Speech Communication 115 (2019) 51–66

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Golden speaker builder – An interactive tool for pronunciation training

Shaojin Ding[a,1], Christopher Liberatore[a,1], Sinem Sonsaat[b], Ivana Lučić[b], Alif Silpachai[b], Guanlong Zhao[a], Evgeny Chukharev-Hudilainen[b], John Levis[b], Ricardo Gutierrez-Osuna[a,*]

[a] Department of Computer Science and Engineering, Texas A&M University, United States
[b] Department of English, Iowa State University, United States

A B S T R A C T

The type of voice model used in Computer Assisted Pronunciation Instruction is a crucial factor in the quality of practice and the amount of uptake by language learners. As an example, prior research indicates that second-language learners are more likely to succeed when they imitate a speaker with a voice similar to their own, a so-called "golden speaker". This manuscript presents Golden Speaker Builder (GSB), a tool that allows learners to generate a personalized "golden-speaker" voice: one that mirrors their own voice but with a native accent. We describe the overall system design, including the web application with its user interface, and the underlying speech analysis/synthesis algorithms. Next, we present results from a series of listening tests, which show that GSB is capable of synthesizing such golden-speaker voices. Finally, we present results from a user study in a language-instruction setting, which show that practising with GSB leads to improved fluency and comprehensibility. We suggest reasons for why learners improved as they did and recommendations for the next iteration of the training.

1. Introduction

Pronunciation teaching often includes practice with a teacher, who can guide learners individually and provide feedback in the correct manner and amount when necessary (Hincks, 2003). Yet this is often time-consuming and expensive when the educational institutions' benefits are taken into consideration. Additionally, this does not match up well with the way that teachers usually approach pronunciation teaching. Research shows that most teachers approach pronunciation teaching in an ad-hoc manner, that is, they address pronunciation issues mostly in presence of a salient error or an error causing a communication problem. This is mostly either because teachers do not have sufficient training (Burgess and Spencer, 2000) or self-confidence (Couper, 2017; MacDonald, 2002) in pronunciation teaching. Another common belief among teachers is that pronunciation improvement will take care of itself with sufficient input and it does not require teaching in the way that other language skills do. This is a belief that was motivated by the principles of communicative language teaching which emphasized fluency over accuracy (Levis and Sonsaat, 2017).

However, providing instruction and feedback on immediate production in pronunciation teaching is an essential pedagogical requirement for learners' improvement, even though it can demand extensive instructional interventions (Warren et al., 2009). One solution to the lack of time and training of teachers is computer-assisted pronunciation training (CAPT) systems, which have been utilized to support learners to study autonomously and help teachers provide learners with individual feedback without using large amounts of time in class (Egan and LaRocca, 2000; Eskenazi, 1999; Rypa and Price, 1999). CAPT may also be motivating for many learners, both because of their interest in technology and because of learning preferences that make working with a computer program more comfortable than interacting with a real person. CAPT gives learners the chance to work on their pronunciation in a stress-free environment, at their own time and pace. For instance, pronunciation is a skill that may require extensive listening and repetition. Some learners may feel uncomfortable about asking for a repetition in class more than once, but with a CAPT program it is easier to make use of extensive repetition (Hardison, 2004). All said, CAPT offers great promise for individualized pronunciation instruction, more consistent practice, and greater comfort in learning (Levis, 2007).

With advancements in speech technologies such as automatic speech recognition (ASR) and speech synthesis, CAPT can also provide practice opportunities that a face-to-face class cannot. For example, the use of speech visualizations that adapt to each person's speech (Bliss et al., 2018), the use of multiple voices in perceptual training (Barriuso and Hayes-Harb, 2018; Thomson, 2011; Thomson, 2012), or the use of personalized voices (Probst et al., 2002) all provide learning opportunities that classroom pronunciation training cannot. The latter idea (i.e., personalized voices) has resurfaced several times in the CAPT literature. It was first proposed nearly thirty years ago by Nagano and Ozawa (Nagano and Ozawa, 1990). In their pioneering study, Japanese learners were asked to practice with a model of their own voice that had been modified to match the prosody of a reference English speaker. Post-training utterances from these learners were rated as more native-like than those for a second group of learners who instead had practiced with the reference English voice. More than a decade later, Probst et al. (2002) published a study in Speech Communication where L2 learners

* Corresponding author.
E-mail addresses: shjd@tamu.edu (S. Ding), rgutier@cse.tamu.edu (R. Gutierrez-Osuna).
[1] These authors contributed in equal parts to the development of Golden Speaker Builder.

https://doi.org/10.1016/j.specom.2019.10.005
Received 16 July 2019; Received in revised form 8 October 2019; Accepted 28 October 2019
Available online 1 November 2019

Table 1
Feedback types in the oral classroom and CALL environment (Heift, 2004) (p. 418).

Feedback type	Oral classroom	CALL
Explicit correction	You mean…	Correct answer
Recast	Teacher reformulation	Correct answer
Clarification	What do you mean?	Try again!
Meta-linguistic feedback	Explanation of error type	Explanation of error type
Elicitation	Ellipsis	Highlighting
Repetition	Intonation	Highlighting

were asked to practice with a native speaker voice that had different characteristics. Participants who imitated a well-matched voice (i.e., one with characteristics similar to their own voice) improved more than those who imitated a poor match. This result led the authors to suggest that each learner has an ideal speaker voice to imitate, a so-called "Golden Speaker." Nearly ten years later, and in an article also published in Speech Communication (Felps et al., 2009), we proposed that each learner's "Golden Speaker" should be their own voice, resynthesized to have a native accent. Most notably, in that study we presented an accent-conversion technique that was able to correct not only the learner's prosody (as Nagano and Ozawa had done) but also their segmental errors (i.e., phoneme substitutions, additions and deletions). Missing from our study, however, was a validation of the technique on pronunciation-training experiments. This is a clear next step. A decade since the first paper has shown that refining the accent-conversion technique for successful deployment in pronunciation training was more challenging that expected. The improvenment we have seen in accent-conversion quality makes us optimistic for further successful deployment of the Golden Speaker algorithms.

The manuscript describes a web application (Golden Speaker Builder; GSB) and the underlying speech analysis/synthesis algorithms that allow L2 learners to generate their own personalized voices. In a first step, we conduct a series of listening tests to determine the extent to which the synthesized voices mirror the learner's own voice with an American English accent. Then, we validate GSB in a language-instruction setting with a population of Korean L2 learners of English. The study was guided by two research questions:

- **RQ1**: What is the effect of using the GSB on learners' improvement of their comprehensibility and fluency?
- **RQ2**: What features of the GSB did learners find useful, and what did they find in need of improvement?

2. Review of the literature

2.1. Feedback in second language pronunciation acquisition

Feedback is an essential factor in L2 learning of all kinds and includes a range of implicit and explicit approaches. Feedback refers to "information learners receive in response to their communicative efforts" (Mackey and Abbuhl, 2005) (p. 210). Researchers emphasize the role of feedback in Second Language Acquisition (SLA) by arguing that positive evidence (i.e., input) is not sufficient unless learners are also provided with negative feedback (Gass et al., 1998). Similarly, Swain and Lapkin (1995) report that input alone is not sufficient for SLA; output should also accompany input because output fosters deeper engagement with language than input alone. Swain (2000) emphasizes the importance of output by stating "output may stimulate learners to move from the semantic, open-ended, strategic processing prevalent in comprehension to the complete grammatical processing needed for accurate production" (p. 99).

As noted in Heift (2004), because of the differences of medium, the computer-assisted language learning (CALL) environment and oral classroom settings are different from each other in terms of the way they provide feedback. For instance, a teacher saying, "what do you mean?" as a clarification request is replaced by a command sentence on the computer screen, "try again!"

Types of feedback presented in Table 1 can also be provided to second language learners in CAPT programs employing ASR and speech synthesis technologies. For instance, ASR-based programs may provide a pronunciation score based on detected pronunciation errors in an utterance (Mak et al., 2003) which can be classified as similar to clarification in an oral classroom. These scores may lead learners to repeat their performance until they get a satisfactory score. Some programs attempt to identify specific mispronounced parts of an utterance indicating where there is a problem (Kanters et al., 2009) while others highlight the individual sounds that are mispronounced and provide metalinguistic explanations about how to produce given sounds correctly.

Another type of feedback which may lead to improvement in pronunciation is a "recast"—a correct restatement of the mispronounced utterance. In relationship to oral feedback in pronunciation teaching, two studies by Lyster (1998, 2001) are noteworthy. Lyster studied French immersion classrooms to analyze feedback strategies employed by teachers along with learners' uptake—that is, their immediate repair, based on feedback they received. Lyster found teachers preferred using recasts for grammatical and phonological errors whereas they made use of elicitation for lexical errors. Lyster also reported that the use of recasts led to the highest rate of uptake for phonological errors. Based on these findings, he suggested that reformulation of the erroneous utterance might be sufficient to correct a pronunciation error successfully. Similarly, Nicholas et al. (2001) supported recasts being classified as an implicit type of feedback since they make learners aware of the new items to be learned without impeding the flow of conversation.

Recasts in CALL can be interpreted as the imitation of a correct utterance, mostly pronounced by a native speaker. Imitation exercises have been found to be helpful for pronunciation improvement as previous research found that this type of learning improves learners' perception (Eskenazi, 2009). However, questions about who to imitate have led the way to new research in pronunciation. Probst et al. (2002) focused the discussions about what voice a language learner should imitate, that is, what factors lead to a "*golden speaker*" for learners to imitate. Their research suggested that foreign language learners imitating speakers whose voice features are similar to theirs would find pronunciation learning easier. In other words, the golden speaker voice would serve as a recast for the learner's production. The authors also suggested that speech rate may be a primary contributor more to speech similarity. Other research also shows that learners' imitation preferences may depend on their language background and proficiency as well as learning stage. For instance, speed of utterance preferences of learners may go from slower to faster once they feel comfortable with pronunciation features of an utterance (Wang and Lu, 2011). Probst et al. (2002) concluded that a CAPT program should provide learners multiple golden speakers to listen to; Wang and Lu (2011) suggested that this means that learners should be given a chance to control voice modification features such as different speech rates and pitch formants, based on the learners' own preferences.

2.2. Self-imitation in pronunciation training

A handful of studies have examined the possibility of modifying the learner's own voice and using it for pronunciation training (Hirose et al., 2003; Peabody and Seneff, 2006; Bissiri and Pfitzinger, 2009; Bissiri et al., 2006; De Meo et al., 2012; Pellegrino and

Vigliano, 2015). In early work, Nagano and Ozawa (1990) evaluated a prosodic-conversion method to teach English pronunciation to Japanese learners. One group of students was trained to mimic utterances from a reference English speaker whereas a second group was trained to mimic utterances of their own voices, previously modified to match the prosody of the reference English speaker. Post-training utterances from the second group of students were rated as more native-like than those from the first group. More recently, Bissiri and Pfitzinger (2009) and Bissiri et al. (2006) used prosodic modification to teach German lexical stress to Italian speakers. Receiving feedback in the form of the learner's own voice (resynthesized to match the local speech rate, intonation and intensity of a reference German speaker) was shown to be more effective than receiving feedback in the voice of the reference German speaker. Providing feedback in the learner's own voice also had a motivating effect, with several participants asking to continue the training, whereas participants in the control group showed no particular interest.

Pronunciation training with prosodic modifications of the learner's utterances has been shown to improve not only accentedness but also intelligibility. De Meo et al. (2012) evaluated the effectiveness of two forms of training (imitation and self-imitation) to teach suprasegmental patterns of Italian to Chinese learners. Participants in the self-imitation condition heard their own voice, resynthesized to match the native model, whereas those in the imitation condition followed traditional imitation exercises. Native listeners were then asked to classify learners' post-training productions as belonging to one of four speech acts: requests, orders, granting, and threats. Classification performance was significantly higher for utterances from participants in the self-imitation group. Similar improvements in communicative effectiveness were obtained in a later study with Japanese learners of L2 Italian Pellegrino and Vigliano, 2015). These studies show that (1) *prosodic* accent conversions are an effective tool to teach pronunciation to L2 learners and (2) the effect is robust across several L1-L2 combinations. Incorporating segmental accent conversion–the next logical step in this new genre of technology–is the major contribution of our work.

2.2.1. Algorithms for segmental accent conversion

In contrast with the self-imitation literature, where no studies exist that incorporate segmental adjustments of the learner's own voice, the speech-processing literature offers a few studies on speech modification of segmental errors in non-native speech. These studies have shown that segmental modifications are more effective at reducing the perceived accent of an utterance than prosody modification alone, both within regional accents of the same language (Yan et al., 2007) and across languages (Felps et al., 2009).

In early work, Yan et al. (2007) developed a method to transform vowels of three major regional English accents (British, Australian, and General American). The authors built a statistical model of vowel formant ratios from multiple speakers, and then extracted empirical rules to modify pitch patterns and vowel durations across the three accents. Using this model, the authors then adjusted formant frequencies, pitch patterns and vowel durations of an utterance to match a desired target accent. In an ABX test, 78% of Australian-to-British accent conversions were perceived as having a British accent, and 71% of the British-to-American accent conversions were perceived to have an American accent. In both cases, changing prosody alone (pitch and duration) led to noticeable changes in perceived accent, though not as significantly as formant modifications. The method hinged on being able to extract formant frequencies, so it cannot be easily extended to larger corpora because formant frequencies are ill-defined for unvoiced phones and cannot be tracked reliably even in voiced segments.

A few studies have attempted to blend L2 and L1 vocal tract spectra instead of completely replacing one with the other. In one such study, Huckvale and Yanagisawa (2007) reported improvements in intelligibility for Japanese utterances produced by an English text-to-speech (TTS) after blending their spectral envelope with that of an utterance of the same sentence produced by a Japanese TTS. Felps et al. (2009) proposed a method that was suitable for voiced as well as unvoiced phones. The authors split short-time spectra into a spectral envelope and flat glottal spectra. Then, they replaced the spectral envelope of an L2 utterance with a frequency-warped spectral envelope of a parallel L1 utterance and recombined it with the L2 glottal excitation. Listening tests showed a significant reduction in accent following segmental modification. More recently, Aryal et al. (2013) presented a voice morphing strategy that can be used to generate a continuum of accent transformations between an L2 speaker and a native speaker. The approach decomposes the speech Cepstrum into spectral slope and spectral detail, then generates accent conversions by combining the spectral slope of the L2 speaker with a morph of the spectral detail of the native speaker. This morphing technique provides a tradeoff between reducing the accent and preserving the voice identity of the L2 learner, and it may serve as a behavioral shaping strategy in computer assisted pronunciation training.

Accents originate from differences in articulation, which suggest that articulatory information may be useful in accent conversion. To explore this possibility, Felps et al. (2012) used concatenative speech synthesis to replace mispronounced diphones in an L2 utterance with other L2 diphones whose articulatory configuration was similar to a reference native utterance. The approach reduced the perceived non-native accents by 20%, but performed poorly when tasked with finding phonemes that the L2 did not utter. To address this problem, Aryal and Gutierrez-Osuna (2015) proposed a statistical parametric approach, which trains a Gaussian Mixture Model-based articulatory synthesizer for the L2 speaker, then drives it with articulatory data from a reference native utterance mapped to the L2 articulatory space via a Procrustes transform. In listening tests, the authors found that the method reduced the perceived non-native accents while preserving the voice quality of the L2 speaker. However, these methods require articulatory data, which is impractical for pronunciation training.

2.2.2. Accent conversion vs. voice conversion

Accent conversion is closely related to the problem of voice conversion (VC) (Mohammadi and Kain, 2017). Voice conversion transforms utterances from a source speaker to appear as if a (known) target speaker had produced them. To be successful, the conversion should match multiple identity cues of the target speaker, including but not limited to vocal tract configurations, prosody, pitch range, accent/dialect, and speaking rate. Ideally, the only information retained from the source utterance is its linguistic content, i.e., what has been said. Accent conversion goes one step further, since it attempts to capture both the linguistic content and the pronunciation of the source utterace, and combine it with the voice quality of the target speaker (i.e., those aspects associated with the target speaker's physiology), to create a new voice that sounds like the target speaker speaking with the source speaker's pronunciation. Therefore, accent conversion is a more challenging problem than voice conversion since ground truth for the output voice (i.e., the L2 learner's voice with a native accent) is not available.

2.3. Comprehensibility and fluency

Comprehensibility, along with accentedness and intelligibility, as operationalized by Munro and Derwing (1995), refer to partially independent measures of speech understanding. Comprehensibility is a measure of the amount of effort a listener puts forth in understanding and is partially tied to pronunciation, but is also a function of discourse patterns, lexico-grammar, and fluency measures. Accentedness is a measure of the perceived difference of a speaker's pronunciation from a reference accent. Intelligibility is a measure of how a listener actually understands a speaker, whether in decoding words, understanding the message, or understanding the intentions (Levis, 2018). It is not typically measured on a scale. Fluency is not directly connected to pronunciation accuracy, but is instead a measure of how automatically speech is produced. This paper focused on comprehensibility and fluency, each of which was mea-

sured using a 10-point Likert scale (0–9) in which the two endpoints of the scale were specified but the midpoints were not.

2.3.1. Comprehensibility

Comprehensibility refers to the amount of cognitive effort put forth by listeners in understanding speech (Derwing and Munro, 2015). Highly comprehensible speech is thus easy to understand, taking little extra effort. The difference between comprehensibility and accentedness is important to keep in mind in evaluating the success of pronunciation training because comprehensibility may be a better predictor of communicative success than accentedness (Derwing and Munro, 1997).

Unlike accentedness ratings, comprehensibility ratings correlate with a wide range of features beyond pronunciation. Isaacs and Trofimovich (2012) showed this in an examination of factors that were implicated in different ratings of comprehensibility. In their study, the researchers specified 19 quantitatively scored speech measures, including pronunciation features related to segmentals and suprasegmentals, fluency features, features related to vocabulary and grammatical complexity, and discourse features related to the construction of oral texts. They analyzed and coded the speech samples of 40 French learners of English, and the scores based on their analysis were subjected to a correlation with the comprehensibility ratings of naive native speaker (NS) raters. They found that most of the features and categories correlated with differences in comprehensibility ratings. This suggests that changes in one feature alone may not necessarily improve comprehensibility and that evaluations of comprehensibility are not connected to pronunciation directly. Rather, comprehensibility judgments also include other features of speech.

Other studies also suggest that comprehensibility is not based only on pronunciation. In one study, Tyler (1992) used two transcribed presentations, one originally given by a non-native speaker (NNS) and one by an NS. To remove the effect of pronunciation, both presentations were read aloud by another NS. The NNS presentation was rated as being less clear and more difficult to follow (that is, it was less comprehensible). The researcher argued that the use of unexpected, nonparallel discourse markers, unclear anaphoric reference, and over-use of coordination were the cause of the difficulties.

This does not mean, however, that pronunciation is irrelevant to improvements in comprehensibility, Derwing et al. (1998) found that instruction on prosodic skills and general fluency resulted in higher comprehensibility for L2 learners' spontaneous speech, while equivalent instruction on segmentals did not result in spontaneous speech improvement. Gordon and Darcy (2016) confirmed this finding, albeit for a shorter treatment. Derwing and Rossiter (2003) similarly found that comprehensibility ratings for an approach focusing more heavily on suprasegmentals showed greater improvement than a segmental approach.

2.3.2. Fluency

Fluency, another feature assessed in this study, has been used with a variety of meanings: general proficiency (Fillmore, 1979) and smooth delivery (Lennon, 1990; Riggenbach, 2000) are two of the most common. Fluency is connected to a wide variety of temporal features of speech (i.e., speech rate, the use of pauses, and repairs), the use of formulaic language (Ejzenberg, 2000), whether phrases are logically constructed (Nakatani and Hirschberg, 1994), phonological features of speech (Wennerstrom, 2000), interactive characteristics of speech in conversation (Riggenbach, 1991), perceived smoothness of speech by listeners (Derwing et al., 2006), mean length of run (see (Lennon, 1990)), and automaticity of speech production (Segalowitz, 2007). Automaticity in turn is connected to phonological memory and attention control (Segalowitz, 2007; O'brien et al., 2007).

Fluency is not independent of accentedness and comprehensibility but is indirectly related to both. For example, comprehensibility ratings correlate with elements related to fluency (Isaacs and Trofimovich, 2012). Speech rate is also predictive of fluency judgments (Cucchiarini et al., 2000; Kormos and Dénes, 2004), and similar judgments of fluency may be given for speech at different rates. Listeners are sensitive to whether speech is fluent, and speech that is heard as too fast or too slow may also be heard as more accented or as less comprehensible (Derwing et al., 1998).

In relation to research on pronunciation, fluency may be measured by evaluating speech features such as speech rate or articulation rate, or it may be measured using Likert scales to capture perceptual evaluations by asking listeners to assign a score using a value between the two ends of a scale.

2.4. Effects of instruction

A robust finding of pronunciation instruction is that it works. Three recent studies show that whether instruction comes from human teachers or in CAPT, significant improvements are the norm. In the first, Saito (2012) looked at 15 pre-/post-test design studies to see whether instruction led to improved pronunciation and found that explicit attention to pronunciation typically led to improvement. Improvement was more common in controlled tasks and less common in spontaneous speech.

In a second study, Lee et al. (2014) conducted a meta-analysis of 86 studies to explore the success of pronunciation instruction. Instruction typically resulted in a relatively large degree of improvement, especially when the instruction was carried out over longer time periods, when there was consistent feedback to learners, and again in more controlled tasks (such as reading aloud or imitation). This is perhaps not surprising since most studies have used controlled tasks. Relevant to this study, most studies employed university students.

In a corresponding narrative analysis, Thomson (2012) and Thomson and Derwing (2014) analyzed most of the studies in Lee et al. (2014), but focusing instead on criteria from research for what pronunciation training should be like. The studies were mostly about segmental improvement. The kind of instruction was usually underspecified. Few studies (9%) have focused on improvements in comprehensibility and intelligibility. This study examines improvements in comprehensibility, but most results that show improvements in global ratings privilege prosody rather than segmentals.

In all three reviews, few studies used delayed post-tests, so it was unclear whether improvement continued past the intervention. These analyses suggest that interventions should be successful, and that explicit attention to pronunciation should lead to improvement. However, they do not indicate whether more implicit feedback based on a Golden Speaker voice will be sufficient to show improvement in comprehensibility and fluency.

3. System description

To answer the Research Questions presented earlier, we developed Golden Speaker Builder (GSB), an online interactive tool that allows L2 learners to build a personalized pronunciation model: their own voice producing native-accented speech (i.e. a "golden speaker"). To build their "golden speaker", L2 learners follow three steps. In the first step, the learner records a keyword for each phone (e.g., for phoneme /ʒ/, the learner records the keyword "vision") under the guidance of an instructor to ensure that the utterance has near-native production. After recording each keyword, the learner segments the phone using a graphical display of the waveform. In the second step, the learner records several sentences, which are used to estimate the learner's pitch statistics. In a final step, the learner selects a native speaker as a source model, and GSB resynthesizes the native speaker's sentences using the recorded phone segments and prosody statistics of learner. The process can be completed in less than thirty minutes and generates a Golden Speaker voice that produces intelligible speech with the voice quality of the L2 learner, and the prosody of the source native speaker normalized to the pitch range of the L2 learner.

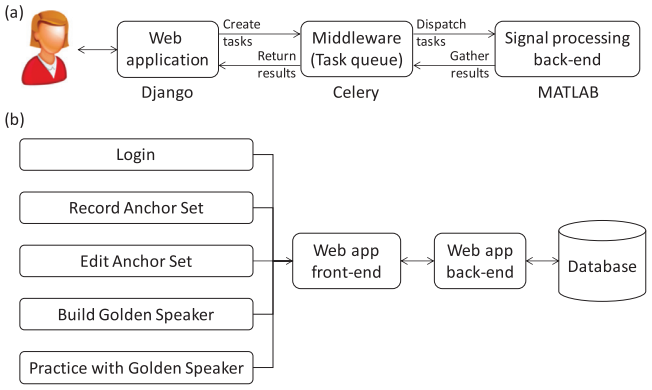

Fig. 1. (a) Overall software architecture. (b) Architecture of the web application.

The software architecture of GSB is shown in Fig. 1. GSB consists of three components: a web application, a signal processing back-end, and a middleware to connect the signal processing back-end to the web application. The web application provides a graphical interface for the learner, responds to the learner's requests, and stores the learner's data (i.e., login information, speech recordings, and golden speakers) onto a database – see Fig. 1b. The signal processing back-end runs the accent conversion algorithms, which generates synthesized speech for each Golden Speaker model. Finally, the middleware layer provides communication between the web application and the signal processing back-end via an asynchronous task queue. Detailed descriptions of each component are included in the following subsections.

3.1. Web application

We implemented the web application using the Django framework.[2] The web-app front-end was written in HTML5 and Javascript, and decorated with Bootstrap,[3] whereas the web-app back-end was written in Python with Django internal modules. User data is managed by an SQLite database engine[4] on a standard Linux file system. We hosted the web application through Nginx.[5] To follow the workflow described below, we provide five functional modules: Login; Record Anchor Set; Edit Anchor Set; Build "Golden Speaker"; and Practice with "Golden Speaker".

The **Login** module provides registration and login functions. To use GSB, learners must register an account using their email, and login with their registered account and password. We implemented this module using Auth0 authentication,[6] and connected Auth0 to the SQLite database to save the users' account information. This module guarantees the privacy of learners' information and ensures that each learner can only operate on their own information and data.

The **Record Anchor Set** module enables learners to record keywords and prosody sentences, later used to build a Golden Speaker model. As shown in Fig. 2, the learner must record a keyword for each of the 40 phones in American English (CMU phone set[7]). Once a user records a keyword, the interface allows the learner to segment the phone segment (or "Anchor") by highlighting the corresponding region of the speech waveform. Separate tabs are used for consonants, vowels, and pitch sentences. Consonants are arranged according to their place and manner of articulation, and vowels are arranged according to their frontness and height (not shown). This arrangement allows the teacher and learner to review the basic organization of speech sounds in English, as the learner records the various keywords. The "Pitch Sentences" tab includes 30 sentences representative of conversational speech (e.g., "What time does the bus leave for the airport?") that were deliberately selected to provide good coverage of various prosodic contexts, and a free-speech exercise in which the learner first watches a 3-minute short film[8] and then records a 1–2 min audio summary. Recordings for all the keywords and pitch sentences are saved on the file system, whereas the segmentation information is saved in the database. In a final step, both the recordings and the segmentation information are sent to the signal processing back-end.

Table 2
Keyword selection. The following is a list of keywords used to build anchor sets for L2 learners in the GSB application. Phoneme names are shown on the left column in ARPABET notation, and the words used to elicit the phoneme on the left.

AA	*father*	**CH**	*cheat*	**HH**	*heat*	**NG**	*sing*	**TH**	*think*
AE	*ash*	**D**	*deep*	**IH**	*if*	**OW**	*oh*	**UH**	*push*
AH	*us*	**DH**	*this*	**IY**	*east*	**OY**	*toy*	**UW**	*boot*
AO	*horse*	**EH**	*"s"*	**JH**	*jeep*	**P**	*poke*	**V**	*vote*
AW	*ouch*	**ER**	*earth*	**K**	*keep*	**R**	*reads*	**W**	*weeds*
AX	*sofa*	**EY**	*ace*	**L**	*leads*	**S**	*See*	**Y**	*yes*
AY	*ice*	**F**	*feed*	**M**	*make*	**SH**	*sheep*	**Z**	*zoo*
B	*boat*	**G**	*gust*	**N**	*no*	**T**	*tea*	**ZH**	*vision*

We selected one keyword per phoneme to capture an "ideal" example of that phoneme or its main characteristic, e.g., the dominant allophone of that phoneme. Voiceless aspirated stops are more distinct than unvoiced aspirated stops, and were chosen preferentially for that reason. Additionally, final stops were avoided, as well as final rhotics and velarized approximants (e.g. "dark L"). The full selection of keywords is shown in **Table 2**.

The **Edit Anchor Set** module allows learners to make changes to a previously recorded "Anchor Set". This includes re-recording specific keywords or pitch sentences, and making corrections to the segmentations. Learners also have the option to rename, copy, and delete the Anchor Sets from their profile. Once an Anchor Set is modified, the updated recordings and segmentation information are automatically sent to the signal processing back-end.

The **Build Golden Speaker** module allows learners to select one of several Native Speaker (NS) voices, each containing hundreds of sentences, and pair it with one of their own Anchor Sets. Once a particular NS voice, Anchor Set, and list of sentences has been selected, this information is sent to the signal-processing back-end to build the Golden Speaker model.

The **Practice with Golden Speaker** module allows the learner to practice pronunciation with any of the previously-built Golden Speakers. For example, we used a *backward buildup* exercise as one technique for pronunciation practice, where the learner practices a long sentence starting from the last phrase and adding complexity in a backwards fashion. As an example, given the practice sentence "*We're going to the supermarket to buy vegetables for dinner,*" the learner produces the phrase "*for dinner*," then the phrase "*to buy vegetables for dinner*" and so forth.

3.2. Speech processing back-end

To build Golden Speakers, the signal processing back-end uses a Sparse, Anchor-Based Representation (SABR) reported in prior work (Liberatore et al., 2015; Liberatore et al., 2018). The motivation behind SABR is to separate speaker-dependent cues (*how* something was said) from speaker-independent ones (*what* was said). SABR performs this decomposition by representing speech as a sparse, weighted sum of

[2] https://www.djangoproject.com/.
[3] https://getbootstrap.com/.
[4] https://www.sqlite.org/.
[5] https://www.nginx.com/.
[6] https://auth0.com/.
[7] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[8] "Spellbound" by Ying Wu and Lizzia Xu; available at youtube.com/watch?v = W_B2UZ_ZoxU

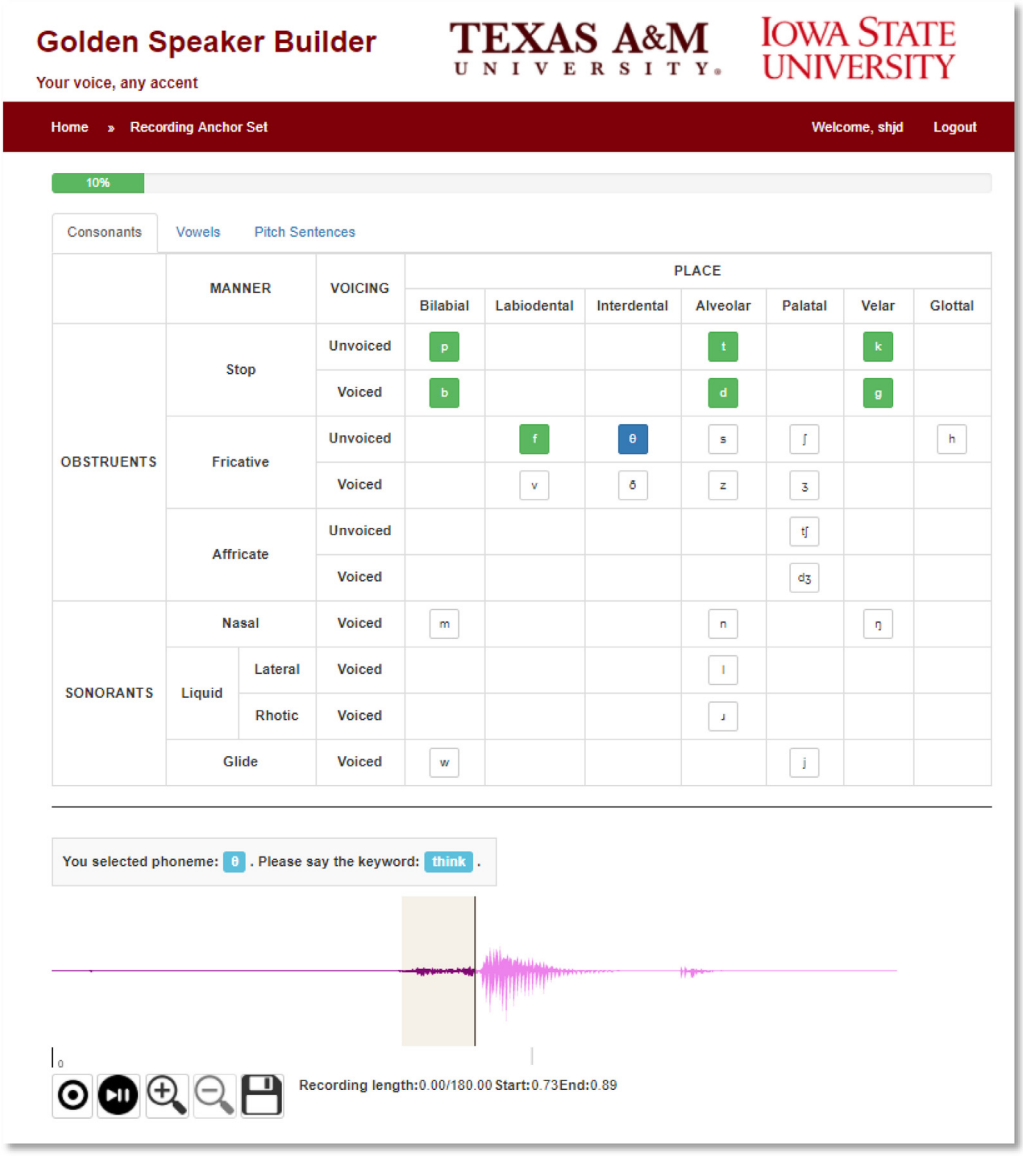

Fig. 2. Graphical user interface for recording consonants in American English. In the example shown, the learner has already recorded keywords for all the stop consonants (highlighted in green), has recorded the phone */θ/* (highlighted in blue) and is in the process of selecting the appropriate section in the speech waveform shown at the bottom of the page.

acoustic "anchors":

$$X \cong A_S W_S \tag{1}$$

where each column in matrix X represents an analysis window (i.e., a vector of MFCCs), A_S is a matrix of anchors for speaker S (one anchor per phone), and W_S is the utterance's weight matrix. If there are M acoustic frames in an utterance, N acoustic features (i.e., MFCCs), and K speaker anchors, then $X \in \mathrm{R}^{N\times M}$, $A_s \in \mathrm{R}^{N\times K}$, and $W_S \in \mathrm{R}^{K\times M}$.

Due to the sparse nature of speech, a natural way to perform the decomposition is via sparse coding: minimize the reconstruction error $\|X - A_S W_S\|$ while also minimizing the number of basis vectors used in the decomposition. In SABR, we use the nonnegative Lasso (Tibshirani, 1996):

$$\min_{W_S}\left(\|X - A_S W_S\| + \lambda||W_S||_1\right) s.t.\ W_S > 0 \tag{2}$$

where λ is a regularization term that balances the reconstruction and sparsity criteria, and $\|\cdot\|_1$ is the L1 norm (i.e., Manhattan distance). To solve for the Lasso, our implementation uses the Least Angle Regression (LARS) (Efron et al., 2004) algorithm.

Given anchor sets A_S and A_T for source and target speakers, respectively, SABR provides a straightforward way to perform voice conversion: for each source utterance X_S, compute the weight matrix W_S relative to the source anchors A_S, then combine it with the target anchors:

$$\hat{X}_T = A_T W_S \tag{3}$$

In the case of GSB, source anchors are precomputed in advance for each of the native speaker voices, whereas target anchors are obtained from the learner's Anchor Set. First, we compute the STRAIGHT (Kawahara, 2006) spectral envelope and compress it to 25 MFCCS (25 Mel-filterbanks, 25 coefficients, 8 kHz cutoff). Then, we separate energy ($MFCC_0$) and use the remaining coefficients ($MFCC_{1-24}$) in Eq. (3). After converting these coefficients, we append the source $MFCC_0$ and back-project the MFCCs into the STRAIGHT spectrum. Finally, we transform the pitch track F_0^S to match the target speaker's pitch range using log mean and variance scaling:

$$\log \hat{F}_0^T = \sigma_T \frac{\log\left(F_0^S\right) - \mu_S}{\sigma_S} + \mu_T \tag{4}$$

where μ_S, μ_T and σ_S, σ_T are the mean and variance of the log of the source and target speaker's pitch distributions, respectively.

3.2.1. Residual warping

Eq. (3) can lead to "muffled" speech that has low quality and lacks spectral detail since the original encoding in Eq. (1) discards the residual component R_S:

$$X_S = A_S W_S + R_S \tag{5}$$

which typically has a magnitude of 1.5 dB (Liberatore et al., 2015). To improve synthesis quality, one may be tempted to add the source residual R_S back into Eq. (3). Unfortunately, the residual R_S oftentimes carries speaker-specific information. As a result, naïvely adding it to the reconstructed target spectrum $\hat{X}_T$ alters the voice identity of the "Golden Speaker", moving it away from that of the learner.

To address this issue, GSB adds the residual reconstruction error R_S to the reconstructed target spectrum $\hat{X}_T$ via an intermediate function $F(\cdot)$:

$$\hat{X}_T = A_T W_S + F(R_S) \tag{6}$$

which transforms residuals from the source acoustic space to the target acoustic space. Namely, for each pair of source-target anchors A_S^k and A_T^k, we select the frequency warp that minimizes the SSE of the warped source and target anchors. Then, at runtime, we use the SABR weights W_S to compute a warping function for each individual frame.

Following Panchapagesan and Alwan (2009), we use a piecewise linear warping function that has two free parameters: an inflection point ω_0 (normalized frequency), and the slope p of the warping from 0 to ω_0:

$$f_{pw}(\omega;\omega_0,p) = \begin{cases} p\omega, & 0 \le \omega \le \omega_0 \\ p\omega_0 + \left(\frac{1-p\omega_0}{1-\omega_0}\right)(\omega-\omega_0), & \omega_0 < \omega \le 1 \end{cases} \tag{7}$$

When using cepstral coefficients, the transform in Eq. (7) can be expressed as a linear transform. Following (Panchapagesan and Alwan, 2009), we compute this transform as a product of a Discrete Cosine Transform (DCT) matrix C and its warped inverse (IDCT) $\hat{C}$. Assuming M filters in an MFCC filterbank, N cepstral coefficients, and a warping function $f(\omega)$, matrices $C \in \mathcal{R}^{N\times M}$ and $\hat{C} \in \mathcal{R}^{M\times N}$ can be computed as:

$$C_{m,k}{}^T = \left[\alpha_k \cos\left(\pi k \omega_m\right)\right]_{\substack{1 \le m \le M \\ 0 \le k \le N-1}} \tag{8}$$

$$\hat{C}_{m,k} = \left[\alpha_k \cos\left(\pi k f(\omega_m)\right)\right]_{\substack{1 \le m \le M \\ 0 \le k \le N-1}} \tag{9}$$

where α_k is a term to ensure that the DCT is unitary, and ω_m is the normalized frequency for the mth Mel filter. The linear warping of the MFCCs is $T = C\hat{C}$, where $T \in \mathcal{R}^{N\times N}$. Substituting $f_{pw}(\cdot)$ from Eq. (7) into Eq. (9), the transform becomes a function of ω_0 and p:

$$T(\omega_0, p) = C\hat{C}(\omega_0, p) \tag{10}$$

For each pair of source-target anchors A_S^k and A_T^k, we create a transform T_k by selecting ω_0 and p to minimize the SSE of the transformed source and target anchors:

$$T_k = \underset{T(\omega_0,p)}{\operatorname{argmin}} \sum \left(T(\omega_0,p)A_S^k - A_T^k\right)^2 \tag{11}$$

Following Pitz and Ney (2005), we constrain the inflection frequency $\omega_0 \in [0.4, 0.8]$ and the warping slope $p \in [0.8, 1.2]$. The resulting residual warping VC method is similar to Weighted Frequency Warping (Erro et al., 2010).

The final transform is the weighted sum of the individual anchor transforms T_k, where we add a single row $W_{k+1} = 1 - W_{1\ldots k1}$ to ensure the weights sum to 1, and set the corresponding warp $T_{k+1} = I$. For each source frame $X_{S,\,i}$, SABR weight vector $W_{S,\,i}$, and the frame residual $R_{S,\,i}$, we estimate the target speaker's spectrum $X_{T,\,i}$ as:

$$\hat{X}_{T,i} = A_T W_{S,i} + \left(\sum_{k=1}^{K+1} W_{S,i,k} T_k\right) R_{S,i} \tag{12}$$

Because of the sparsity imposed in Eq. (2), the resulting residual transform matrix favors weights on or near the diagonal, a cepstral VTLN property noted by Pitz and Ney (2005).

3.3. Middleware

GSB uses an asynchronous task queue, Solem (2016), as the middleware to communicate between the web application and the signal processing back-end. Each time the user submits a request containing signal processing operations, the web application creates a task worker and pushes it into the asynchronous task queue. Tasks in the queue are then dispatched to an available worker, which in turn calls the appropriate signal processing function in the back-end. Once the task is complete, results are sent back to the web application through the asynchronous task queue, and the worker is set to be available.

Two types of signal-processing tasks are included in GSB: (1) building a SABR model for a given Anchor Set, and (2) synthesizing speech for a "Golden Speaker". Tasks of the first type are dispatched after a complete Anchor Set is recorded and saved. This involves passing all the recordings (keywords, pitch sentences) and segment information to the signal processing back-end, saving the SABR model (i.e., target anchors A_T and pitch statistics μ_T, σ_T) to the file system, and passing the corresponding path to the web application so it can be stored in the database. The run time to build a SABR model is 10 min, largely due to the STRAIGHT speech analysis (~5 s processing time for 1 s of speech). Tasks of the second type are dispatched when the user submits a request to build a "Golden Speaker". This involves passing the following information to the signal-processing backend: the teacher's SABR model (computed far in advance), the learner's SABR model (computed from the Anchor Set), and a list of sentences the learner wants to synthesize. Once these sentences have been re-synthesized as a "Golden Speaker", the recordings are saved to the Linux file system, and the corresponding path is returned to the web application so it can be stored in the database. The run time for this type of task is approximately 10 s/sentence.

4. listening tests

We conducted a series of perceptual listening tests to determine how successful GSB was in generating golden speaker voices. First, we conducted a voice-identity test to assess whether the golden speaker captures the learner's voice quality, which is the most significant goal to achieve. Next, we conducted an accentedness test to determine if the GSB syntheses have native-like accents, a goal that is also critical for our application. Finally, as a common practice in speech-synthesis related tasks, we evaluated the audio quality of the syntheses through a standard MOS test.

4.1. Speech corpus

The speech corpus used for these perceptual listening tests consisted of recordings from L1 speakers (the "teacher" voices), L2 speakers (the "learner" voices) and golden speaker voices of the L2 speakers using the L1 speakers as models. For this purpose, first we recorded two American English speakers (CBL: male; GMA: female) as teacher voices. Each speaker produced 100 utterances from the ARCTIC corpus (Kominek and Black, 2004), from which we built the SABR models, and an additional 24 utterances to be used as "training" utterances for participants in the pronunciation training experiment (reported in section 5). To generate SABR models for each teacher, we extracted phoneme labels using the Montreal forced-aligner (McAuliffe et al., 2017). Namely, for each

phoneme in the GSB "Record Anchor Set" interface ($N = 40$), we extracted a single phoneme anchor corresponding to the centroid of all frames in the 100 utterances that were labeled with the corresponding phoneme.

Next, we recruited 18 L2 learners of American English to participate in the pronunciation training study; see Section 5.1 for details. Each L2 learner recorded a set of keywords and prosody sentences, from which we built their corresponding SABR model. Then, L2 learners practiced with the 24 training utterances and recorded them pre- and post-treatment. Two of the L2 learners did not finish the study and another one L2 learner did not record their post-test sentences. Consequently, we have speech data from 15 learners (8 males, 7 females). Of these, we used speech data from 6 learners[9] (3 males, 3 females) for the perceptual listening tests reported here. To obtain golden-speaker voices, we paired the 3 male L2 learners with the male L1 teacher voice (CBL), and the 3 female L2 learners with the female L1 teacher voice (GMA).

4.2. Perceptual studies

For each pair of L1-L2 speakers, we evaluated the golden-speaker voice against a control. The golden-speaker condition (**GS**) used a SABR model for the L2 learner where each phoneme anchor was obtained from the corresponding keyword segment, as originally segmented by the L2 learner–see Fig. 2, as well as the prosody sentences (forced aligned with the Montreal forced-aligner). The control condition used a golden-speaker that only applies a pitch transformation (**PT**) (Martin, 2004; Genevalogic 2006) to the L1 teacher voice to match the pitch range of the learner.

We conducted the perceptual listening tests on Amazon Mechanical Turk to evaluate the non-native voice identity, accentedness, and acoustic quality of the two golden-speaker voices.[10] Recordings in each listening test were randomly ordered. We also included 12 calibration utterances in each listening test to detect if listeners were not attending adequately to the task (Buchholz and Latorre, 2011). If so, we removed their responses from the sample.

4.2.1. Voice identity

We evaluated the voice identity of the syntheses using a Voice Similarity Score (VSS) test Felps et al., 2009; Kreiman and Papcun, 1991). Namely, participants listened to pairs of utterances and were required to ((1) decide whether the two utterances were from the same speaker, and (2) then rate their confidence in the decision on a 7-point scale, as in Pelham and Blanton (2012). For each utterance pair, one was a testing utterance randomly sampled from one of the two golden-speaker voices; the other was a reference utterance randomly sampled from either the corresponding source or target speaker. The VSS was then computed by collapsing the above two fields into a 15-point scale from −7 (definitely different speakers) to +7 (definitely the same speaker). Listeners ($n = 30$) rated the VSS of 108 utterance pairs. We used 48 pairs of utterances for each synthesis condition (GS and PT)—8 pairs per L1-L2 speaker pair (4 AC-L1, 4 AC-L2), and 12 pairs of unmodified source and target utterances to ensure participants did not cheat. Following Felps et al. (2009), we played utterances in reverse to reduce the influence of accents in the perception of voice identity.

Results are shown in Fig. 3. For GS voices, listeners were quite confident that the syntheses and the original L1 recordings are from different speakers and they were sligtly confident that the syntheses and the original L2 recordings are from the same speaker (GS system, AC-L1: −4.41; AC-L2 2.00; $p \ll 0.001$, single-tailed T-tests). In contrast, listeners were quite confident that syntheses from pitch transformation were from the same speaker as the original L1 recordings and were somewhat confident that they were from different speakers than the original L2 recordings (PT system, AC-L1: 4.46; AC-L2: −2.94; $p \ll 0.001$, single-tailed T-tests). Both the AC-L1 and AC-L2 distributions were significantly different for the GS and PT systems ($p \ll 0.001$, two-tailed T-test). Thus, PT syntheses were perceived as being very close to the L1 speaker and very different from the L2 learners, whereas GS voices were rated as being very different from the L1 speaker, and close to the voice of the L2 learners, indicating a good identity match.

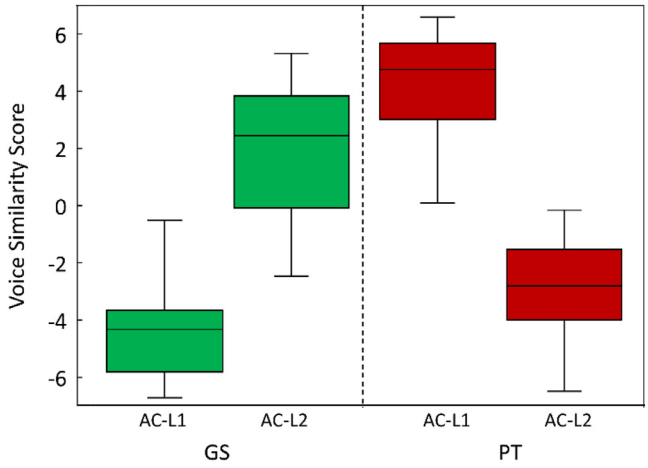

Fig. 3. Voice identity ratings. The range is from −7 (definitely different speakers) to +7 (definitely the same speaker).

4.2.2. Accentedness

Following Munro and Derwing (1995), we used a scaled-rating test to establish the degree of accentedness of individual utterances. Listeners ($n = 27$) rated the foreign accentedness (1-No foreign accent, 9-Very strong foreign accent) of 120 utterances. The utterances were from either of the two test conditions above (GS, PT), from the source native speakers, or from the target foreign speakers. We used 30 utterances for each of the test conditions and the target foreign speakers (5 utterances for each of 6 learners, 30 utterances). For both of the source native speakers, we selected 15 utterances to ensure a class balance in the test.

Results are summarized in Fig. 4(a). As expected, original utterances from the L1 speakers received the lowest ratings for foreign accentedness (1.11), whereas those from the L2 learners received highest ratings (7.44). PT achieved similar ratings as the original L1 utterances (1.17; $p = 0.236$, two-tailed t-test), which is to be expected since pitch-transformed utterances are identical to L1 utterances except for their pitch range. Finally, the GS voice was rated as being significantly less accented (2.42) than the L2 utterances (7.44; $p \ll 0.001$, two-tailed t-test) but not as much as L1 utterances (1.11; $p \ll 0.001$, two-tailed t-test) or the PT utterances (1.17; $p \ll 0.001$, two-tailed t-test). In summary, the GS voice showed a significant decrease (~84%) in foreign accentedness compared to the original L2 speech.

4.2.3. Acoustic quality

We evaluated the acoustic quality of the two golden-speaker voices using a Mean Opinion Score (MOS) test. Listeners ($n = 28$) rated the MOS (1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent) of 120 utterances. We used the same test conditions as in the foreign accentedness test in the prior section.

Results are summarized in Fig. 4(b). Listeners rated original utterances from L1 speakers and pitch transformation as having the highest acoustic quality (L1: 4.66, PT: 4.56). Surprisingly, though, listeners gave

[9] We randomly selected 6 learners from the original set of 15 learners to ensure that listeners could complete the test within a reasonable time (within 30 minutes) to avoid fatigue.

[10] Following Aryal and Gutierrez-Osuna (2015) , "Reduction of non-native accents through statistical parametric articulatory synthesis," *The Journal of the Acoustical Society of America,* vol. 137, no. 1, pp. 433-446, 2015., all listeners were required to pass an American accent identification test prior to participating in the studies.

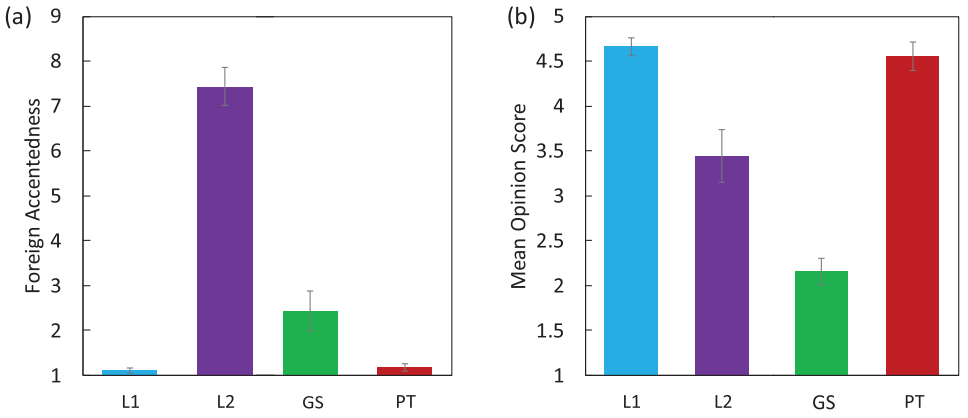

Fig. 4. (a) Foreign accentedness ratings. The rating ranges from 1 (no foreign accent) to 9 (very strong foreign accent). (b) Mean opinion score (MOS) of acoustic quality ratings with 95% confidence interval. The MOS scale is from 1 (bad) to 5 (excellent).

the L2 recordings a much lower MOS than the L1 (3.44; $p \ll 0.001$, two-tailed *t*-test), despite the fact that they were the original natural speech recordings, which indicates that ratings of acoustic quality are influenced by accentedness. Finally, listeners rated the GS voices as having lower quality (2.16) than the PT voices (4.56; $p \ll 0.001$, two-tailed *t*-test), due to distortions introduced in the accent-conversion algorithm. We anticipated this result, since the pitch transformation technique does not alter the speech spectrogram and distortions are minimal due to the use of the STRAIGHT vocoder, which produces high-quality speech analysis and reconstruction.

5. User study

We conducted a user study to validate GSB in a language-instruction setting with a population of Korean L2 learners of English. The study followed a quasi-experimental pre-, immediate post- and delayed post-test at a midwestern university in the USA. Learners took a pre-test followed by three weeks of CAPT using the GSB, followed by an immediate post-test one week after training and a delayed post-test three weeks after training. Learners were interviewed after each test session.

5.1. Participants

There were two groups of participants in this study: learners and raters. Learners were 15 Korean learners of English (eight male) majoring in various fields of study. Learners were recruited from undergraduate and graduate ESL courses when one of the researchers introduced the study in a classroom visit. Initially, 18 learners signed up to participate the study; however, we did not include the data from three of these participants since they missed at least two training sessions.

Raters included 95 native-English speaking undergraduate students majoring in different areas at the same university. These raters were part of two groups since comprehensibility ($n = 50$), and fluency ($n = 45$) were each rated by a separate group of raters. All raters were recruited from first- and second-year composition classes through the introduction of the study by one of the researchers in a classroom visit. Learners and raters were recruited through convenience sampling; that is, we collected data from all students who were willing to participate.

5.1.1. Pronunciation challenges for Korean speakers in English

We chose to use Korean speakers because of the high likelihood that they would have both segmental and suprasegmental difficulties with English. We also chose Korean learners because different Korean learners often have similar types of difficulties, even at more advanced levels of English proficiency. Among the most notable differences between the English and the Korean sound systems are that Korean vowels do not have a tense vs. lax distinction, and voiced and voiceless sounds are not regarded as different (Lee, 2001).

L1 Korean learners find both segmental and suprasegmental features of English challenging. Lee (2001) lists the vowel and consonants sounds of English most likely to cause issues. Among vowels, /ɔ/ is problematic, as it does not exist in Korean, so Korean speakers of English tend to assimilate it to a pure /o/ (Cho, 2004). Additionally, English /ʌ/ is often pronounced by Koreans as /ɑ/, while English /æ/ is assimilated to Korean /e/. The Korean sound system does not include the sound /ɝ/, which is frequently confused with /ɔ/. Therefore, differentiating words such as “work” and “walk” is difficult both in perception and production.

For consonant sounds, Korean learners of English do not have a voiced vs. voiceless distinction as in English. Therefore, word pairs such as “log” and “lock”, “raised” and “raced”, “beach” and “peach”, etc., are often confused (Lee, 2001). Voiced and voiceless distinctions are also not found in stops and affricates. Korean has three phonemic voiceless stops (such as /p/, pH/ and /pp/) for the bilabial, alveolar and velar places of articulation where English has two phonemes distinguished by voicing. The same pattern holds for the post-alveolar affricate /tʃ/. The lack of phonemic stop-fricative distinctions in Korean also leads to challenges with /b/-/v/ and /f/-/p/, as in “defend” and “depend” (Cho, 2004). Another common challenge is the English distinction between /ɹ/ and /l/, mapping to a single Korean phoneme. Other consonant sounds not found in Korean are /z/, /ð/, and /θ/, and they are frequently assimilated to /dʒ/, /d/, and /s/, respectively. Apart from having difficulties with consonant sounds because they are not present in the Korean sound system, Korean learners of English also have difficulties with certain similar consonant sounds in specific environments. So, /ʃ/ and /tʃ/ are part of Korean but are not found in syllable codas. As a result, Korean learners often add either /ɹ/ or /ə/ to English words ending in these sounds to match Korean syllable structure constraints (Lee, 2001).

Prosodically, in Korean each syllable has similar emphasis, and each word in a sentence has the same prominence. This may sometimes cause it to be characterized as monotonous-sounding (Cho, 2004). Korean and English also differ in the ways that they use intonation, and especially in how English uses flexibly-placed lexical prominence to call attention to information structure. Korean also has an accentual phrase that is defined by varied tonal patterns that do not map to equivalent patterns in English (Jun, 1995).

5.2. Materials

Materials used in this study included recordings of Korean learners' speech collected through a read-aloud task as well as three interviews done during the pre-, immediate post- and delayed post-tests.

Read-aloud Task. The read-aloud task included 48 sentences (Appendix A), 24 of which were modified from sentences taken from Carnegie Mellon University Arctic speech synthesis corpus (Kominek et al., 2003). The reasons for modifying the original sentences

were twofold: (a) to make them more readable by removing or changing problematic words such as proper names, and simplifying difficult sentence structures including infrequent syntactic patterns which are commonly used only in literary texts; and (b) to include words which were likely to contain sounds that were problematic for Korean learners. The other 24 sentences were adapted from United States State Department English as a Second Language materials[11] and posts on social media so that we had a representation of conversational sentences. These sentences were also modified in some cases to include words that contained problematic vowels and consonant sounds for Korean learners.

Interviews. The interviews included varying numbers of questions depending on the interview time (pre, post, delayed post). The purpose of these questions was to understand the educational background of learners, their use of English, and why they were interested in taking part in pronunciation training. Immediate and delayed post-test questions collected data about learners' use of and experience with the GSB and their self-evaluation of improvements as a result of the GSB training.

5.3. Procedures

Learners. In the pre-test learners were first interviewed about their personal and educational backgrounds, their use of English, and their interest in the pronunciation training. This helped us get a sense of what the learners thought about their own pronunciation and why they wanted to improve it. They then recorded sounds of English by producing key words and pitch sentences in the GSB tool (see Section 3.1) with one of the researchers present to guide them through the process. Finally, learners recorded a free speech sample by narrating a short video.[12] This video was chosen because it had an uncomplicated story line and required use of words that learners would be familiar with. Once learners recorded English sounds and sentences to estimate their voice pitch, they read aloud 48 sentences, 24 of which were used in the training.

In the week following the pre-test, learners started a three-week training program. During each week, learners came to a computer lab on the university campus three times. Each time, the learners spent thirty minutes using the training interface with headphones. The interface included the 24 training sentences, each of which was created from a synthesis of the learner's own voice with that of a native speaker. Learners practiced with 8 sentences in the first week, 16 sentences in the second week (reviewing the sentences from week 1 and adding 8 more), and all 24 sentences in the third week. The training interface included three types of exercises: say-listen-repeat, listen-repeat, and backward build-up (see Fig. 5). After becoming familiar with all three types of exercises, learners were free to use any format that they found useful. Learners were told to use the instructions provided in the training program but were encouraged to consult any of the research assistants in the room if they did not understand how to use something. After the first week, few learners asked any questions. In addtion, eye-tracking was used for all learners during the training, but is not reported in this paper.

Following the three weeks of training, learners took part in the immediate and delayed post-tests. Immediate post-tests were given in the week following the training, delayed post-tests were given three weeks after the training. In both of these tests, learners first recorded the 48 sentences, retold the story in the video, and were then interviewed.

Raters. Comprehensibility and fluency were each rated by a separate group of raters. Because raters could only rate between 260 and 360 sentences in the rating time, we chose to focus only on the first week's sentences (eight training sentences). We included the pre-test, post-test and delayed post-test sentences for each of the 15 learners, along with a set of six distractors from native speakers to verify the consistency of ratings. This meant that raters were ideally rating 360 sentences (15×3×8). The rating procedures for comprehensibility and fluency were the same. All sentences were uploaded to rating software developed by one of the researchers, and sentences were presented randomly. Because the rating task was completely randomized, the total number of sentences listened to by each rater for each dependent task varied (i.e. not all listeners listened to every sentence because of differences in how long it took to complete the task). In addition, three sentences from native speakers were included to check for rater attentiveness.

Raters listened to and evaluated as many as sentences as they could in the 50 min provided for the rating task. Raters evaluated each sentence they listened to based on a 10-point Likert scale. 0 represented a poor rating and 9 represented an excellent rating for each dependent variable. Before raters started the rating task, they listened to four training sentences so that they became familiar with the task. They were encouraged to use the whole scale and could ask questions if they did not understand anything.

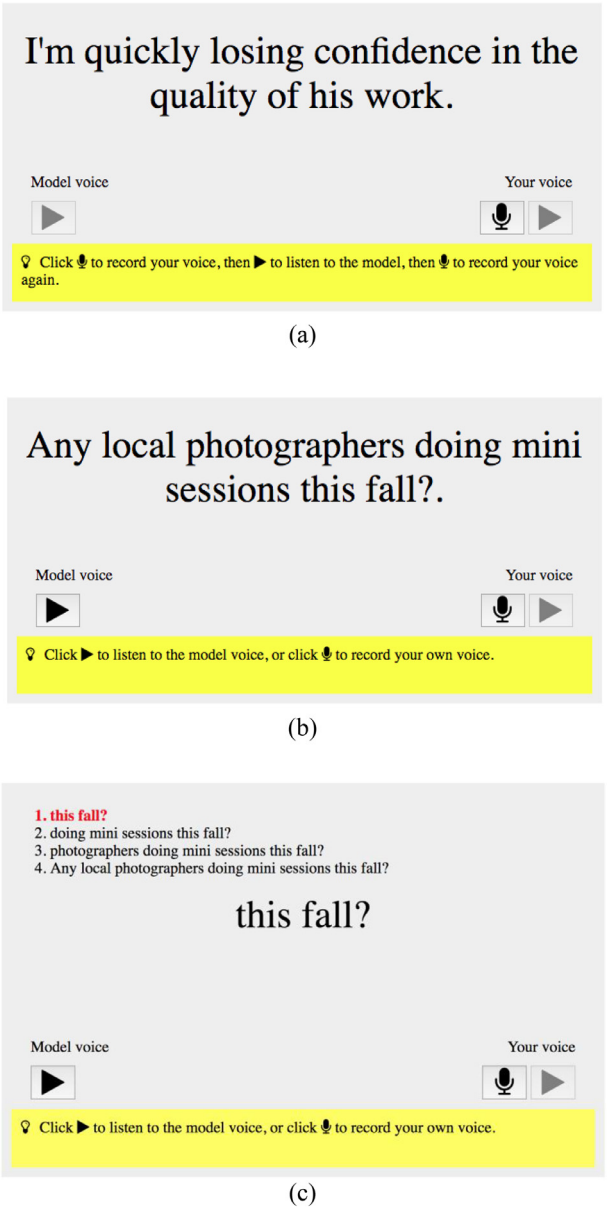

Fig. 5. Training interface using syntheses from the GSB tool: (a) say-listen-repeat exercise; (b) listen-repeat exercise; (c) backward build-up exercise.

[11] https://americanenglish.state.gov/materials-teaching-english.
[12] https://www.youtube.com/watch?v=TuNdTpjXkJ0.

Table 3
Statistical analysis of comprehensibility and fluency.

	Comprehensibility	Fluency
Effect of Time, $\chi^2(2)$	17.7 ($p < 0.001$)	27.8 ($p < 0.001$)
Pre-test score, mean [95% CI]	4.5 [4.2, 4.8]	3.2 [2.9, 3.6]
Immediate post-test ccore, mean [95% CI]	5.0 [4.7, 5.3]**	4.5 [4.2, 4.8]**
Delayed post-test score, mean [95% CI]	4.8 [4.5, 5.1]*	4.4 [4.1, 4.7]**

Post-hoc comparisons:
* $p < 0.1$.
** $p < 0.001$.

5.4. Data analysis

Inter-rater reliability was assessed using Cronbach's α (a correlation-based metric). In a reliable rating procedure, if one rater assigns a higher value to recording X than to recording Y, then other raters evaluating the same pair of recordings would be similarly expected to assign a higher value to recording X than to recording Y. If this is indeed the case, Cronbach's α would be high, and distributions of ratings can be reliably compared to answer RQ1.

Since, as it will be reported below, our dataset yielded high Cronbach's α, our main analyses proceeded to compare distributions of ratings across conditions. Analyses were based on fitting linear mixed-effects regression models to predict dependent variables (i.e. ratings of comprehensibility and fluency) based on the two factors: Training (trained vs. untrained sentences) and Time (pre-test, immediate post-test, and delayed post-test). After checking the normality assumption by running the Shapiro-Wilk test for each dependent variable, four nested models were fit to the data: (1) an intercept-only model; (2) a model adding a fixed effect for Time; (3) a model adding a fixed main effect for Training; and (4) a model adding an interaction between Time and Training. All models included random by-talker intercepts and random slopes for Training. Gains in goodness of fit of successive models were evaluated using chi-square tests. Fixed-effect parameters of the full model were used to estimate means of the dependent variables at different levels of Time and Training. Wald estimates of the confidence intervals (CIs) for means were then derived from the model.

5.5. Results: improvement of comprehensibility and fluency

5.5.1. Comprehensibility

A total of 8004 comprehensibility ratings were obtained from 50 listeners. Each recording was rated by an average of 22.3 listeners). Interrater reliability was assessed using correlations between ratings (Cronbach's $\alpha = 0.919$). High Cronbach's α indicates that distributions of ratings can be analyzed statistically to ascertain differences between conditions.

Comprehensibility ratings were normally distributed ($W = 1$, $p < 0.001$). Two nested linear mixed-effects regression models were fit to the data to predict the rating of comprehensibility: Model 1 was an intercept-only model, and Model 2 added a fixed effect for Time. Both models included random by-participant intercepts and random slopes for Time. Model 2 resulted in a significantly better fit to the data than Model 1: $\chi^2(2) = 17.7$, $p < 0.001$; see Table 3. This suggests that speakers' comprehensibility significantly changed over time. Estimated mean ratings and the 95% confidence intervals (CIs) were then derived from Model 2. At pre-test, the mean rating was 4.5 (CI: [4.2, 4.8]); at immediate post-test, 5.0 (CI: [4.7, 5.3]); at delayed post-test, 4.8 (CI: [4.5, 5.1]). Post-hoc pairwise comparisons of least-square means revealed a significant difference between pre-test and immediate post-test ($p < 0.001$). The difference between the immediate post-test and the delayed post-test was only marginally significant ($p = 0.069$), and so was the difference between the pre-test and the delayed post-test ($p = 0.059$). This suggests that participants did improve their comprehensibility from pre-test to immediate post-test, but we are unable to tell with certainty whether their gains were retained or partially lost by the time of the delayed post-test.

5.5.2. Fluency

To explore whether there was an improvement in fluency, a total of 6798 fluency ratings were obtained from 45 listeners. Each recording was rated by an average of 18.9 listeners. As with the measure of comprehensibility reported above, the high value of Cronbach's $\alpha = 0.963$ indicated that this measure was highly reliable.

Fluency ratings were normally distributed ($W = 1$, $p < 0.001$). Two nested linear mixed-effects regression models were fit to the data to predict the rating of fluency: Model 1 was an intercept-only model, and Model 2 added a fixed effect for Time. Both models incorporated a random by-participant intercept with a random slope for Time. Model 2 resulted in a significantly better fit to the data than Model 1: $\chi^2(2) = 27.8$, $p < 0.001$; see

Table 3. This suggests that fluency changed over time. Estimates of means and 95% CIs were derived from Model 2. At the pre-test, the mean rating of fluency was 3.2 (CI: [2.9, 3.6]); at the immediate post-test, 4.5 (CI: [4.2, 4.8]); and the delayed post-test, 4.4 (CI: [4.1, 4.7]). Post-hoc comparisons revealed that there was a significant difference between the pre-test and both post-tests ($p < 0.001$), while there was no difference between the immediate post-test and the delayed post-test ($p = 0.561$). Gains in fluency between the immediate post-test and the delayed post-test were retained.

To summarize, comprehensibility and fluency both were rated as improving from the pre-test. Trained sentences showed significant improvements in fluency from pre-test to post-test and maintained the improvement at the delayed post-test. Clearly, the training regimen, in which language learners practiced the trained sentences for three weeks, had an effect on how smoothly they were able to produce them.

5.6. Results: learners' GSB experience

To answer research question 2, "What features of the GSB did learners find useful or in need of improvement?", we interviewed learners following their immediate post- and delayed post-tests. Although both interviews included similar questions (Appendix B), delayed post-test interview included an additional question in which learners were asked to listen to two sentences from their pre- and post-test productions.

When learners were asked about the value of the pronunciation training and the ways they improved their speaking and pronunciation, they named several features. The feature that all learners except for one mentioned was fluency. Fourteen learners stated that GSB was helpful in making their speech sound more fluent and smoother. In fact, eight of these learners noticed how fluent they sounded after they listened to their pre- and post-test sentences during the delayed post-test interview. Learners' perceived improvement in fluency is also supported by our quantitative findings which showed a significant improvement between pre- and post-test. Learners (Excerpts 1 and 2) usually reported how 'choppy', 'cut' or 'slow' they sounded in their pre-test sentences whereas how 'quick' or 'smooth' they were in their post-test productions.

Excerpt 1:

Learner: actually this one is much more better than first.

Interviewer: okay, what is better about it?

Learner: this one, second one.

Interviewer: but what about it is better? What makes it better?

Learner: the first one is just uh how to say that, flow, ***the flow sounds like cut***.

Interviewer: okay ***so choppy***.

Learner: and the second one isn't, ***more better fluency***.

Excerpt 2:

Learner: uh, oh. I think ***my spoken English is more quick.***

Interviewer: more quick, okay.

Learner: yeah more quick and um I think ***my fluency is better***.

Connections between the words was something that some learners mentioned when they talked about fluency; they believed being able to connect words to each other instead of saying them one by one made their speech sound more smooth and more natural (See Excerpt 3). As a result, fluency and connected speech features were co-occurring topics learners touched on. Connected speech was something that some learners noticed clearly during their GSB training. They referred to the 'linking' between words and how they did not notice the connection between sounds before. They stated that they tried to use the GSB voice as a model to be able to produce the linking between words. One of the learners (Excerpt 4) said she knew about connected speech but she did not care about it until her practice with the GSB because she thought connected speech created a noticeable difference between her own pronunciation and that of the model voice. This awareness led her to care about something that she had not cared about before.

Excerpt 3:

Learner: so far more smooth and sounds more naturally.

Interviewer: Okay and anything else other than those?

Learner: mmm, I think just like I changed the way I speaked. Like well first before the training I said all words, speaking really clearly. And after the training like ***more connected and more smooth***.

Excerpt 4:

Interviewer: what are those things that you noticed with this model voice?

Learner: some something like when the ***words connected together very strongly.***

Interviewer: Okay so you have trouble with connected speech. Did you notice that before? Your, did you not know it before?

Learner: actually ***I didn't care about it before***. But I do care right now. After this,

Interviewer: why did it make you to care about it?

Learner: um, I think it's the ***big difference with my voice and model voice***.

Another pronunciation feature that was mentioned by most learners ($n = 12$) was intonation. Learners often stated how monotonous their speech was compared to the model voice and they did not have much 'ups and downs' or 'highs and lows' in their speech when they spoke English (Excerpt 5). Learners often explained the difference between their intonation and that of English by explaining how Korean works in general. They explained the change between 'high and low' as not something existing in Korean (Excerpt 6). When we asked learners if they would recommend practicing with the GSB to the others, one learner specifically commented on the benefit of hearing his own voice and how it helped with noticing the flow and intonation of the language: "*…it is a good opportunity to listen to your actual voice and then you can practice your pronunciation and you can actually be* ***aware of your voice or flows and intonation***".

Excerpt 5:

Interviewer: did you feel any changes during the training in your pronunciation? Anything you think you are doing better now?

Learner: oh I could some um realize that in terms of like um do question or some, so sometimes I need to ***tone down and tone up in terms of different question types***. That would be helpful to speak in English.

Interviewer: so you improved your intonation with those questions?

Learner: Mm-hmm. Yes I think so.

Excerpt 6:

Interviewer: okay, so how was yours different from the model voice?

Learner: um many ***Koreans pronunciation is not really high or low. just stable*** because Korean yeah, Korean language is kind of that. So um it was helpful to practice how to which part is good and ***what goes off and which part is goes down***.

Interviewer: Mm-hmm. So you started to think about those things?

Learner: Mm-hmm.

Learners also mentioned how GSB helped them notice the stress in individual words and sentences ($n = 6$). In addition, they mentioned how it helped with the improvement of certain sounds of English. However, the benefit of the GSB in improving segmentals was likely from practicing extensively for three weeks rather than hearing a voice similar to theirs. Extensive fluency practice may impact segmental improvement simply because of practice. Because the learners mostly talked about improvements in fluency and prosody, improvements in segmental quality may have been a side-effect of practice in general, and not connected to practicing with a golden speaker voice.

Three different exercise types were included in the design: say-listen-repeat, listen-repeat, and backward build-up exercises. Several learners ($n = 9$) stated their favorite exercise type was backward build-up because it gave them a chance to practice pronunciation in smaller chunks of speech. They could listen to the phrases in a sentence separately and this helped them in three ways: a) focus on parts they had more difficulties with, b) listen to words individually, c) focus on tones [i.e., intonation], and d) control the speed better (See Excerpt 7). One of the learners specifically mentioned the normal speed of sentences was too fast for him and backward build-up gave him the chance to practice things step by step, thus helping him with the flow of speech.

Excerpt 7:

Learner: Mm, I think all of them is great for practicing, but mmm, big words made the small words helpful.

Interviewer: okay, why?

Learner: Mm, all because the two the big words I could ***follow the speed***, and I understand how to ***pronounce the tones***.

Excerpt 8:

Learner: The difficult part was it was too fast. It was too fast to me and it's ***difficult to follow uh the full sentence***. And the easy part was, I don't know in the third practice, ***the step by step practice*** it was good to learn how to pronounce and how ***to make some flows***. Something like that.

In addition to the benefits for their pronunciation, most learners ($n = 10$) talked about the benefits of GSB for their listening skills—about how it helped them improve their listening or how it helped them listen critically and notice the problems in their pronunciation. Comments about listening improvement were similar to the comments about pronunciation in the sense that they performed better in hearing the connections between words or were better at catching up with the speed of speech. However, comments about listening critically showed how listening to a voice similar to one's own can help with perceiving the differences between one's self and the target pronunciation. One of the learners said "*I did not realize that there was a problem for me, but when I practicing it, I just realize that oh, model voice is correct and so yeah.*"

Learners in the study were also asked about further development of the GSB. One of the topics they commented on frequently was the voice quality. They suggested the voice quality could be improved. Some students stated that the model voice in the GSB was not very much like them and some others said there were parts of some sentences that the voice was not clear or very easy to understand. One learner said "*Uh it was good but one thing, um the models voice sometimes like vague. A little noise, so sometimes I can, I could not figure it out. The clear sounds from*

model voice." A similar comment from another learner was "*not clear sounds. So at the time I could not um figure out how to pronounce it like exactly because model voice sometimes very fast and sometimes vague*."

Another place for improvement lay in the design aspects of the GSB because some learners said having only three types of exercises or having a limited number of sentences to work with made their experience boring at times. Thus, adding more exercise types and sentences would be helpful. Another thing recommend by the learners was to be able to control the speed of speech because it was too fast for some learners and it made their effort to focus on pronunciation more challenging. Similar to that, learners also asked to practice individual words instead of only by phrases as in the backward build-up exercises. Suggestions about pronunciation improvement and support of visualization (such as including pictures and videos) were among the other recommendations for the improvement of the GSB.

6. Discussion

6.1. Analysis of the perceptual studies

The perceptual study indicates Golden Speaker Builder accomplished our goal of building a speaker voice that is suitable for self-imitation pronunciation training: the identity of the golden speaker voice (GS) is closer to that of the L2 learners than to the L1 source speakers, and has reasonable acoustic quality. Although the syntheses based on pitch transformation (Martin, 2004; Genevalogic 2006) achieved lower foreign accentedness and higher acoustic quality than GSB syntheses, pitch transformation failed to capture the L2 learners' identity, which is critical for self-imitation pronunciation training. Additionally, we found that a compounding factor in evaluating synthesis results is that of the rated acoustic quality. While GS had lower MOS than the original L1 speech, the original L2 recordings were also rated significantly lower (3.44 MOS). Since the L1 and L2 speakers were recorded under identical conditions, we suspect listeners regarded disfluencies and foreign accents in L2 speech as being of lower acoustic quality than native speech. Post-test feedback from some listeners supports this explanation: some were unsure if the low intelligibility was due to the speaker or to the overall low acoustic quality.

6.2. Analysis of the user study

In this study, we looked at the effectiveness of an interactive CAPT program on 15 Korean learners' improvement (as measured by ratings of comprehensibility and fluency) and what they thought about their learning experience with the program. Our study also explored if learning would be retained over a longer time period, as measured by a delayed post-test. The results showed a significant improvement in learners' comprehensibility and fluency for the trained sentences. Although ratings for both comprehensibility and fluency in the delayed post-test were slightly lower than the post-test, neither dropped to the level of the pre-test. Our qualitative findings especially supported the quantitative findings on fluency improvement because learners thought the GSB training was most helpful for their fluency.

The improvement in comprehensibility is similar to the controlled production results for Munro and Derwing (1998), who found that the comprehensibility of read-aloud speech improved after both segmentally-based and prosodically-based training. Their amount of practice was greater than in our study (12 weeks vs. 3 weeks) and the presence of a human instructor presumably allowed for more directed feedback than we provided. According to Isaacs and Trofimovich (2012), comprehensibility includes features related to discourse cohesion, grammar and vocabulary use, fluency, and pronunciation. Our study looked only at the results of fluency and pronunciation for their contributions to comprehensibility because the learners read sentences, and the grammar and vocabulary choices were made for them in the sentences. The only things that could improve were pronunciation and fluency

It seems clear from our results that implicit feedback, using only the model voice for computer-assisted recasts, may have limited the improvement. Calling learners' attention to particular sounds that may be problematic, or offering real-time mispronunciation feedback on specific portions of the speech signal, may help learners to make better use of a model voice. It is also possible that including visualizations of prosody, especially intonation, vowel lengthening, and juncture, would help learners to attend more carefully to features of pronunciation that are not noticed using implicit feedback. Hardison (2004), in training L2 learners to hear and produce French intonation, provided visual feedback. This directed feedback led to improvement in intonation and in untrained features.

When learners were asked for their opinions of their GSB experience, many learners reported how practice with the GSB helped them hear that their intonation and stress were different than the model voice and they believed they improved these features. Learners said the model voice allowed them to learn prosodic features of the language. While this is encouraging, it does not offer clear support for GSB; the use of any native-like voice prosody may have been equally or more effective. Because there was no control group, we cannot speak to this question.

One concern raised by learners was the speed of the model voice. It was initially too fast for many learners, even though it sounded like a normal speech rate for a native speaker. Fast speech can create problems for learners to catch the words and imitate speech. However, research shows that it does not necessarily mean that slower speech will lead to greater comprehensibility. It is more important to have a speech rate which is similar to a learner's, or just slightly faster, rather than a slower one (Probst et al., 2002; Munro and Derwing, 2001).

The only feedback learners received in the training was the synthesized version of their own voice, and we hypothesized it would help learners in perceiving their pronunciation problems and pronouncing in a more target-like way. Some learners said the GSB model voice did not sound quite like them; for others, learners said they did not hear all words clearly in some sentences, which could be due to either synthesis quality or speed. The voice quality issue is indeed not a new problem, as other studies also showed some distortions in parts of their synthesized speech (Sundström, 1998; Yoon, 2007). But there is a possibility that the synthesized speech, either in quality or speed, may have limited what learners could pay attention to.

6.3. Limitations

An important limitation in drawing conclusions from this study is that we did not have a control group to compare to the group which was trained with the GSB. In this case, a control group would be a synthesized voice that was created with two native voices so that both synthesized voices would be equally modified. Our plan is to include a control group for future iterations so that we see whether the voice created with the GSB or any voice model led to equal or better improvement.

A second limitation was our attention to only the sentence-level read aloud task. Our intention was to control for sentence type and rate all three weeks of the sentences. We do not know whether the sentences for Weeks 2 and 3 showed the same improvement. We also do not know if the training could have led to improvements in spontaneous speech where attention to discourse production, to vocabulary and grammar choices, and to fluency over longer stretches of speech would be more noticeable.

6.4. Future directions

Learners' suggestions about the GSB and our quantitative results show points to be taken into consideration for future iterations and design features that should be improved for the GSB tool. Changes that would improve the GSB experience regard both the quality of the golden

speakers and design issues with the learning interface that can lead to more improvement for learners.

First and foremost, the voice quality of the GSB tool must be addressed. It not only should match learners' voice quality more closely, it should also include multiple options for voice matching so that learners are more motivated to practice with it. This may increase the chances of improvement in segmentals and comprehensibility. Learners should also be able to control the speech rate, making it slower or faster depending on their needs. It is likely that learners will use the speed control to slow down and increase rate in practicing for different purposes. In addition, giving learners the ability to work on small chunks of speech through selection on a waveform would also allow them target a particular part of speech depending on their personal difficulties. The strong preference for the backward buildup task in this study indicates that learners both want to work on longer and shorter stretches of speech as they try to improve. Screen capturing technology would also help researchers see where learners perceive their difficulties to be.

In this study, all learners' voices were synthesized with the same native speaker's voice, thus learners were not given a chance to synthesize their speech with a native speaker of their choice. This was a practical decision because after recording multiple native voices, most voices demonstrated consistent levels of vocal fry (or creak) that ultimately limited their usefulness for synthesis. Giving learners the chance to choose a speaker for themselves may be helpful in terms of increasing their motivation; however, previous research shows that learners cannot always choose the speaker whose speech parameters are closest to themselves (Probst et al., 2002).

The GSB learning interface can also be developed more with different exercises types (such as directed perception tasks), feedback that highlights individual problems, learning aids such as brief explanations about how to work on pronunciation features, and guidance on what features are most important in a particular sentence. It would also be helpful to incorporate a directed perception test to help identify challenges before starting.

Conclusions

This study suggests that a CAPT program which utilizes feedback from a voice model can be helpful for the improvement of fluency (through attention to suprasegmental features of pronunciation) and for comprehensibility. Learners themselves reported an increase in their awareness for their use of intonation, stress, and connected speech in English. It may be that other types of feedback could be even more effective in promoting improvement.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Read-aloud sentences for Pre-, Post-, and Delayed Post-tests

1 I can't believe I gave up civilization for this.
2 If I was right again I still would not apologize.
3 The girls stared purposefully into each other's faces.
4 Who made you judge and jury? You're not in charge.
5 It's fairly clear to me that he didn't recognize it.
6 He thought he had seen it, but there was nothing on the rock.
7 My friend was actually talking about butterflies.
8 The singing voice approached rapidly, then faded away.
9 I'm quickly losing confidence in the quality of his work.
10 It was a temptation, but he resisted it for a while.
11 Without their friends, they wouldn't be acting so brave.
12 For a time the exciting thrill of his adventure was gone.
13 I'm looking forward to a week at the beach for vacation.
14 So, where do you want to eat lunch before English class?
15 Did you get to watch the football game last night?
16 How do I convey my emotions without emojis?
17 Any local photographers doing mini sessions this fall?
18 Do you mind if we stop by the post office on the way home?
19 It's been a real pleasure for the students to meet you.
20 We're out of food. Can you pick something up?
21 That sounds familiar! I know just how you feel.
22 You shouldn't have stayed up so late watching TV.
23 If anyone is into watches, check out my new web page.
24 Would you be able to help me find the secretary's office?
25 Each insult added to the value of the claim.
26 He was worth absolutely nothing to the world.
27 It seems strange for a zookeeper to think something like that."
28 We were met by powerful opposition when we made our plans public."
29 It was a curious coincidence, almost like someone planned it."
30 The fourth and fifth days passed without any developments."
31 After the car crash, his face was streaming with blood."
32 I discovered that the promise was unexpectedly fulfilled."
33 She spoke with genuine sympathy in her face and voice."
34 He obeyed the pressure of her hand, and changed directions."
35 Every bone in her aged body seemed broken or dislocated.
36 He began to follow the footprints of the dog."
37 You should try something new, what do you have to lose?
38 Why don't you call for a reservation while I change my shoes?
39 I bought a bunch of vegetables at the farmer's market.
40 Blaze has the best veggie pizzas. Just thought I'd share.
41 What casual restaurants in town have free Wi-Fi?
42 It's hard to learn a foreign language as you get older.
43 I just bought a ticket to New York for Thanksgiving.
44 Sorry, my phone has a terrible signal here.
45 My favorite hobbies are photography and folk dancing.
46 What time does the bus leave for the airport?
47 Check out our page. We offer free estimates and low rates.
48 There's a schedule change tomorrow because of the flood.

Appendix B. Post-test and Delayed post-test Interview Questions

Post-Test Interview

1 In what ways was the pronunciation training valuable to you? In what ways do you feel you have improved?
2 What was it like practicing with the golden speaker model?
3 How long and how often did you practice?
4 Was the visual feedback helpful?
5 Do you feel like your ability to listen to English speech has improved?
6 Do you feel like your pronunciation has improved? In what ways?
7 Which types of pronunciation were the most difficult to improve?
8 Did you notice any other pronunciation or language items that you had difficulty with during your practice? What were they?
9 What was difficult about practicing with the "golden speaker"?
10 What kind of suggestions would you give for trying this in the future?
11 What did you notice when you were practicing?
12 Was it easy to repeat the sentences at the same speed?
13 Was it easy to get the consonant sounds correctly?
14 Was it easy to get the vowel sound correctly?
15 What kinds of things did you pay most attention to?
16 What kind of thins did you practice most and why?
17 How do you like the interface of the "Golden Speaker"?
18 How easy was it to use the website to practice?
19 How comfortable were you using the website?
20 Did you have any technical problems?
21 Would you recommend that others try out the golden speaker builder?

Delayed Post-test Interview

1 Since finishing the training, in what ways was the pronunciation training continued to be valuable to you?
2 Have you continued to use the training materials?
3 Has the training affected how you approach your English pronunciation?
4 Do you feel like your ability to listen to English speech has improved?
5 Do you feel like your pronunciation has improved? In what ways?
6 Which types of pronunciation continue to be difficult to improve?
7 Have you noticed any other pronunciation or language items that have been difficult after your practice? What were they?
8 What things would you suggest for more effective practice?
9 What kind of suggestions would you give for trying this in the future?
10 What features do you most remember about practicing – consonants, vowels or other features of speech?
11 Was it helpful to have someone helping you to practice?
12 What kinds of things do you remember paying attention to?
13 Are there any things you have tried to change in your own speech since the training?
14 Would you recommend that others try out the golden speaker builder?

References

Aryal, S., Gutierrez-Osuna, R., 2015. Reduction of non-native accents through statistical parametric articulatory synthesis. J. Acoust. Soc. Am. 137 (1), 433–446.
Aryal, S., Felps, D., Gutierrez-Osuna, R., 2013. Foreign accent conversion through voice morphing. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Lyon, France.
Barriuso, T.A., Hayes-Harb, R., 2018. High variability phonetic training as a bridge from research to practice. CATESOL J. 30 (1), 177–194.
Bissiri, M.P., Pfitzinger, H.R., 2009. Italian speakers learn lexical stress of German morphologically complex words. Speech Commun. 51 (10), 933–947.
Bissiri, M.P., Pfitzinger, H.R., Tillmann, H.G., 2006. Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis. Proceedings of the 11th Australian International Conference on Speech Science & Technology. University of Auckland, New Zealand.
Bliss, H., Bird, S., Cooper, P.A., Burton, S., and Gick, B., "Seeing speech: ultrasound-based multimedia resources for pronunciation learning in Indigenous Languages," 2018.
Buchholz, S., Latorre, J., 2011. Crowdsourcing Preference tests, and How to Detect Cheating. INTERSPEECH.
Burgess, J., Spencer, S., 2000. Phonology and pronunciation in integrated language teaching and teacher education. System 28 (2), 191–215.
Cho, S.-.M., 2004. An acoustic study of the pronunciation of Korean vowels uttered by Japanese speakers. Speech Sci. 11.
Couper, G., 2017. Teacher cognition of pronunciation teaching: teachers' Concerns and Issues. TESOL Q. 51 (4), 820–843.
Cucchiarini, C., Strik, H., Boves, L., 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. J. Acoust. Soc. Am. 107 (2), 989–999.
De Meo, A., Vitale, M., Pettorino, M., Cutugno, F., Origlia, A., 2012. Imitation/self-imitation in computer-assisted prosody training for Chinese learners of L2 Italian. In: Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference.
Derwing, T.M., Munro, M.J., 1997. Accent, intelligibility, and comprehensibility: evidence from four L1s. Stud. Second Lang. Acquis. 19 (1), 1–16.
Derwing, T.M., Munro, M.J., 2015. Pronunciation Fundamentals: Evidence-Based Perspectives For L2 Teaching and Research. John Benjamins Publishing Company.
Derwing, T.M., Rossiter, M.J., 2003. The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. Appl. Lang. Learn. 13 (1), 1–17.
Derwing, T.M., Munro, M.J., Wiebe, G., 1998. Evidence in favor of a broad framework for pronunciation instruction. Lang. Learn. 48 (3), 393–410.
Derwing, T.M., Thomson, R.I., Munro, M.J., 2006. English pronunciation and fluency development in Mandarin and Slavic speakers. System 34 (2), 183–193.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32 (2), 407–499.
Egan, K., LaRocca, S., 2000. Speech recognition in language learning: a must. In: Proceedings of the InSTILL 2000, pp. 4–9.
Ejzenberg, R., 2000. The juggling act of oral fluency: a psycho-sociolinguistic metaphor. In: Perspectives on Fluency. University of Michigan, pp. 287–313.
Erro, D., Moreno, A., Bonafonte, A., 2010. Voice conversion based on weighted frequency warping. IEEE Trans. Audio Speech Lang. Process. 18 (5), 922–931.
Eskenazi, M., 1999. Using a computer in foreign language pronunciation training: what advantages? Calico J. 447–469.
Eskenazi, M., 2009. An overview of spoken language technology for education. Speech Commun. 51 (10), 832–844.
Felps, D., Bortfeld, H., Gutierrez-Osuna, R., 2009. Foreign accent conversion in computer assisted pronunciation training. Speech Commun. 51 (10), 920–932.
Felps, D., Geng, C., Gutierrez-Osuna, R., 2012. Foreign accent conversion through concatenative synthesis in the articulatory domain. IEEE Trans. Audio Speech Lang. Process 20 (8), 2301–2312.
Fillmore, C.J., 1979. On fluency. In: Individual Differences in Language Ability and Language Behavior. Elsevier, pp. 85–101.
Gass, S.M., Mackey, A., Pica, T., 1998. The role of input and interaction in second language acquisition introduction to the special issue. Modern Lang. J. 82 (3), 299–307.
Genevalogic. (2006). *SpeedLingua*. Available: http://home.speedlingua.com.
Gordon, J., Darcy, I., 2016. The development of comprehensible speech in L2 learners. J. Second Lang. Pronunc. 2 (1), 56–92.
Hardison, D.M., "Generalization of computer assisted prosody training: quantitative and qualitative findings," 2004.
Heift, T., 2004. Corrective feedback and learner uptake in call. ReCALL 16 (2), 416–431.
Hincks, R., 2003. Speech technologies for pronunciation feedback and evaluation. ReCALL 15 (1), 3–20.
Hirose, K., Gendrin, F., Minematsu, N., 2003. A pronunciation training system for Japanese lexical accents with corrective feedback in learner's voice. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).
Huckvale, M., Yanagisawa, K., 2007. Spoken language conversion with accent morphing. Proceedings of ISCA Speech Synthesis Workshop.
Isaacs, T., Trofimovich, P., 2012. Deconstructing comprehensibility: identifying the linguistic influences on listeners' L2 comprehensibility ratings. Stud. Second Lang. Acquis. 34 (3), 475–505.
Jun, S.A., 1995. A phonetic study of stress in Korean. J. Acoust. Soc. Am. 98 (5) 2893.
Kanters, S., Cucchiarini, C., Strik, H., 2009. The goodness of pronunciation algorithm: a detailed performance study. In: Proc. Speech and Language Technology in Education (SLaTE 2009).
Kawahara, H., 2006. STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds. Acoust. Sci. Technol. 27 (6), 349–353.
Kominek, J., Black, A.W., 2004. The cmu arctic speech databases. Proceedings of the Fifth ISCA Workshop on Speech Synthesis.
Kominek, J., Black, A.W., Ver, V., 2003. CMU ARCTIC Databases For Speech Synthesis. Proceedings of the Fifth ISCA ITRW on Speech Synthesis. Pittsburgh, PA.
Kormos, J., Dénes, M., 2004. Exploring measures and perceptions of fluency in the speech of second language learners. System 32 (2), 145–164.
Kreiman, J., Papcun, G., 1991. Comparing discrimination and recognition of unfamiliar voices. Speech Commun. 10 (3), 265–275.
Lee, J., Jang, J., Plonsky, L., 2014. The effectiveness of second language pronunciation instruction: a meta-analysis. Appl. Linguist. 36 (3), 345–366.
Lee, K.-.N., 2001. Automatic generation of pronunciation variants for Korean continuous speech recognition. J. Acoust. Soc. Korea 20 (2), 35–43.
Lennon, P., 1990. Investigating fluency in EFL: a quantitative approach. Lang Learn 40 (3), 387–417.
Levis, J., Sonsaat, S., 2017. Pronunciation in the clt era. In: The Routledge handbook of English Pronunciation, pp. 267–283.
Levis, J., 2007. Computer technology in teaching and researching pronunciation. Annu. Rev. Appl. Linguist. 27, 184–202.
Levis, J., 2018. Intelligibility, Oral communication, and the Teaching of Pronunciation. Cambridge University Press, Cambridge: Cambridge.
Liberatore, C., Aryal, S., Wang, Z., Polsley, S., Gutierrez-Osuna, R., 2015. SABR: sparse, anchor-based representation of the speech signal. In: Proceedings of the INTERSPEECH, pp. 608–612.
Liberatore, C., Zhao, G., Gutierrez-Osuna, R., 2018. Voice conversion through residual warping in a sparse, anchor-based representation of speech. In: Proceedings of the IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP).
Lyster, R., 1998. Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. Lang. Learn. 48 (2), 183–218.
Lyster, R., 2001. Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. Lang. Learn. 51, 265–301.
MacDonald, S., 2002. Pronunciation-views and practices of reluctant teachers. Prospect 17 (3).
Mackey, A., Abbuhl, R., 2005. Input and interaction. In: Mind and Context in Adult Second Language Acquisition: Methods, Theory and Practice, pp. 207–233.
Mak, B., et al., 2003. PLASER: pronunciation learning via automatic speech recognition. In: Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing-Volume 2. Association for Computational Linguistics, pp. 23–29.
Martin, P., 2004. WinPitch LTL II, a multimodal pronunciation software. In: Proceedings of the STIL/ICALL Symposium 2004.
McAuliffe, M., Socolof, Michaela, Mihuc, Sarah, Wagner, Michael, Sonderegger, Morgan, 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. presented at the Interspeech.
Mohammadi, S.H., Kain, A., 2017. An overview of voice conversion systems. Speech Commun. 88, 65–82 04/01/ 2017.
Munro, M.J., Derwing, T.M., 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Lang. Learn. 45 (1), 73–97.
Munro, M.J., Derwing, T.M., 1998. The effects of speaking rate on listener evaluations of native and foreign-accented speech. Lang. Learn. 48 (2), 159–182.
Munro, M.J., Derwing, T.M., 2001. Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. Stud. Second Lang. Acquis. 23 (4), 451–468.

Nagano, K., Ozawa, K. , 1990a. English speech training using voice conversion. In: Proceedings of the First International Conference on Spoken Language Processing.
Nagano, K., Ozawa, K. , 1990b. English speech training using voice conversion. In: Proceedings of the First International Conference on Spoken Language Processing (ICSLP). Kobe, Japan.
Nakatani, C.H., Hirschberg, J., 1994. A corpus-based study of repair cues in spontaneous speech. J. Acoust. Soc. Am. 95 (3), 1603–1616.
Nicholas, H., Lightbown, P.M., Spada, N., 2001. Recasts as feedback to language learners. Lang. Learn. 51 (4), 719–758.
O'brien, I., Segalowitz, N., Freed, B., Collentine, J., 2007. Phonological memory predicts second language oral fluency gains in adults. Stud. Second Lang. Acquis. 29 (4), 557–581.
Panchapagesan, S., Alwan, A., 2009. Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC. Comput. Speech Lang. 23 (1), 42–64.
Peabody, M., Seneff, S., 2006. Towards automatic tone correction in non-native mandarin. Chin. Spok. Lang. Process. 602–613.
Pelham, B.W., Blanton, H., 2012. Conducting Research in psychology: Measuring the Weight of Smoke. Cengage Learning.
Pellegrino, E., Vigliano, D., 2015. Self-imitation in prosody training: a study on Japanese learners of Italian. In: Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE). Leipzig, Germany.
Pitz, M., Ney, H., 2005. Vocal tract normalization equals linear transformation in cepstral space. IEEE Trans. Speech Audio Process. 13 (5), 930–944.
Probst, K., Ke, Y., Eskenazi, M., 2002. Enhancing foreign language tutors–in search of the golden speaker. Speech Commun. 37 (3–4), 161–173.
Riggenbach, H., 1991. Toward an understanding of fluency: a microanalysis of nonnative speaker conversations. Discourse Process. 14 (4), 423–441.
Riggenbach, H., 2000. Perspectives on Fluency. University of Michigan Press/ESL.
Rypa, M.E., Price, P., 1999. VILTS: a tale of two technologies. CALICO J. 385–404.
Saito, K., 2012. Effects of instruction on L2 pronunciation development: a synthesis of 15 quasi-experimental intervention studies. TESOL Q. 46 (4), 842–854.
Segalowitz, N., 2007. Access fluidity, attention control, and the acquisition of fluency in a second language. TESOL Q. 41 (1), 181–186.
Solem, A., "Celery: distributed Task Queue," 4.0.0 ed: Celery, 2016, http://www.celeryproject.org/.
Sundström, A., 1998. Automatic prosody modification as a means for foreign language pronunciation training. In: Proceedings of the ISCA Workshop on Speech Technology in Language Learning (STILL 98), Marholmen, Sweden, pp. 49–52.
Swain, M., Lapkin, S., 1995. Problems in output and the cognitive processes they generate: a step towards second language learning. Appl. Linguist. 16 (3), 371–391.
Swain, M., 2000. The output hypothesis and beyond: mediating acquisition through collaborative dialogue. Sociocultural Theory and Second Language Learning 97, 114.
Thomson, R.I., Derwing, T.M., 2014. The effectiveness of L2 pronunciation instruction: a narrative review. Appl. Linguist. 36 (3), 326–344.
Thomson, R.I., 2011. Computer assisted pronunciation training: targeting second language vowel perception improves pronunciation. Calico J. 28 (3), 744.
Thomson, R.I., 2012. Improving L2 listeners' perception of English vowels: a computer-mediated approach. Lang. Learn. 62 (4), 1231–1258.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. 267–288 Series B.
Tyler, A., 1992. Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. TESOL Q. 26 (4), 713–729.
Wang, R., Lu, J., 2011. Investigation of golden speakers for second language learners from imitation preference perspective by voice modification. Speech Commun. 53 (2), 175–184 February 2011.
Warren, P., Elgort, I., Crabbe, D., 2009. Comprehensibility and prosody ratings for pronunciation software development. Language Learning & Technology 13 (3).
Wennerstrom, A., 2000. The role of intonation in second language fluency. In: Perspectives on Fluency. University of Michigan, pp. 102–127.
Yan, Q., Vaseghi, S., Rentzos, D., Ching-Hsiang, H., 2007. Analysis and synthesis of formant spaces of british, australian, and american accents. IEEE Trans. Audio Speech Lang. Process. 15 (2), 676–689.
Yoon, K., 2007. Imposing native speakers' prosody on non-native speakers' utterances. 현대영미어문학 25 (4), 197–215.