Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2024, Volume 28, Issue 1
pp. 1–16

ARTICLE

Corrective feedback accuracy and pronunciation improvement: Feedback that is 'good enough'

Alif Silpachai, Radboud University
Reza Neiriz, Iowa State University
MacKenzie Novotny, Iowa State University
Ricardo Gutierrez-Osuna, Texas A&M University
John M. Levis, Iowa State University
Evgeny Chukharev, Iowa State University

Abstract

It is unclear whether corrective feedback (CF) provided by L2 computer-assisted pronunciation training (CAPT) tools must be 100% accurate to promote an acceptable level of improvement in pronunciation. Using a web-based interface, 30 native speakers of Chinese completed a pretest, a computer-based training session to produce nine sound contrasts in English, and a posttest. The study manipulated feedback accuracy using a modified "Wizard of Oz" protocol in which a phonetically-trained human listener in a separate room provided CF on the trainees' productions, but the trainees thought that the computer-based system provided the CF. The computer system presented a set of three sound contrasts with 100% accuracy, three with 66% accuracy (with one of three human responses changed randomly), and three with 33% accuracy (with two of three human feedback responses being changed). The trainees' pre- and posttest productions were rated for accuracy by native speakers of English. For trained items, productions were not significantly different when the trainees received CF with 100% or 66% accuracy, but both resulted in greater improvement than feedback with 33% accuracy. An important implication for L2 pronunciation training software is that machine feedback can be beneficial even when it is 'good enough' (i.e., not 100% accurate).

Keywords: *Corrective Feedback, Second Language Pronunciation, English as a Second Language, CAPT*

Language(s) Learned in This Study: *English*

APA Citation: Silpachai, A., Neiriz, R., Novotny, M., Gutierrez-Osuna, R., Levis, J. M., & Chukharev, E. (2024). Corrective feedback accuracy and pronunciation improvement: Feedback that is 'good enough'. *Language Learning & Technology*, *28*(1), 1–16. https://hdl.handle.net/10125/73582

Introduction

Explicit teaching of second language (L2) pronunciation is almost always successful (J. Lee et al., 2015) in improving the accuracy of L2 learners' speech in controlled speaking contexts (Saito & Plonsky, 2019). There is less robust but still promising evidence that human- or computer-instructed pronunciation teaching improves intelligibility and comprehensibility in spontaneous speech (Thomson & Derwing, 2015), especially with the teaching of suprasegmentals (Derwing et al., 1998; Gordon & Darcy, 2016; Zhang & Yuan, 2020). A key factor in such improvement is corrective feedback (CF).

One goal of CF is to help learners reduce their pronunciation errors. In L2 learning, lack of CF may result in learners producing utterances in which "their interlanguage representations become automatized routines" (A. H. Lee & Lyster, 2017, p. 374). This means that learners' interlanguage pronunciation errors may not be actively adjusted when CF is absent. Consequently, learners may not notice errors as they focus

on expressing meaning.

CF may also serve to lower learners' anxiety when speaking an L2, which might indirectly help them improve their L2 pronunciation. E. J. Lee (2016) observed that when learners received CF from teachers, they had reduced anxiety when speaking English. The author implied that reduced anxiety might have helped some learners pay attention to and process the CF.

There is strong evidence that providing understandable and accurate feedback is an important factor in improvement (Martin & Sippel, 2021; Saito & Lyster, 2012). Such feedback can come from teachers (Dlaska & Krekeler, 2013), from other learners (Dai & Wu, 2021; Martin & Sippel, 2021), from computers providing model voices (Ding et al., 2019), or through computerized mispronunciation detection (MPD) or automatic speech recognition (ASR) algorithms (McCrocklin, 2019). Despite major advances in MPD algorithms, it is unclear what levels of CF accuracy such systems must achieve in order to improve learners' pronunciation. The goal of this paper is to address this gap. Our study is important because 100% accuracy may not be necessary, since L2 learners are generally not fully accurate when judging their own goodness of pronunciation, either in perception or production, hence the notion of providing CF that is 'good enough.'

Background

Studies have shown that CF can help L2 learners improve their pronunciation and comprehensibility. In particular, CF that specifies the error location and provides a correction may be especially effective (see Table 1 of Lyster et al., 2013, p. 4, for more types of CF) by helping learners "develop self-awareness of their pronunciation difficulties ... when these difficulties occur so that they can learn to self-correct or self-monitor" (Darcy, 2018, p. 29). For example, in a study of human-directed pronunciation improvement, Saito and Lyster (2012) reported that native speakers of Japanese improved their pronunciation of English /ɹ/ (as in *right*) after four days of working with an instructor who provided recasts (error-free restatements of a word containing a pronunciation error committed by a learner) in form-focused instruction.

Although human listeners are the gold standard in pronunciation evaluation (Derwing & Munro, 2015), human CF may not always be optimal. Attention distraction, fatigue, mood, cognitive biases, and simple inconsistency in attending to the same features of speech, as well as dozens of other issues, can lead to variations in the quality of human CF. A potential solution is to use computer-based tools, as they can provide consistent CF endlessly. In a CAPT (Computer Assisted Pronunciation Teaching) study of CF on intonation to L2 learners of French, Hardison (2004, p. 48) argued that consistent and repetitive CF had advantages over human feedback because the same CF provided by humans would have required many hours of instructor time. Hirata (2004) obtained similar results by using fundamental frequency feedback for Japanese pitch and length contrasts.

Most types of computer-based CF have focused on segmental errors and using the output of an ASR system (i.e., a written representation of what was said) as CF to improve the pronunciation of sounds (Ellis, 2006; García et al., 2020; McCrocklin, 2019). However, different approaches have been used to provide CF. Sung (2008) observed that Korean learners of English improved their pronunciation after two weeks of training using a software tool (Dr. Speaking) that had been developed specifically for native Korean speakers. The software provided CF on segmental errors through ASR and spectrogram comparisons with a model utterance. Unlike Sung (2008), García et al. (2020) reported that English learners of Spanish improved their pronunciation after using a web application (iSpraak) for 15 weeks. The application provided CF in the form of an accuracy score (0–100%) by comparing an ASR transcription of the learner's utterance against the instructor-provided text.

Nevertheless, MPD and ASR tools are unlikely to achieve 100% accuracy (Derwing et al., 2000; McCrocklin & Edalatishams, 2020). Accuracy is always greater for L1 than L2 speech, even when L2 speech is intelligible to human listeners. In Derwing et al. (2000), L2 speech was recognized only 70% of the time by a commercial version of Dragon Naturally Speaking, while in McCrocklin and Edalatishams' (2020) analysis of Google Voice, L2 speech was recognized with approximately 90% accuracy. Malakar and Keskar (2021) compared the accuracy of 63 phoneme-recognition systems. When considering

recognition at the phoneme level (which is more challenging than at the word level), accuracies range from 63% to 95%, averaging about 75%.

Previous studies have also indicated that CF does not need to be 100% accurate to promote improvement (Bashori et al., 2022; McCrocklin, 2016; Mroz, 2018). Instead, a 'good enough' level of accuracy may be useful and sufficient in a wide variety of language learning environments, such as those that lack pronunciation teachers or those in which learners are self-taught (so-called "informal language learning contexts"; see Dressman & Sadler, 2020).

This Study

This exploratory study used a single-session training experiment to estimate how different levels of accuracy in CF affected improvements in L2 pronunciation. In particular, we compared systems with three levels of CF accuracy: 100% (i.e., the "perfect" system, all else being equal), 66%, and 33% accuracy (henceforth Systems 100, 66, and 33, respectively). We predicted that learners' pronunciation improvements would be greatest for System 100 and lowest for System 33.

Research Questions

Our study had two research questions. The primary one asked whether increased levels of CF accuracy would increase pronunciation improvements. We explored this question using a modified "Wizard of Oz" protocol (see Browne, 2019), based on *The Wizard of Oz* (Fleming, 1939), in which a trained human listener (i.e., the wizard) provided the feedback, which we treated as ground truth (i.e., 100% accurate). The human feedback was then computationally modified to create three systems: System 100 (in which no changes were made to the feedback), System 66 (in which one of three human responses was changed to be inaccurate), and System 33 (in which two of three human responses were modified to be inaccurate). Our secondary question was whether learning on trained items would transfer to untrained items. For this reason, the experimental protocol also included three untrained items for each sound pair involved in the training to determine whether the participants would perform similarly on the trained and untrained items.

Methodology

Participants

Thirty native speakers of Chinese, international students at a large midwestern U.S. university, participated in this study for a small payment in accordance with IRB protocols at the host university. The average age of the participants was 27.6 (SD = 4.48) and ranged from 22 to 37 years old. Of the 30 participants, 28 spoke Mandarin Chinese natively, one spoke Cantonese natively, and one spoke both languages. Of the 28 participants who spoke Mandarin, three specified their first dialects as the Chongqing, Sichuan, and Henan dialects, respectively.

Stimuli

Stimuli included nine sound contrasts chosen because they are often difficult for Chinese speakers to produce, as determined by some of the authors' experiences in English as a Second Language (ESL) classrooms as well as previous research (Qian, 2018). The nine sound contrasts were divided into three groups, three sound contrasts for each training set. Each set included a mixture of consonant and/or vowel contrasts but was not organized further as they did not differ in difficulty or learnability. In addition, each set was presented using Systems 33, 66, and 100 for different participants, which resulted in counterbalancing all training sets (Table 2). Because we were examining the effects of system accuracy on pronunciation improvement, the functional load (Brown, 1988) of each contrast, or its likelihood to influence comprehensibility, was not considered in how the sound contrasts were distributed.

Each contrast was contextualized in eight short phrases, for a total of 72 phrases. The phrases were first created by one of the authors, who has decades of experience in developing pronunciation materials, and then were revised in consultation with the remaining authors. These contrasts and phrases are presented in

Table 1, with phrases that were presented for feedback in normal print and those that appeared only during testing phases in italics.

Table 1

Phrases Containing Target Sound Contrasts and the 72 Phrases Used in the Experiment

Sound Contrast	Phrases Used for the Study
Group 1	
/w/ - /v/ (Set 1)	very wild, well versed, valuable wine, every week, driving west, *vanishing wealth*, *withering vines*, *virtual wonders*
/i/ - /ɪ/ (Set 2)	thirteen wishes, Lisa's lips, simply easy, six peaches, thin-crust pizza, *a green ship*, *fourteen fish*, *sweet kisses*
/æ/ - /ɛ/ (Set 3)	the best laughs, a test track, bland bread, the saddest men, never passive, *a fast ending*, *pack the rest*, *better sand*
Group 2	
/ŋ/ - /n/ (Set 4)	cutting the lawn, a long pen, the phone rang, sing it again, the gun goes bang, *wrong question*, *wings or fins*, *ten rings*
/s/ - /θ/ (Set 5)	seven birthdays, thirteen signs, thirty some years, bath soaps, wealthy sons, *some thanks*, *false faith*, *a slow month*
/z/ - /s/ (Set 6)	lose the list, seven or zero, sign for the zoo, the rising sun, raise your voice, *zebra crossing*, *a super magazine*, *a crazy person*
Group 3	
/ɑ/ - /ʌ/ (Set 7)	a hot tub, lots of luck, not a duck, funny socks, stuck in the car, *a couple of dollars*, *our monthly talk*, *a big pot of money*
/u/ - /ʊ/ (Set 8)	they should be stewed, useful wool clothes, took it sooner, loosely understood, could be the truth, *mushy and gooey*, *pushed him through*, *a blue cushion*
/ð/ - /d/ (Set 9)	smooth dishes, those diamonds, that lion's den, don't bathe for an hour, dance clothing, *dismal weather*, *Dana's father*, *breathe more deeply*

Note. Of the 72 phrases, the 27 ones in italics (three per set) are transfer items.

Procedures

We developed a web-based interface to present stimuli, record participants, and display CF. The interface presented the stimuli and tasks in three stages: pretest, training, and posttest. Each stage started with audio and written instructions, as well as example items related to that stage. Participants were seated comfortably in a sound-attenuated room. A Blue Yeti microphone was used to record their productions. The recording process took between 60 and 90 minutes. Participants were monitored by a research assistant through a remote desktop application to ensure that the data collection ran smoothly.

During the pretest stage, participants were first asked to complete a task that familiarized them with the procedure using four practice stimuli that did not target experimental sound contrasts. After the familiarization, participants began the pretest. Each of the 72 phrases was presented individually to each participant to read aloud and record using the web interface. Participants were allowed to listen to their recordings as many times as they wished before moving on to the next phrase. This was done so they would follow the same procedures for testing and training, and to allow them to check recording quality. Once they submitted each phrase, an audio file was saved to the server in WAV format.

Participants were informed that they would receive CF from three pronunciation feedback systems, all of which looked the same but with different levels of accuracy. They did not receive any information specifying the level of accuracy for each system. The order of the systems was counterbalanced in how they were presented (Table 2). After training with each system (after each set of three contrasts), participants rated the system using a 10-point Likert scale that indicated how accurate they thought the system was in evaluating their pronunciation. We included this rating to determine the degree to which participants were aware of differences in feedback accuracy.

Table 2

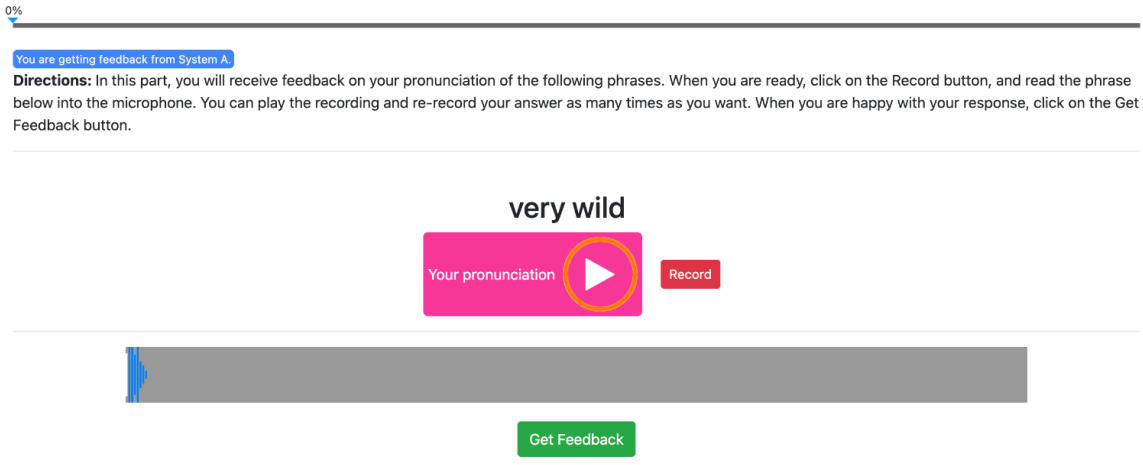
Counterbalancing Measures for System Accuracies and Sound Contrast Sets

Sound Contrasts	System 100	System 66	System 33
Set 1 (contrast 1–3)	Subjects 1–10	Subjects 11–20	Subjects 21–30
Set 2 (contrast 4–6)	Subjects 21–30	Subjects 1–10	Subjects 11–20
Set 3 (contrast 7–9)	Subjects 11–20	Subjects 21–30	Subjects 1–10

During the training stage, participants were again familiarized with the procedure using four practice stimuli that did not target experimental sound contrasts. For all participants, the procedure of the familiarization tasks was identical to the procedure of the training system they were first exposed to (e.g., if a participant was first exposed to System 100, the feedback of the familiarization task was the same as that received through training on System 100; if the first system presented was System 66, feedback in the familiarization task also followed System 66's accuracy for CF). After the familiarization tasks, participants began training. The training consisted of 45 phrases from Table 1 (the first five phrases for each sound contrast), with the remaining 27 phrases in Table 1 (the final three phrases for each sound contrast, listed in italics) being used as transfer items. Participants were shown one phrase at a time and were asked to record it and submit it for CF, as illustrated in Figure 1. As done in the pretest stage, they were allowed to record the phrase as many times as they wished before submitting it for CF.

Figure 1

Illustration of the Web Interface After a Participant Recorded a Production of the Phrase "Very Wild"



Once participants submitted the recording for each of the training phrases using the interface, they continued to the feedback phase. They produced a training phrase (as in Figure 1), and this recording was presented to the wizard, who provided CF to the participant on the two target phonemes, as illustrated in

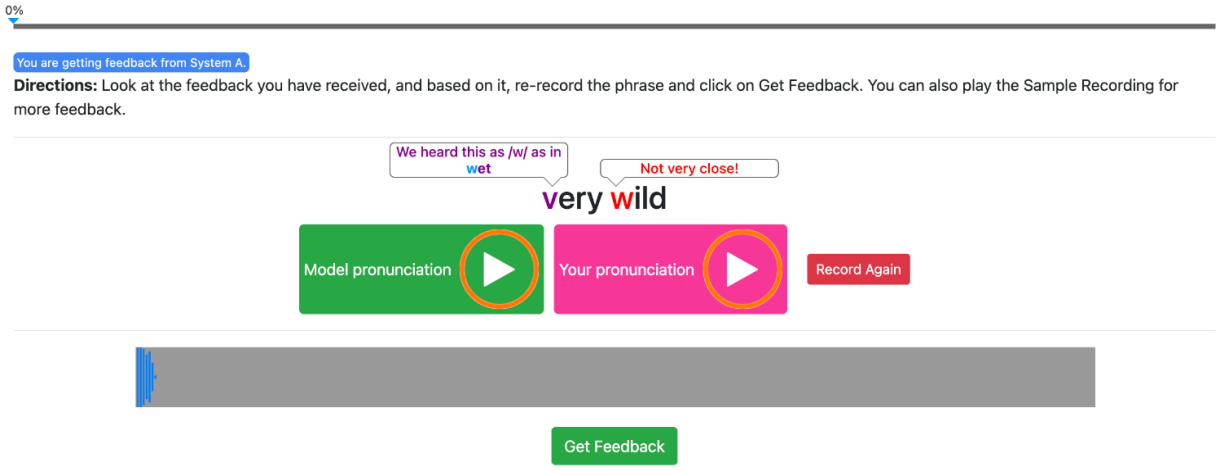
Figure 2. The CF was limited to three options:

1. If a sound had been produced correctly, the orthographic representation of the sound was highlighted, and a floating bubble indicated that the pronunciation was correct.
2. If a sound corresponded to the other phoneme in the pair, a floating bubble indicated they had made this mistake (e.g., "We heard this as /w/ as in *wet*").
3. If a sound did not match either of the target phonemes, a floating bubble indicated that their pronunciation did not match either target (e.g., "Not very close!").

All training items were presented twice during training. If the item was produced incorrectly, it was presented again immediately. If the item was produced correctly, it was presented again at a later time.

Figure 2

Illustration of the Web Interface After a Trainee Has Received CF

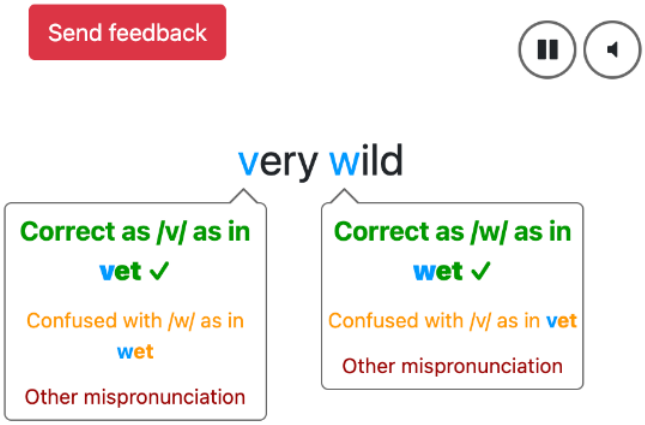


CF was provided using the "Wizard of Oz" protocol in the human-computer-interaction literature (see Browne, 2019), in which participants believed the CF was produced by an automatic system, but it was actually provided by "the wizard" who was a human listener with advanced training in phonetics and pronunciation teaching. The CF levels of accuracy, 100%, 66%, and 33%, were accomplished by leaving the wizard's feedback untouched (System 100) or by computationally adjusting the CF accuracy from the human listener, resulting in Systems 66 and 33. Each level of feedback accuracy was applied to one training set of three sound contrasts in a random order (as shown in Table 2).

Each time a participant submitted a recording for CF, an orthographic transcription of the item (with the target phonemes highlighted) was displayed on the wizard's interface along with the audio recording. The wizard then chose one of three CF options for each of the highlighted phonemes, as can be seen in Figure 3. The recording was played on a loop for the wizard until he selected the appropriate option, which was immediately sent to the participant. On average, the latency between the participant submitting a recording and receiving CF was 8.54 seconds.

Figure 3

Web Interface for the Human Listener (Wizard), Who Selected "Correct" for the Production of /v/ and /w/ as CF to the Learner



After completing the training, participants recorded the same 72 pretest items a second time. The items were presented in the same order as in the pretest phase. The procedure was identical to the one used on the pretest. Each participant thus created 144 recordings in total (72 during the pretest phase and 72 during the posttest phase) for a grand total of 4,320 recordings (144 recordings × 30 participants).

Assessment of Trainees' Productions

Twenty-seven native speakers of English were recruited to rate the pre- and posttest productions of the target phonemes. Raters were university students who had taken (or were finishing) either an introduction to linguistics or introduction to world languages course. Raters listened to two productions that were identical in content, one from the pretest and one from the posttest. They were blind to which of the two samples in each pair was recorded at pre- or posttest. Raters worked for about one hour in a rating session and could return for additional sessions. Each block received ratings from two different raters. Several raters completed more than one session, but there was at least 24 hours between sessions, and they rated new items during each session.

The raters used an interface which grouped the recordings into blocks of eight items targeting the same sound contrast, randomly selecting items from the pool of participants and phrases. This setup produced 540 blocks. Raters made forced-choice pairwise comparisons between audio samples: For each pair of recordings, they indicated which recording contained a better exemplar of a target phoneme. For example, in the phrase "valuable wine," they would only evaluate the [v] or the [w] but not both. In addition, on a drop-down menu, raters also reported their confidence for each decision on a 5-point scale from 0 (*no confidence*) to 4 (*strong confidence*). They were also provided a checkbox option to flag missing recordings or recordings with quality issues. When audio quality issues were reported, these ratings were excluded from further analysis.

Data Analysis

Judgments of speech samples were encoded as -1 if the pretest recording was rated as superior to the posttest recording and +1 if the posttest recording was rated as superior. The two variables (binary choice and confidence rating) were multiplied to create a single variable called "confidence improved" (or CI) with nine response levels, from -4 indicating the rater was highly confident the learner's performance was better at pretest to +4 indicating the rater was highly confident the learner's posttest recording was better. A rating

of 0 (the midpoint of the scale) indicated that the rater was not confident in making a choice between the recordings (i.e., the learner's performance was not noticeably different in the two recordings). The CI variable was treated as an ordinal variable in statistical modeling.

Results

CF Systems and Learners' Pronunciation

Four nested generalized linear mixed-effects models were fitted to the data. The first of the four models was an intercept-only model, the second model added a fixed main effect of system (System 100, System 66, System 33), the third model added a fixed main effect for item training status ("trained" if the item was included in the training intervention, "untrained" if the learners were not exposed to the item during training), and the fourth model added the interaction effect between system and item training status. All models predicted the ordinal response variable "confidence improved" (as defined in the "Data Analysis" subsection) and included random intercepts for item, learner, and rater.

Likelihood-ratio tests were performed to assess the change of the goodness of fit with the addition of each fixed predictor. The results of the test showed that each consecutive model fit the data significantly better than the previous model: $\chi^2(1) = 7.85$, $p = .005$; $\chi^2(2) = 11.16$, $p = .004$; $\chi^2(2) = 13.01$, $p = .002$. This indicates that there were statistically significant differences (a) among the three systems in terms of pronunciation improvement and (b) between trained and untrained items and that (c) the differences among systems varied significantly between trained and untrained items.

To better understand these results, post-hoc comparisons between the three systems were performed with Tukey adjustments for multiple comparisons. These comparisons showed that, for trained items, System 100 was not significantly different from System 66 ($p = .900$), but it was significantly different from System 33 ($p < .001$), and System 66 was significantly different from System 33 ($p < .001$). However, for untrained items, there were no significant differences among the three systems ($p > .600$).

While small p-values ($p < .001$) give grounds for rejecting the null hypothesis and thus suggest statistically significant differences between the systems, large p-values only indicate that the null hypothesis could not be rejected, which is not the same as confirming the null hypothesis. Therefore, a large p-value is not interpretable in terms of whether there is a real difference between two conditions (i.e., there is no way to distinguish the true absence of an effect from a Type II error). The Bayesian approach to statistical inference (Wagenmakers, 2007) does not have this limitation and has been successfully applied to pronunciation training research in previous work (e.g., Silpachai et al., 2021). For this reason, we also fitted a Bayesian mixed-effects model with the same structure as the final (fourth) model described above, from which a posterior distribution of the "confidence improved" variable was obtained.

Figure 4 and Table 3 show the estimated cell means for "confidence improved" by system and item training status with 95% probability intervals (PIs). The results for trained items were different from the results for untrained ones. The estimates showed that, for trained items, the 95% PIs for Systems 100 and 66 do not include zero, indicating that these two systems were both effective in producing pronunciation improvement after training. However, the 95% PI for System 33 does include zero, indicating that System 33 was not effective in promoting pronunciation improvement. For untrained items, the 95% PIs for all three systems included zero, suggesting lack of effect of any of the systems for improvement of untrained items.

Finally, we estimated Bayes factors (Wagenmakers, 2007) from the model to quantify the strength of evidence in favor of the null hypothesis (that training did not lead to pronunciation improvement) and the alternative hypothesis (that training led to pronunciation improvement) for all three systems, and for trained and untrained items. The Bayes factor (null/alternative) of 18.90 suggests that "very strong" evidence exists for the lack of difference between Systems 100 and 66. An estimated Bayes factor (null/alternative) of 24.31 also suggests that there is "very strong" evidence that Systems 100 and 66 were different from System 33 but only for trained items. For untrained items, there is no evidence beyond "anecdotal" for either the

presence or absence of a difference between Systems 100/66 and System 33, as shown by the Bayes factor (null/alternative) of 0.93.

In other words, Systems 100 and 66 significantly outperformed System 33 on the trained items, and there is very strong evidence that System 66 performed as well as System 100. As far as the transfer of learning gains to untrained items goes, it is not clear whether there was any improvement on the untrained items, as seen from the Bayes factors. Collecting more data in follow-up studies might help shed more light on how the accuracy of feedback affects the ultimate transfer of learning gains to untrained items.

Although Systems 100 and 66 resulted in similar improvement, participants' ratings of the systems suggest that the participants recognized differences in the accuracy with which the systems gave them feedback. Participants rated System 100 (M = 8.1, SD = 1.5) as more accurate than System 66 (M = 6.8, SD = 1.9) and System 66 as more accurate than System 33 (M = 5.9, SD = 2.4). A one-way ANOVA showed that these differences were statistically significant: $F(2) = 10.72$, $p < .001$. This indicates that participants could perceive differences in CF accuracy between the three systems.

Figure 4

Estimated Cell Means by System With 95% PIs

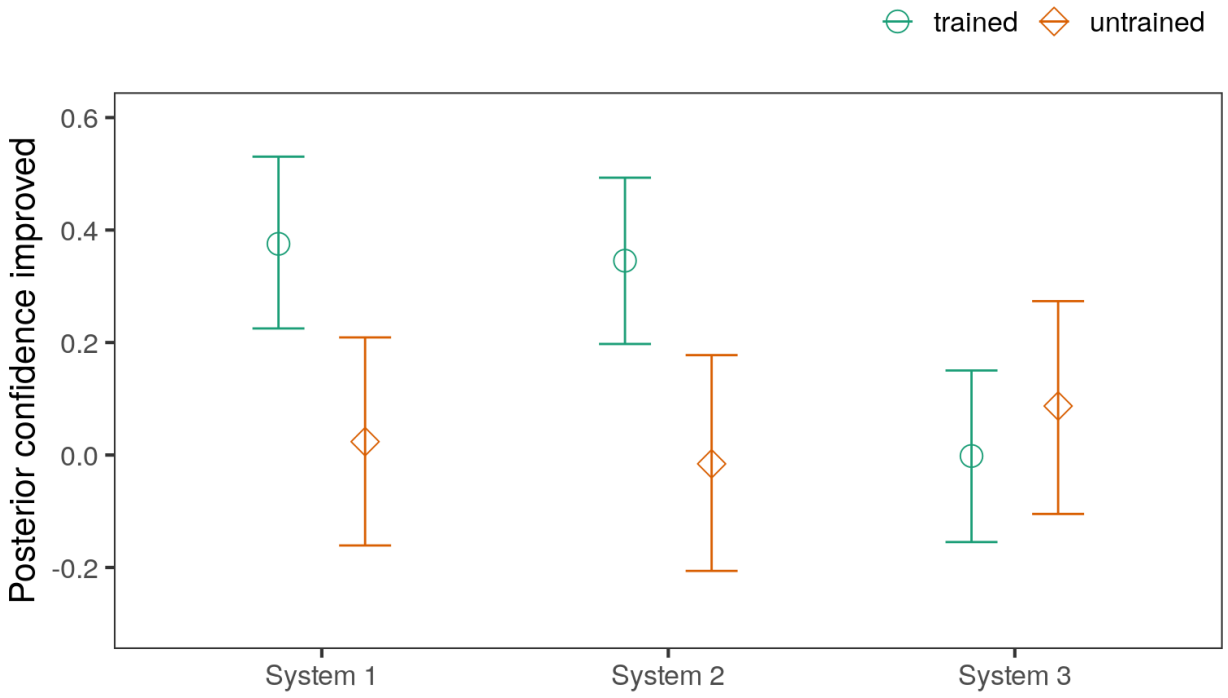


Table 3

Estimated Cell Means by System With 95% PIs

	Trained Items		Untrained Items	
	Mean	95% PI	Mean	95% PI
System 100	0.38	[0.23; 0.53]	0.02	[-0.16; 0.21]
System 66	0.35	[0.20; 0.49]	-0.02	[-0.21; 0.18]
System 33	0.00	[-0.15; 0.15]	0.09	[-0.10; 0.27]

Discussion

In our single-session training experiment, our primary research question was concerned with whether the accuracy of CF from three systems influenced how accurately L2 learners produced trained segmentals. The results suggest that 'good enough' CF accuracy promotes L2 pronunciation improvement. For trained items, the accuracy of a CF system mattered for L2 pronunciation improvement, but it only mattered up to a certain point. Systems 100 and 66, both of which provided more accurate than inaccurate CF, led to greater improvements in pronunciation than System 33, which provided more inaccurate than accurate CF. Thus, these results suggest that CF does not have to be perfect but must be 'good enough' (i.e., in this study, it had to be at least 66% accurate) to be successful in improving participants' pronunciation. In regard to our secondary research question, which was about transfer to untrained items, the accuracy level of a CF system did not matter for improvement. It is possible that, because this was only a single-session experiment, the amount of time necessary to show transfer was insufficient. Thus, transfer, if there was any, could not have been observed no matter how good the feedback was.

Why Was System 100 as Effective as System 66?

Speculations can be made about why we found no difference between Systems 100 and 66. First, it may be that a 66% accuracy level is sufficient for learners to tolerate and filter out the inadequacies of a CF system (McCrocklin, 2016). In a similar finding, McCrocklin (2014) found that Windows Speech Recognition promoted learner autonomy and improvement of pronunciation even though students recognized that the system was not always accurate but was "high enough to facilitate pronunciation practice for intermediate to advanced learners" (p. 93).

Second, there might be an upper limit to the usefulness of CF accuracy, that is, CF will be beneficial for learners at ever higher levels of accuracy, but the benefits may reach a plateau limited by certain characteristics of the learners. We know that production accuracy for L2 learners may depend on perception accuracy (Flege, 1995), the ability to produce sounds accurately all the time, or confusions created by orthography. In regard to perception, Brown (1998) found that advanced level Japanese learners of English had very poor perception of the /ɹ/-/l/ contrast (around 30%) but that many learners could produce the /ɹ/-/l/ contrast more accurately than they perceived it. It may also be that production accuracy can only improve so much and that errors will still be evident. In Brown's (1998) study, the learners produced more accurately than they perceived, but even the most successful learners only achieved about 60% accuracy in production. In regard to orthography, all of our target sounds were orthographically presented in phrases. This may have caused some loss of accuracy over other types of production tasks (such as delayed repetition). Orthography may promote its own difficulty to L2 pronunciation learners, especially when they are faced with an opaque orthography like English (Bassetti & Atkinson, 2015). As a result, it may be that orthography, either in the target words or example words, affected their production accuracy. A final reason for the lack of difference between Systems 100 and 66 is that the gap between 66% and 100% accuracy might not be noticeable to learners. Because of learners' uncertainty about their pronunciation accuracy in the L2 in all cases, it might be difficult for them to consistently distinguish between the feedback given to

them by a perfect system and a 'good enough' system. In this respect, L2 learners differ from L1 speakers who will notice when the system underperforms in accuracy. While the ability to distinguish between perfect and 'good enough' might be higher as proficiency increases, other evidence shows that advanced L2 speakers can still struggle to identify even a majority of their errors (Dlaska & Krekeler, 2008).

Why Were Systems 100 and 66 More Effective Than System 33?

It is possible that Systems 100 and 66 were more effective than System 33 because learners perceived that their CF was correct most of the time (i.e., 100% or 66%) whereas System 33 was incorrect most of the time. This explanation assumes that learners are sufficiently aware of the correctness of their English pronunciation to judge CF reliability, leading them to have trust in the CF they receive. In McCrocklin's (2014) study, some of the learners using the ASR systems believed the feedback they received was wrong too often, and so they sought out other ASR systems (e.g., Dragon, Alexa) to evaluate their speech. Although we know of no studies on learners' trust in the accuracy of CAPT systems, we believe that this may be a factor in why System 33 had worse results: Its CF was seen as untrustworthy.

Why Did System Accuracy Not Affect the Pronunciation of Untrained Items?

The finding that system accuracy did not affect the pronunciation of untrained items indicates that the training was not effective in helping participants to generalize their learning to novel items. A few factors may have prevented such generalization. First, the training was likely too short for the trainees to cement their learning. The inherent quality of CF might have been high enough in Systems 66 and 100, but the time of exposure was very limited. Second, there may have been too many sound contrasts for a single session, making it too cognitively demanding for the trainees to hold such diverse sound characteristics in their minds.

Other Factors

While other factors may have influenced the equivalence of Systems 100 and 66 in this study, the usefulness of an automatic CF system depends on both the accuracy of the system and the learner. In regard to users of the systems, there is a large amount of inherent variability in how they interact with the feedback they receive. A comparison with native speakers is instructive. Native speakers are likely to evaluate the performance of an MPD system more accurately because they both hear and produce their native phonemes accurately. Furthermore, they know their production is accurate and thus evaluate errors from the MPD system as deficiencies in the system, not as evidence of deficiencies in their own speech. Learners with noticeable segmental errors, on the other hand, are unlikely to perceive or produce L2 phonemes with anything close to 100% accuracy or 100% confidence (e.g., as in Brown, 1998). This suggests that deficiencies in the MPD system may be interpreted by learners as deficiencies in their own production or perception, as long as these deficiencies are not blatant. Previous research has shown that L2 speakers are aware that their pronunciation affects their ability to interact successfully in the L2 context (e.g., Cheng et al., 2021) and that their pronunciation may even make them feel stigmatized (Gluszek & Dovidio, 2010). However, learners may not be aware of which errors make their speech harder to understand, so instead they focus on the sounds that others may tell them about (Derwing & Rossiter, 2002). This is in line with what is known from other research, in which learners are not always able to identify differences between their own speech and that of a model speaker (Silpachai et al., 2021). Learners also have varying awareness of their own errors and how they actually produce sounds in careful or connected speech. In this study, for example, most participants had trouble distinguishing /n/-/ŋ/, but they also seemed to struggle to understand what was wrong in their production, even with 100% accurate feedback. On the other hand, learners may be aware of certain problematic phonemes that they are not confident of but that they produce accurately. In one example, Derwing (2003) found that the voiceless interdental fricative /θ/ (as in *think* and *both*) is a sound that most learners of English identify as a problem in their speech. In other words, /θ/ is a well-known shibboleth in English pronunciation. In our study, participants were surprisingly quite accurate in producing /θ/, perhaps because of their advanced proficiency, but when the feedback was modified to say their accurate pronunciations were wrong, we observed them continuing to practice their pronunciation before re-

recording the words containing /θ/. The result of this mismatch between inaccurate systems and accurate pronunciation is that learners may benefit from less than perfect CF, and if they believe the system is largely accurate, they will come to trust that the overall system is giving them good feedback.

Limitations

This was a single-session training study, and a longer period of training, either under self-study conditions or with human instructional help, could show other informative trends. However, it is encouraging that the CAPT CF was effective even in a short training study and even with less than 100% accuracy. Another limitation is that all trainees received training on the same set of likely problems, but not all actually had the same problems. A diagnostic pretest followed by adaptive training would help to ensure that practice included only errors that learners truly had trouble with.

Another potential limitation is that each learner received CF from a single human listener. In a pilot using two listeners conducted before the full study, the two wizard listeners almost always agreed on the feedback provided, and a single rater allowed for unambiguous feedback. Consequently, there may have been a reduction in reliability or consistency of rating as well as in accuracy and fairness. However, the typical classroom has only a single instructor, who provides feedback unevenly because of many factors, and our system provided directed feedback on all targeted sounds, making the accuracy of the wizard's original feedback more systematic than that provided by an instructor in classroom contexts.

Conclusion

Our findings indicate that more accurate CF does not necessarily result in greater L2 pronunciation improvement. Instead, L2 learners seem to benefit both from feedback that is 66% accurate and from feedback that is human-like in accuracy, at least for items on which they have received CF. In contrast, there is a difference between these two levels of accuracy and the CF associated with System 33. Based on our results, we conclude that there is indeed a 'good enough' level of feedback accuracy and that CF accuracy does not need to be human-like to produce improvements in pronunciation.

Because a pronunciation training computer application does not need to provide CF with 100% accuracy, software developers can still provide CF benefits using a CF tool with a 'good enough' level of accuracy. A priority for future research in this area should be to establish what level of accuracy is 'good enough' for L2 learners and whether different sound contrasts benefit from different levels of accuracy. It is likely that varying levels of accuracy may be required for learners of differing proficiency levels and that learners whose pronunciation has fewer errors may require more accurate feedback because they are likely to notice deficiencies in less accurate systems. In contrast, learners whose pronunciation is more variable may find benefits from less accurate systems because of their own inconsistent pronunciation. Future studies might also consider using more than one model voice, focusing on fewer sound contrasts (e.g., Lively et al., 1993, 1994, and Logan et al., 1991, who focused on only /ɹ/ and /l/) and extending the period of training over multiple sessions.

Furthermore, future research may consider investigating the relationship between proficiency and the ability to distinguish between a perfect system and a 'good enough' system. Presumably, highly proficient learners should be able to distinguish between systems more easily compared to learners with lower proficiency. If this is the case, the accuracy of CF that is 'good enough' may change based on the L2 learner population.

Acknowledgements

This study was funded by NSF collaborative grant 2016984 awarded to Iowa State University and Texas A&M University. The authors would like to thank a graduate research assistant, Mahdi Duris, and undergraduate research assistants, Jamie Smith, Zoë DeKruif, and Jennifer Godbersen, for their help during data collection. The authors also thank Jens Roeser for his assistance with statistical analysis.

References

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2022). 'Look, I can speak correctly': Learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology. *Computer Assisted Language Learning*, 1–29. https://doi.org/10.1080/09588221.2022.2080230

Bassetti, B., & Atkinson, N. (2015). Effects of orthographic forms on pronunciation in experienced instructed second language learners. *Applied Psycholinguistics*, *36*(1), 67–91. https://doi.org/10.1017/s0142716414000435

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly*, *22*(4), 593–606. https://doi.org/10.2307/3587258

Brown, C. A. (1998). The role of the L1 grammar in the L2 acquisition of segmental structure. *Second Language Research*, *14*(2), 136–193. https://doi.org/10.1191/026765898669508401

Browne, J. T. (2019). Wizard of Oz prototyping for machine learning experiences. In *Extended abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). https://doi.org/10.1145/3290607.3312877

Cheng, L., Im, G. H., Doe, C., & Douglas, S. R. (2021). Identifying English language use and communication challenges facing "entry-level" workplace immigrants in Canada. *Journal of International Migration and Integration*, *22*(3), 865–886. https://doi.org/10.1007/s12134-020-00779-w

Dai, Y., & Wu, Z. (2021). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning*, 1–24. https://doi.org/10.1080/09588221.2021.1952272

Darcy, I. (2018). Powerful and effective pronunciation instruction: How can we achieve it? *The CATESOL Journal*, *30*(1), 13–45.

Derwing, T. (2003). What do ESL students say about their accents? *Canadian Modern Language Review*, *59*(4), 547–567. https://doi.org/10.3138/cmlr.59.4.547

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research* (Vol. 42). John Benjamins Publishing Company.

Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly*, *34*(3), 592–603. https://doi.org/10.2307/3587748

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, *48*(3), 393–410. https://doi.org/10.1111/0023-8333.00047

Derwing, T. M., & Rossiter, M. J. (2002). ESL learners' perceptions of their pronunciation needs and strategies. *System*, *30*(2), 155–166.

Ding, S., Liberatore, C., Sonsaat, S., Lučić, I., Silpachai, A., Zhao, G., Levis, J. M., Chukharev-Hudilainen, E., & Gutierrez-Osuna, R. (2019). Golden speaker builder – An interactive tool for pronunciation training. *Speech Communication*, *115*, 51–66. https://doi.org/10.1016/j.specom.2019.10.005

Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation. *System*, *36*(4), 506–516.

Dlaska, A., & Krekeler, C. (2013). The short-term effects of individual corrective feedback on L2 pronunciation. *System*, *41*(1), 25–37. https://doi.org/10.1016/j.system.2013.01.005

Dressman, M., & Sadler, R. W. (Eds.). (2020). *The handbook of informal language learning*. John Wiley & Sons.

Ellis, R. (2006). Researching the effects of form-focused instruction on L2 acquisition. *AILA Review*, *19*(1), 18–41. https://doi.org/10.1075/aila.19.04ell

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.

Fleming, V. (Director). (1939). *The Wizard of Oz* [Film]. Metro-Goldwyn-Mayer.

García, C., Nickolai, D., & Jones, L. (2020). Traditional versus ASR-based pronunciation instruction: An empirical study. *CALICO Journal*, *37*(3), 213–232. https://doi.org/10.1558/cj.40379

Gluszek, A., & Dovidio, J. F. (2010). Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the United States. *Journal of Language and Social Psychology*, *29*(2), 224–234. https://doi.org/10.1177/0261927X09359590

Gordon, J., & Darcy, I. (2016). The development of comprehensible speech in L2 learners: A classroom study on the effects of short-term pronunciation instruction. *Journal of Second Language Pronunciation*, *2*(1), 56–92. https://doi.org/10.1075/jslp.2.1.03gor

Hardison, D. M. (2004). Generalization of computer assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, *8*(1), 34–52. http://dx.doi.org/10125/25228

Hirata, Y. (2004). Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts. *Computer Assisted Language Learning*, *17*(3–4), 357–376. https://doi.org/10.1080/0958822042000319629

Lee, A. H., & Lyster, R. (2017). Can corrective feedback on second language speech perception errors affect production accuracy? *Applied Psycholinguistics*, *38*(2), 371–393. https://doi.org/10.1017/S0142716416000254

Lee, E. J. (2016). Reducing international graduate students' language anxiety through oral pronunciation corrections. *System*, *56*, 78–95. https://doi.org/10.1016/j.system.2015.11.006

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255. https://doi.org/10.1121/1.408177

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, *96*(4), 2076–2087. https://doi.org/10.1121/1.410149

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1–40. https://doi.org/10.1017/S0261444812000365

Malakar, M., & Keskar, R. B. (2021). Progress of machine learning based automatic phoneme recognition and its prospect. *Speech Communication, 135*, 37–53. https://doi.org/10.1016/j.specom.2021.09.006

Martin, I. A., & Sippel, L. (2021). Is giving better than receiving?: The effects of peer and teacher feedback on L2 pronunciation skills. *Journal of Second Language Pronunciation*, *7*(1), 62–88. https://doi.org/10.1075/jslp.20001.mar

McCrocklin, S. M. (2014). *The potential of Automatic Speech Recognition for fostering pronunciation learners' autonomy* (Publication No. 3641050) [Doctoral dissertation, Iowa State University]. ProQuest Dissertations Publishing.

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of Automatic Speech Recognition. *System*, *57*, 25–42. https://doi.org/10.1016/j.system.2015.12.013

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation*, *5*(1), 98–118. https://doi.org/10.1075/jslp.16034.mcc

McCrocklin, S., & Edalatishams, I. (2020). Revisiting popular speech recognition software for ESL speech. *TESOL Quarterly*, *54*(4), 1086–1097. https://doi.org/10.1002/tesq.3006

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals*, *51*(3), 617–637. https://doi.org/10.1111/flan.12348

Qian, M. (2018). *An adaptive computational system for automated, learner-customized segmental perception training in words and sentences: Design, implementation, assessment* (Publication No. 13418755) [Doctoral dissertation, Iowa State University]. ProQuest Dissertations Publishing.

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning*, *62*(2), 595–633. https://doi.org/10.1111/j.1467-9922.2011.00639.x

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Silpachai, A., Rehman, I., Barriusio, T. A., Levis, J., Chukharev-Hudilainen, E., Zhao, G., & Gutierrez-Osuna, R. (2021). Effects of voice type and task on L2 learners' awareness of pronunciation errors. *Proceedings of Interspeech, 2021,* 1952–1956. http://dx.doi.org/10.21437/Interspeech.2021-701

Sung, E. (2008). The effects of computer-assisted pronunciation training in the production of English coda consonants and consonant clusters. *Multimedia-Assisted Language Learning*, *11*(3), 46–68.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326–344. https://doi.org/10.1093/applin/amu076

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Zhang, R., & Yuan, Z. M. (2020). Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition*, *42*(4), 905–918. https://doi.org/10.1017/S0272263120000121

About the Authors

Alif Silpachai is a postdoctoral researcher at the Centre for Language Studies at Radboud University. His main research focuses on the production and perception of tone and second language pronunciation. Alif Silpachai is the corresponding author.

E-mail: alif.silpachai@ru.nl

ORCiD: https://orcid.org/0000-0002-6316-3806

Reza Neiriz has his PhD in Applied Linguistics and Technology from Iowa State University. His research interests include computer-assisted language testing and teaching with a focus on testing different aspects of oral communication ability.

E-mail: rneiriz@iastate.edu

ORCiD: https://orcid.org/0000-0002-2285-9575

MacKenzie Novotny is a PhD student in Applied Linguistics and Technology at Iowa State University. She received an MA degree in TESL/Applied Linguistics in 2021 from Iowa State University. Her research interests include computer-assisted pronunciation training, automatic pronunciation assessment, and second language speech perception.

E-mail: mnovotny@iastate.edu

ORCiD: https://orcid.org/0009-0007-8302-2377

Ricardo Gutierrez-Osuna is a Professor with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA. His current research interests include voice and accent conversion, speech and face perception, wearable physiological sensors, and active sensing.

E-mail: rgutier@cse.tamu.edu

ORCID: https://orcid.org/0000-0003-2817-2085

John M. Levis is a Professor in Applied Linguistics and TESL at Iowa State University. He is the author of *Intelligibility, oral communication and the teaching of pronunciation* (Cambridge University Press) and is a co-developer of L2 ARCTIC: A Nonnative English Speech Corpus.

E-mail: jlevis@iastate.edu

ORCiD: https://orcid.org/0000-0001-7405-5969

Evgeny Chukharev is an Associate Professor in Applied Linguistics and Technology at Iowa State University. His research program investigates topics in language processing, first and second language acquisition, and language change.

E-mail: evgeny@iastate.edu

ORCiD: https://orcid.org/0000-0001-7930-5787