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

Cristian Tejedor-García[1] , David Escudero-Mancebo[1],
César González-Ferreras[1], Enrique Cámara-Arenas[2], Valentín Cardeñoso-Payo[1]

[1]Department of Computer Science [2]Department 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 redundant: the same instructions are usually expressed once in simple technical terms, and then in a friendlier, more impressionistic 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 vocalic 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 /ɑː, æ, ʌ/, the second incorporated /e/, and the third introduced /ɪ, iː/. Sessions II and III involved a partial revision of previous material in the form of new contrasts.

Participants carried out several **task-types** in five different **modes** (see second screenshot of figure 1). The score accumulated by each participant in each mode is also shown (0% - 100%). Access to the different modes was made successive and guided: users could only move to the next mode after reaching a score of 60% or more in the previous one. Each mode was structured in a fixed number of **task-tokens** (see table 1). A task-token was completed when the challenge presented on screen was overcome. All challenges implied interaction between user and application, and they could end in either success or failure.

Theory mode (figure 1, third screenshot) presents users with a 5-10 minute video that provides articulatory and perceptive information about the vowels of each lesson. Our tool makes use of IPA phonetic symbols, under the assumption that – considering the deep nature of the English writing system – any particular aural memory will benefit in terms of recollection from attachment to a specific non-ambiguous visual form. All videos were made by an experienced university teacher of English phonetics who has taught this subject for more than fifteen years. The instructions contained in the videos were given in the L1 of the users, Spanish. Participants could watch the video as many times as they considered necessary, and at least once in order to access the next mode.

Table 1: *Number of task-tokens categorized by mode. THE, EXP, DIS, PRO and MIX are Theory, Exposure, Discrimination, Pronunciation and Mixed modes, respectively.*

Mode	***THE***	***EXP***	***DIS***	***PRO***	***MIX***
# Task-tokens	1	3	10	10	9

Exposure, the second mode (figure 1, fourth screenshot) constitutes, together with the Theory mode, a fundamental part of feedback. Three task-tokens of minimal pairs were presented. In this task-type, both the orthographical and phonetic forms were given. Output model pronunciations were synthetically generated by Google's offline Text-To-Speech tool for Android. This tool synthesizes any text written in any of the languages supported by the device, with adjustable values for rate and pitch. Users listen to each word of the pair five times alternatively, and the word is produced more slowly each time. This mode offers users a first-hand unmediated aural experience of each contrast in order to ease their assimilation. The suggested challenge in this mode is for users to record their own versions of the words and to compare them to the synthesized outputs. Participants were allowed to remain in this mode for as long as they wished, listening and recording at will.

In the **Discrimination** mode (figure 1, fifth screenshot), participants are presented with a written and transcribed minimal pair, while only one of its constituents is synthetically generated; the challenge is to identify which of the words is being pronounced by the TTS. In this task-type, the members of the pair to be synthesized were randomly selected. A total of ten discrimination tokens were presented in each lesson.

The **Pronunciation** mode (figure 1, sixth 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

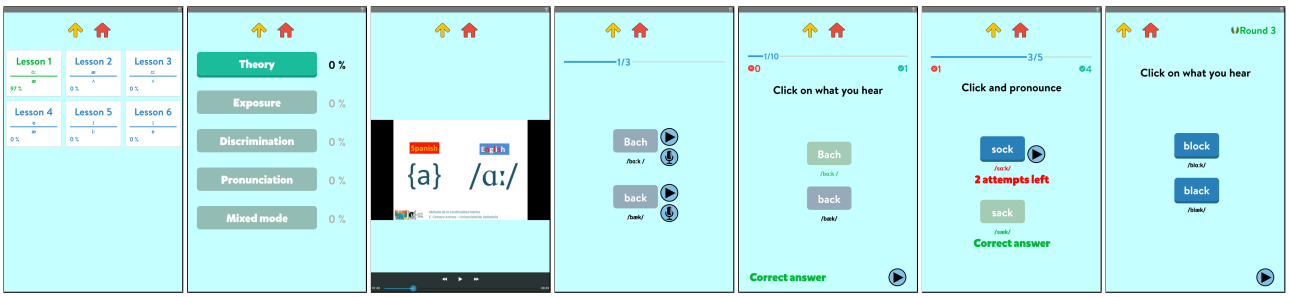

Figure 1: *Screenshots of the sequence steps in a lesson in the application. The second one is the modes menu. The last five screenshots correspond to the five modes: Theory, Exposure, Discrimination, Pronunciation and Mixed modes, respectively.*

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.

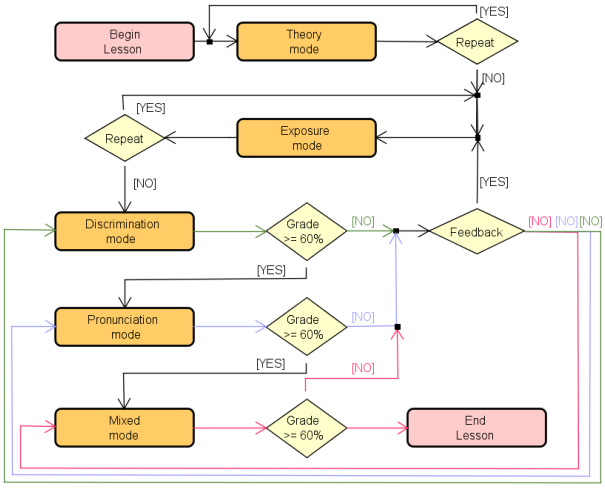

Figure 2: *Flow chart of the training modes in a lesson.*

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)[1]. 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.

[1]Official website (last visited May, 3rd 2017) `http://www.coe.int/t/dg4/linguistic/cadre1_en.asp`

Table 2: *Participant demographic profiles.*

Total	Female	Male	Age: 15-25	Age: 26-45
10	2	8	5	5

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.*

Mode	*THE*	*EXP*	DIS	*PRO*	*MIX*
Time (min)	31.32	16.93	5.48	41.47	19.03
# In.listenings	NA	3570	695	NA	268
# Ex.listenings	NA	1469	299	1479	632
# Productions	NA	NA	NA	4415	1741
# Recordings	NA	902	NA	NA	NA

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].*

	Proposed modes	**Completed modes first-round**	**Mode repetitions**	**Mode repetitions with feedback**	**Mode repetitions without feedback**
DIS	60	51	10	6 [6,0]	4 [3,1]
PRO	60	35	61	40 [23,17]	21 [2,19]
MIX	60	43	34	12 [12,0]	22 [5,17]

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

[1] N. Wiener, *Cybernetics: Control and communication in the animal and the machine*. Wiley New York, 1948.

[2] J. Hattie and H. Timperley, "The power of feedback," *Review of educational research*, vol. 77, no. 1, pp. 81–112, 2007.

[3] R. L. Oxford, "Language learning styles and strategies," *Teaching English as a Second or Foreign Language. M. Celce-Murcia (Ed.)*, p. 359–366, 2001.

[4] C. Cucchiarini, A. Neri, and H. Strik, "Oral proficiency training in dutch l2: The contribution of asr-based corrective feedback," *Speech Communication*, vol. 51, no. 10, pp. 853–863, 2009.

[5] H.-C. Liao, J.-C. Chen, S.-C. Chang, Y.-H. Guan, and C.-H. Lee, "Decision tree based tone modeling with corrective feedbacks for automatic mandarin tone assessment." *Interspeech*, 2010.

[6] S. Bodnar, C. Cucchiarini, B. Penning de Vries, H. Strik, and R. van Hout, "Learner affect in computerised l2 oral grammar practice with corrective feedback," *Computer Assisted Language Learning*, pp. 1–24, 2017.

[7] K. Saito, "Effects of instruction on l2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies," *TESOL Quarterly*, vol. 46, no. 4, pp. 842–854, 2012.

[8] Y. Iribe, S. Manosavanh, K. Katsurada, R. Hayashi, C. Zhu, and T. Nitta, "Introducing articulatory anchor-point to ann training for corrective learning of pronunciation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3716–3720.

[9] K.-H. Wong, W.-K. Lo, and H. Meng, "Allophonic variations in visual speech synthesis for corrective feedback in capt," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5708–5711.

[10] J. Zhao, H. Yuan, W.-K. Leung, H. Meng, J. Liu, and S. Xia, "Audiovisual synthesis of exaggerated speech for corrective feedback in computer-assisted pronunciation training," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8218–8222.

[11] E. Cámara-Arenas, *Native Cardinality: on teaching American English vowels to Spanish students*, ser. Historia y sociedad. Ediciones Universidad de Valladolid, 2013.

[12] E. Cámara-Arenas, "The NCM and the Reprogramming of Latent Phonological Systems: A Bilingual Approach to the Teaching of English Sounds to Spanish Students," *Procedia - Social and Behavioral Sciences*, vol. 116, pp. 3044 – 3048, 2014.

[13] A. Baker, S. Goldstein, and P. Dolgin, *Pronunciation Pairs: An Introductory Course for Students of English. Student's Book*. Cambridge University Press, 1990.

[14] A. Baker and L. Marshall, *Ship or sheep?* Cambridge University Press, 1981.

[15] J. D. O'connor and C. Fletcher, *Sounds English-A pronunciation practice book*. Longman Group UK Limited, 1999.

[16] D. Escudero-Mancebo, E. Cámara-Arenas, C. Tejedor-García, C. González-Ferreras, and V. Cardeñoso-Payo, "Implementation and test of a serious game based on minimal pairs for pronunciation training," *SLaTE*, pp. 125–130, 2015.

[17] C. Tejedor-García, V. Cardeñoso-Payo, E. Cámara-Arenas, C. González-Ferreras, and D. Escudero-Mancebo, "Measuring pronunciation improvement in users of CAPT tool TipTopTalk!" *Interspeech*, pp. 1178–1179, September 2016.

[18] C. Tejedor-García, D. Escudero-Mancebo, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "Improving L2 Production with a Gamified Computer-Assisted Pronunciation Training Tool, TipTopTalk!" *IberSPEECH 2016: IX Jornadas en Tecnologías del Habla and the V Iberian SLTech Workshop events*, pp. 177–186, 2016.

[19] C. Tejedor-García, D. Escudero-Mancebo, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "TipTopTalk! mobile application for speech training using minimal pairs and gamification," *IberSPEECH 2016: IX Jornadas en Tecnologías del Habla and the V Iberian SLTech Workshop events*, pp. 425–432, 2016.

[20] M. Celce-Murcia and J. M. Goodwin, *Teaching Pronunciation*, 4th ed. London. Thomson Learning, 2014.

[21] Z. Handley, "Is text-to-speech synthesis ready for use in computer-assisted language learning?" *Speech Communication*, vol. 51, no. 10, pp. 906–919, 2009.

[22] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[23] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7962–7966.

[24] P. Kerswill and A. Williams, "New towns and koineization: linguistic and social correlates," *Linguistics*, vol. 43, no. 5, pp. 1023–1048, 2005.

[25] S. Loewen, "Focus on form," *Handbook of research in second language teaching and learning*, vol. 2, pp. 576–592, 2011.

[26] Y. Sheen and R. Ellis, "Corrective feedback in language teaching," *Handbook of research in second language teaching and learning*, vol. 2, pp. 593–610, 2011.

[27] P. Roach, *Phonetics*, ser. Oxford English. OUP Oxford, 2001.

[28] E. M. Kissling, "Teaching pronunciation: Is explicit phonetics instruction beneficial for fl learners?" *The modern language journal*, vol. 97, no. 3, pp. 720–744, 2013.