential. We developed two systems, network-based and standalone, for online and offline learners with several features enabling stable recording and motivating learners to practice shadowing repeatedly. These systems are now used in universities in Japan.

2. Network-based system

It allows learners to record shadowing speeches online by themselves. This system is developed using PHP and Javascript and recording is possible in class and at home. A screenshot of this system is shown in Fig. 1. In actual use of this system, a web page of instructions for recording is prepared so that learners can become skillful in recording. A unidirectional earhook microphone is used to suppress babble noise. To hear model utterances, canalphones are recommended to prevent the model utterances from being leaked and recorded (Fig. 2). Each recorded sentence is converted into wave format and sent to the laboratory server automatically so there is no need to collect them again, which is convenient for teachers and engineers. As for automatic assessment, it will be described in Section 4.

3. Standalone system

Network connection is not always available. For such learners, another standalone system was developed, which has more customizability than the network-based system. In our network-based system, customization often requires operations of a system manager, but in the standalone system, a user, i.e., teacher, can customize the system in an easier way. When our network-based system was tested in some universities, due to network traffic, a recording process was sometimes stopped unexpectedly, which discouraged learners. For secure recording, the standalone system is recommended.

4. Online GOP calculation

To motivate learners, the network-based system is capable of calculating the DNN-based GOP score for each shadowing utterance (Fig. 1). The GOP score is calculated using DNN-based acoustic models, which are trained using WSJ corpus [1]. After recording any sentence, learners can know his/her score by a single click. To motivate learners in a class, a tentative web was built to let them know the GOP score distribution of all the learners. The location of each learner was indicated in the distribution (Fig. 3). Every learner can know how well he/she performs shadowing compared to his/her classmates.

5. References

A pipeline for automatic assessment of foreign language pronunciation

Aku Rouhe1, Reima Karhila1, Heini Kallio2, Mikko Kurimo1

1Aalto University, Finland
2University of Helsinki, Finland

firstname.lastname@aalto.fi

Abstract

We illustrate our prototype pipeline for automatic analysis and assessment of foreign language pronunciation. It uses automatic speech recognition as a preprocessing step for phonetic analysis and predicts the grade that human experts would give for the utterance. The work is applied in two educational research projects: DigiTala for L2 Swedish for Finnish-speaking high school students, and SIAK "Say it again, kid!" for L2 English for Finnish-speaking 6-9 year old children.

Index Terms: speech recognition, pronunciation grading, language learning, phonetics, speech analysis

1. Introduction

We demonstrate a pipeline for automatic assessment of foreign language pronunciation. First we describe our work in recognising, validating and segmenting read and spoken prompts into phonemes. Then we present our work in automatic phonetic analysis and its application to predict human experts’ assessments. The pipeline is applied in two demonstrations that have been developed in two projects for two L2 languages.

In the DigiTala project we are currently developing a high stakes spoken language testing process for L2 Swedish of native Finnish students in matriculation examinations. Since expert human reviewing time is scarce and expensive, assistive technology is needed. After our initial work to establish the use of computerised testing environment [1], we have used the pilot systems to collect data and experiences in real usage scenarios.

The SIAK “Say it again, kid!” project approaches foreign language acquisition in early primary education from an interdisciplinary point of view. In the project we have created a L2 pronunciation learning game and tested it in primary schools.

2. Speech recognition for prompted content

The speech processing pipeline at the server starts with an automatic speech recognition. The rest of the system depends on the recognizer to provide an accurate phone-level segmentation, which requires a successful decoding of the input speech. For L2 learners and particularly children, the decoding is a hard task. Thus, our initial work focuses on read or repeated words, where the content of the recordings can be hypothesised. We demonstrate a system which follows reading, validates that the pronunciation matches our expectation and segments the result into phonemes.

We have built acoustic models for L2 Swedish spoken by Finnish-speaking high school students and L2 English spoken by young Finnish-speaking children. Many of these speakers’ data match poorly to the native target language speech corpora, so realistic in-domain data is immensely valuable. Furthermore, most speakers often do not speak fluently, so the conventional language models do not match well, either. We have implemented a heuristic model which tolerates miscues in reading, resembling to the work done in Project LISTEN[2]. An additional constraint is that in the L2 learning game all speech processing operates in realtime.

3. Automatic pronunciation rating

The pronunciation skill level of the speaker is estimated from phoneme recognition results. Individual phoneme segments are classified with bilingual LSTM-RNNs [3] that are trained with native speech data from both the target language and the native language of the users. Scores are computed as a weighted linear sum or other simple regression of phoneme recognition confusion matrices. The regression system is trained in a supervised manner with in-domain utterances of varying skill level annotated by experts.

The English children’s system works on very sparse data, returning a score from 1 to 5 for each utterance. The demonstrated Finnish Swedish takes in 8 utterances consisting of around 200 phonemes, and it gives a rough guess of skill level on a 10-step scale.

4. Conclusions

We have demonstrated an automatic speech processing pipeline which provides assessments of spoken language skill. We have educational research projects using the pipeline in both L2 English and L2 Swedish in schools. In the projects we are gathering large in-domain datasets, which can be used to further improve the speech recognition performance and test and analyse the L2 pronunciation learning. We will also expand the systems to other, more open-ended speaking tasks.

5. Acknowledgements

This work has been supported by the Matriculation Examination Board of Finland, Svenska Folkskolans Vänner Foundation, Tekes Challenge Finland programme and the Academy of Finland’s grants 251170 and 274075. We acknowledge the computational resources provided by the Aalto Science-IT project.

6. References


## Author Index

<table>
<thead>
<tr>
<th>Author</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afshan, Amber</td>
<td>161</td>
</tr>
<tr>
<td>Ahmad, Muneeb</td>
<td>18</td>
</tr>
<tr>
<td>Alfter, David</td>
<td>187</td>
</tr>
<tr>
<td>Alwan, Abeer</td>
<td>161</td>
</tr>
<tr>
<td>Artuso, Stefano</td>
<td>115</td>
</tr>
<tr>
<td>Axtmann, Nico</td>
<td>85</td>
</tr>
<tr>
<td>Bailly, Gérard</td>
<td>127</td>
</tr>
<tr>
<td>Balogh-Ghosh, Jennifer</td>
<td>172</td>
</tr>
<tr>
<td>Baur, Claudia</td>
<td>71</td>
</tr>
<tr>
<td>Berkling, Kay</td>
<td>85</td>
</tr>
<tr>
<td>Bernstein, Jared</td>
<td>172</td>
</tr>
<tr>
<td>Bianco, Maryse</td>
<td>127</td>
</tr>
<tr>
<td>Bosse, Marie-Line</td>
<td>127</td>
</tr>
<tr>
<td>Bouillon, Pierrette</td>
<td>181</td>
</tr>
<tr>
<td>Caines, Andrew</td>
<td>79</td>
</tr>
<tr>
<td>Cámará-Arenas, Enrique</td>
<td>25</td>
</tr>
<tr>
<td>Campbell, Nick</td>
<td>138</td>
</tr>
<tr>
<td>Cardeño-Payo, Valentín</td>
<td>25</td>
</tr>
<tr>
<td>Carson-Berndsen, Julie</td>
<td>1</td>
</tr>
<tr>
<td>Chasaide, Ailbhe Ni</td>
<td>47</td>
</tr>
<tr>
<td>Cheng, Jian</td>
<td>172</td>
</tr>
<tr>
<td>Chiarain, Neasa Ni</td>
<td>138</td>
</tr>
<tr>
<td>Chua, Cathy</td>
<td>71</td>
</tr>
<tr>
<td>Chung, Minhwa</td>
<td>53</td>
</tr>
<tr>
<td>Cowan, Benjamin</td>
<td>138</td>
</tr>
<tr>
<td>Cristoforetti, Luca</td>
<td>115</td>
</tr>
<tr>
<td>Dantsuji, Masatake</td>
<td>42</td>
</tr>
<tr>
<td>Diallo, Alpha</td>
<td>138</td>
</tr>
<tr>
<td>Duan, Richeng</td>
<td>42</td>
</tr>
<tr>
<td>Engwall, Olov</td>
<td>7</td>
</tr>
<tr>
<td>Escudero, David</td>
<td>25</td>
</tr>
<tr>
<td>Evaniini, Keelan</td>
<td>97</td>
</tr>
<tr>
<td>Falavigna, Daniele</td>
<td>115</td>
</tr>
<tr>
<td>Gales, Mark</td>
<td>59</td>
</tr>
<tr>
<td>Gerlach, Johanna</td>
<td>71</td>
</tr>
<tr>
<td>Gilmartin, Emer</td>
<td>138</td>
</tr>
<tr>
<td>Godde, Erika</td>
<td>127</td>
</tr>
<tr>
<td>González-Ferreras, César</td>
<td>25</td>
</tr>
<tr>
<td>Grettler, Roberto</td>
<td>115</td>
</tr>
<tr>
<td>Hommel, Marlisa</td>
<td>177</td>
</tr>
<tr>
<td>Huang, Yuyun</td>
<td>138</td>
</tr>
<tr>
<td>Ito, Kayoko</td>
<td>30</td>
</tr>
<tr>
<td>Jančovič, Peter</td>
<td>91</td>
</tr>
<tr>
<td>Jeon, Hyung-Bae</td>
<td>103</td>
</tr>
<tr>
<td>Kaewtip, Kantapon</td>
<td>161</td>
</tr>
<tr>
<td>Kallio, Heini</td>
<td>190</td>
</tr>
<tr>
<td>Kang, Byung Ok</td>
<td>103</td>
</tr>
<tr>
<td>Karhila, Reima</td>
<td>190</td>
</tr>
<tr>
<td>Kato, Tsuneo</td>
<td>13</td>
</tr>
<tr>
<td>Kawahara, Tatsuya</td>
<td>42</td>
</tr>
<tr>
<td>Khalifa, Albara</td>
<td>13</td>
</tr>
<tr>
<td>Kim, Jaebok</td>
<td>138</td>
</tr>
<tr>
<td>Knill, Kate</td>
<td>59</td>
</tr>
<tr>
<td>Kurimo, Mikko</td>
<td>190</td>
</tr>
<tr>
<td>Kyriakopoulos, Konstantinos</td>
<td>59</td>
</tr>
<tr>
<td>Laarmann-Quante, Ronja</td>
<td>36</td>
</tr>
<tr>
<td>Lee, Yun-Keun</td>
<td>103</td>
</tr>
<tr>
<td>Lee, Yun-Kyung</td>
<td>103</td>
</tr>
<tr>
<td>Li, Ruobing</td>
<td>167</td>
</tr>
<tr>
<td>Lin, Hui</td>
<td>167</td>
</tr>
<tr>
<td>Litman, Diane</td>
<td>109</td>
</tr>
<tr>
<td>Lopes, José</td>
<td>7</td>
</tr>
<tr>
<td>Lulich, Steven M.</td>
<td>161</td>
</tr>
<tr>
<td>Magooda, Ahmed</td>
<td>109</td>
</tr>
<tr>
<td>Malinín, Andrey</td>
<td>144</td>
</tr>
<tr>
<td>Mana, Nadia</td>
<td>115</td>
</tr>
<tr>
<td>Mehret, Carolina</td>
<td>85</td>
</tr>
<tr>
<td>Meshgi, Kourosh</td>
<td>156</td>
</tr>
<tr>
<td>Minematsu, Nobuaki</td>
<td>30</td>
</tr>
<tr>
<td>Mirzaei, Maryam Sadat</td>
<td>156</td>
</tr>
<tr>
<td>Mubin, Omar</td>
<td>18</td>
</tr>
<tr>
<td>Mulholland, Matthew</td>
<td>97</td>
</tr>
<tr>
<td>Nanjo, Hiroaki</td>
<td>42</td>
</tr>
<tr>
<td>Oh, Yoo Rhee</td>
<td>103</td>
</tr>
<tr>
<td>Orlando, Joanne</td>
<td>18</td>
</tr>
<tr>
<td>Ozgun, Kaan Ege</td>
<td>161</td>
</tr>
<tr>
<td>Park, Jeon-Gue</td>
<td>103</td>
</tr>
<tr>
<td>Pilan, Ildiko</td>
<td>187</td>
</tr>
<tr>
<td>Qian, Mengjie</td>
<td>91</td>
</tr>
<tr>
<td>Qian, Yao</td>
<td>97</td>
</tr>
<tr>
<td>Ragni, Anton</td>
<td>144</td>
</tr>
<tr>
<td>Rao, Preeti</td>
<td>121</td>
</tr>
<tr>
<td>Rayner, Manny</td>
<td>71</td>
</tr>
<tr>
<td>Rosenfeld, Elizabeth</td>
<td>172</td>
</tr>
<tr>
<td>Rouhe, Aku</td>
<td>190</td>
</tr>
</tbody>
</table>
Russell, Martin 71, 91
Ryu, Hyuksu 65
Sabu, Kamini 121
Saito, Daisuke 188, 189
Schiavo, Gianluca 115
Shahid, Suleman 18
Skantze, Gabriel 7
Sloan, John 1
Song, Hwa Jeon 103
Strasly, Irene 181
Strik, Helmer 71
Su, Ketong 138
Swarup, Prakhar 121
Tejedor-García, Cristian 25
Tsourakis, Nikolaos 181
Tsuprun, Eugene 97
Tulsiani, Hitesh 121
Vilain, Coriandre 127
Volodina, Elena 187
Wang, Chuan 167
Wang, Xinhao 132, 150
Wang, Yu 144
Wei, Xizi 71, 91
Yamamoto, Seiichi 13
Yamauchi, Yutaka 30, 189
Yang, Seung Hee 53
Yeung, Gary 161
Yue, Junwei 30, 189
Zechner, Klaus 132
Zhao, Yong 138