

2-2018

A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer

Manman Qian
Iowa State University

Evgeny Chukharev-Hudilainen
Iowa State University, evgeny@iastate.edu

John M. Levis
Iowa State University, jlevis@iastate.edu

Follow this and additional works at: https://lib.dr.iastate.edu/engl_pubs

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Educational Methods Commons](#), [Educational Technology Commons](#), [International and Comparative Education Commons](#), and the [Speech and Rhetorical Studies Commons](#)

The complete bibliographic information for this item can be found at https://lib.dr.iastate.edu/engl_pubs/262. For information on how to cite this item, please visit <http://lib.dr.iastate.edu/howtocite.html>.

This Article is brought to you for free and open access by the English at Iowa State University Digital Repository. It has been accepted for inclusion in English Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer

Abstract

Many types of L2 phonological perception are often difficult to acquire without instruction. These difficulties with perception may also be related to intelligibility in production. Instruction on perception contrasts is more likely to be successful with the use of phonetically variable input made available through computer-assisted pronunciation training. However, few computer-assisted programs have demonstrated flexibility in diagnosing and treating individual learner problems or have made effective use of linguistic resources such as corpora for creating training materials. This study introduces a system for segmental perceptual training that uses a computational approach to perception utilizing corpusbased word frequency lists, high variability phonetic input, and text-to-speech technology to automatically create discrimination and identification perception exercises customized for individual learners. The effectiveness of the system is evaluated in an experiment with pre- and post-test design, involving 32 adult Russian-speaking learners of English as a foreign language. The participants' perceptual gains were found to transfer to novel voices, but not to untrained words. Potential factors underlying the absence of word-level transfer are discussed. The results of the training model provide an example for replication in language teaching and research settings.

Keywords

Computer-Assisted Language Learning, Pronunciation, Second Language Acquisition, Perception

Disciplines

Educational Assessment, Evaluation, and Research | Educational Methods | Educational Technology | International and Comparative Education | Speech and Rhetorical Studies

Comments

This article is published as Qian, M., Chukharev-Hudilainen, E., & Levis, J. (2018). A system for adaptive high-variability segmental perceptual training: implementation, effectiveness, transfer. *Language Learning & Technology*, 22(1), 69–96. DOI: [10125/44582](https://doi.org/10.125/44582). Posted with permission.



A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer

Manman Qian, Iowa State University

Evgeny Chukharev-Hudilainen, Iowa State University

John Levis, Iowa State University

Abstract

Many types of L2 phonological perception are often difficult to acquire without instruction. These difficulties with perception may also be related to intelligibility in production. Instruction on perception contrasts is more likely to be successful with the use of phonetically variable input made available through computer-assisted pronunciation training. However, few computer-assisted programs have demonstrated flexibility in diagnosing and treating individual learner problems or have made effective use of linguistic resources such as corpora for creating training materials. This study introduces a system for segmental perceptual training that uses a computational approach to perception utilizing corpus-based word frequency lists, high variability phonetic input, and text-to-speech technology to automatically create discrimination and identification perception exercises customized for individual learners. The effectiveness of the system is evaluated in an experiment with pre- and post-test design, involving 32 adult Russian-speaking learners of English as a foreign language. The participants' perceptual gains were found to transfer to novel voices, but not to untrained words. Potential factors underlying the absence of word-level transfer are discussed. The results of the training model provide an example for replication in language teaching and research settings.

Keywords: *Computer-Assisted Language Learning, Pronunciation, Second Language Acquisition, Perception*

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

APA Citation: Qian, M., Chukharev-Hudilainen, E., & Levis, J. (2018). A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer. *Language Learning & Technology*, 22(1), 69–96. doi:10.125/44582

Introduction

In the early 21st century, three trends in second language (L2) pronunciation instruction emerged. First, teaching pronunciation regained once-lost attention in L2 pedagogy (Derwing & Munro, 2015), and a consensus formed that the development of intelligible speech was the primary goal for pronunciation instruction. Second, learner-centered pedagogies became critical for effective pronunciation acquisition. Specifically, pronunciation instruction increasingly recognized that not all errors were equally important, and that different learners required different instructional emphases. Third, the ever-increasing use of computer-assisted language learning (CALL) tools penetrated all areas of L2 pedagogy, including pronunciation. Computer-assisted pronunciation teaching (CAPT) applications have proved capable of adapting to individual learners, something that is extremely challenging in conventional classroom settings, especially for large class sizes as in many English as a foreign language (EFL) contexts (Bahanshal, 2013; Liang, 2009).

However, in current CAPT applications, little use has been made of the available linguistic and

technology resources (e.g., word frequency lists, spoken language databases, and text-to-speech technology), and many CAPT applications show limited flexibility in diagnosing and treating individual learner problems. Some applications claiming to reduce accent have been over-commercialized into tech-showy packages that attract buyers but do not effectively serve the goal of pronunciation instruction (Neri, Cucchiari, Strik, & Boves, 2002; Thomson, 2013). Arguably, a model is yet to be developed that would help improve pronunciation intelligibility through a flexible approach to pronunciation instruction. Ideally, such a model should build upon the cutting-edge research on the effect of high variability phonetic training (Thomson, 2011, 2012; Wang & Munro, 2004). This article describes an attempt to introduce a first iteration of such a model, exemplified in a prototype CAPT tool that provides perceptual training of English segmental features for L2 learners.

This article is organized as follows: First, we review the relevant literature on the importance of proficiency in L2 segmental perception, the desired pedagogical approaches for L2 segmental instruction, and limitations of the existing perceptual training models. Then, we introduce the design of a novel CAPT tool that bridges the gap between research and practice in L2 perceptual training. Finally, we present an empirical study to assess the effectiveness of the tool and the transfer of learning according to Levis' (2007) CAPT evaluative framework.

Literature Review

Selectively Teaching L2 Segmentals

The importance of segmental features (*segmentals*) to intelligibility is well-documented (e.g., Jenkins, 2000; Munro & Derwing, 2006), but segmentals do not require equal pedagogical attention in all contexts, in that not all segmental errors compromise intelligibility to the same extent (Brown, 1988). A generally accepted principle that informs pedagogical choices is the functional load (FL) principle, which measures “the work which two phonemes (or a distinctive feature) do in keeping utterances apart” (King, 1967, p. 631) and serves to rank the gravity of phonemic errors. [Appendix A](#) lists a 10-point scale error gravity hierarchy (larger numbers represent greater error gravity) for commonly confused British English phonemic contrasts (Brown, 1988). Later empirical research demonstrates a correlation between the FL value of a phoneme and its impact on intelligibility. Munro and Derwing (2006) analyzed the relation between English phoneme error types and comprehensibility ratings by native listener judges and found that errors with sound pairs of high FL (/l-n/, /ʃ-s/, and /d-z/) resulted in greater loss of comprehensibility, while comprehensibility was not impacted as seriously for errors with low FL (/ð-d/ and /θ-f/). The impact of FL on listeners' ability to understand was also sensitive to error frequency. The researchers noted that comprehensibility was worse for sentences with two high FL errors than one and that “sentences that contained only one high FL error were rated significantly worse for comprehensibility than sentences containing three low FL errors” (p. 528).

Segmental errors for non-native speakers tend to be first language (L1) specific (Flege et al., 2006; Jia, Strange, Wu, Collado, & Guan, 2006). For instance, /i-ɪ/ is notoriously problematic for Chinese speakers but does not seem to affect Arabic speakers to the same extent (Swan & Smith, 2002). The influence of L1 on L2 phonological acquisition is well recognized in L2 pronunciation research. According to the speech learning model (Flege, 1995) and the Perceptual Assimilation Model (Best, 1995), learners' ability to perceive and produce L2 phonemes are partly predictable from the acoustic distance between the target phonemes and the learners' native phonemes. A major reason for this is the hypothesis that L2 phonological acquisition attainment is often linked with the age when L2 acquisition starts (Scovel, 1969). As part of the declining process, learners lose the acuity to perceive the phonetic categories of a foreign language (Best & MacRoberts, 2003). Some believe that this loss can be reinforced by continual L1 acquisition, making ears increasingly attuned to native acoustic features (Iverson, Hazan, & Bannister, 2005). As a consequence, when hearing an unfamiliar phoneme, adult learners are inclined to filter the sound through their native phonetic inventory, trying to map the foreign phoneme to L1 representations in close proximity.

However, speakers sharing an L1 may not experience the same difficulties in acquiring segmentals. For example, segmental errors such as /l-n/ substitution with Mandarin speakers are regionally dependent despite having the same L1 as other speakers without /l-n/ substitution errors (Richards, 2012). Munro, Derwing, and Thomson (2015) monitored the performance in English consonant productions of 17 Mandarin speakers and 23 Slavic speakers over the course of two years, during which no explicit pronunciation instruction was provided. Despite similarities (which were expected) among learners of the same L1, the types of difficulties faced by the learners varied dramatically—among the 21 sounds examined for the two L1 groups, only one sound, coda /ld/, was uniformly difficult to the Mandarin participants, whereas for each of the other 20 sounds, a good proportion (at least 20% and on most occasions over 40%) of the participants did not show pronunciation problems, raising questions about a strictly L1-based pedagogy.

There is also evidence that some L2 segmentals may not require explicit teaching. For instance, Munro and Derwing (2008) observed that some Mandarin and Slavic speakers learned to correctly produce English vowels that did not exist in their L1s without instruction (such as /ɪ/, /ʊ/, and /ʌ/) during their initial months of stay in the target language country. Munro et al. (2015) further noted that some segmental problems were self-corrected over time by at least some learners.

All these findings have important implications for the teaching of L2 segmentals, in that teaching materials need to be prepared in both a principled and flexible fashion. Specifically, only the segmentals with high FLs should be selected for pronunciation syllabi. At the same time, materials should also be fine-tuned to address different L1s and for individual learners within an L1 group.

Necessity of Perceptual Training

L2 perceptual training is necessary not only because natural L2 perceptual acquisition is challenging for adult learners due to their potential lack of perception of foreign sounds, but also because perceptual training facilitates oral production. Many researchers view perception as a necessary precursor for production (Bongaerts, van Summeren, Planken, & Schils, 1997; Denes & Pinson, 1963; Kim, 2005; Neufeld, 1988). Research suggests a precedent relationship of perception development to production achievement (e.g., Baker & Trofimovich, 2001; Detey & Racine, 2015; Walden, 2014), such that perceptual insufficiency tends to inhibit production performance (Flege, Bohn, & Jang, 1997; Iverson et al., 2005). The indispensable role of perception to production may be further strengthened by recent neurolinguistic discoveries that the ability to perceive is essential to accurate articulation (Golestani & Pallier, 2007). However, the perception-precedes-production hypothesis has remained controversial, especially with growing work showing that production development can be achieved through interventions other than perceptual training, such as visual feedback (Gick, Bernhardt, Bacsfalvi, & Wilson, 2008; Olson, 2014; Patten & Edmonds, 2015; Suemitsu, Dang, Ito, & Tiede, 2015). Despite the controversies, one phenomenon that holds true is that the development of perception and production is inextricably linked. The mutually facilitative interaction between perception and production has been demonstrated in a great many studies (e.g., Catford & Pisoni, 1970; Linebaugh & Roche, 2013, 2015; Pimsleur, 1963; Wang, Jongman, & Sereno, 2003), including perceptual research conducted specifically at the segmental level (Bradlow, Pisoni, Yamada, & Tohkura, 1997; Bradlow, Yamada, Pisoni, & Tohkura, 1999; Lambacher, Martens, Kakehi, Marasinghe, & Molholt, 2005; Lopez-Soto & Kewley-Port, 2009; Okuno & Hardison, 2016; Rochet, 1995; Rvachew, 1994; Thomson, 2011).

CAPT and High-Variability Phonetic Training in Perceptual Training

Technology has been utilized for pronunciation teaching effectively in a variety of forms. Research has shown that general L2 pronunciation skills can be strengthened by the use of audio recordings (Hardison, 2003, 2005), podcasts (O'Bryan & Hegelheimer, 2007), and text-to-speech (TTS) technology (Kiliçkaya, 2008). Acoustic analysis is often found to benefit L2 pronunciation instruction for a variety of features, including intonation (Levis & Pickering, 2004), stress (Coniam, 2002), rhythm (Coniam, 2002; Varden, 2006), and segmentals (Lambacher, 1999). Automated speech recognition feedback has also been

reported to have a positive effect on improving L2 English production (Hincks, 2003, 2005; McCrocklin, 2016; Walker, Trofimovich, Cedergren, & Gatbonton, 2011). In addition to their effect on L2 English pronunciation learning, technology can also promote non-English phonological acquisition (Chun, Jiang, Meyr, & Yang, 2015: Mandarin tones; Ducate & Lomicka, 2009: German and French pronunciation; Hardison, 2004: French prosody and segmentals; Hirata, 2004: Japanese intonation; Hirata & Kelly, 2010: Japanese segmentals; Kawai & Hirose, 2000: Japanese phonemes; Lord, 2008: Spanish pronunciation; Motohashi-Saigo & Hardison, 2009: Japanese segmentals). Specific to the development of segmental perception, the focus of the current study, computer-mediated auditory training has also been reported as efficacious in a number of studies (Iverson & Evans, 2009; Nishi & Kewley-Port, 2007, 2008; Thomson, 2011, 2012; Wang & Munro, 2004).

Perception training can be effectively carried out using high-variability phonetic training (HVPT), which relies on speech input produced in multiple phonetic contexts by multiple voices (Pisoni & Lively, 1995).¹ A plethora of studies have reported HVPT as effective for perceptual training. Both synthetic (Jamieson & Morosan, 1986, 1989; Strange & Dittmann, 1984) and natural voices (Lively, Logan, & Pisoni, 1993; Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994; Logan, Lively, & Pisoni, 1991) have been used for creating HVPT materials in experiments (although little work has compared the effectiveness of the two stimulus types). Evidence supports the superiority of HVPT over techniques using single-talker stimuli in facilitating L2 perception development (Lambacher et al., 2005; Wang et al., 2003; Wang & Munro, 2004; Wang, Spence, Jongman, & Sereno, 1999). HVPT has also been recognized as effective for promoting the transfer of perceptual gains from trained to untrained words and talkers (e.g., Bradlow et al., 1997; Bradlow, 2008; Carlet & Cebrian, 2014; Iverson et al., 2005). The model has also provided positive perceptual training results on English segmentals such as /ɛ/ and /æ/ in CAPT applications (Thomson, 2011, 2012).

Lack of Flexibility in Existing CAPT Programs

Despite the significant body of research outlined above, the full potential of CAPT has not been realized in perceptual L2 segmental training. For example, adaptive learner models, used in some CALL applications (Chukharev-Hudilainen & Klepikova, 2016) have not been sufficiently applied to CAPT. Levis (2007) argues for the use of adaptive modeling in this context:

The [segmental training] system should assist learners and teachers in prioritizing pronunciation topics by channeling learners toward typical vowel and consonant errors for their language backgrounds. For example, a Korean learner of English would, after setting up a user profile, be directed to pronunciation topics that are problematic for Korean learners. ... Even better than this rather crude channeling mechanism would be an error diagnostic informed by language specific filtering. A diagnostic component in a CAPT system should include perception elements in which learners identify and discriminate among problematic sounds. (p. 188)

Munro et al. (2015) also argue that “an ideal [segmental training] approach would be a ... CAPT system that diagnoses individual learner difficulties and provides remedial exercises in exactly the areas needed” (p. 54). On the other hand, they also pointed out that such an approach did not exist yet while most current CAPT programs followed a one-size-fits-all design due to factors such as disconnections between development and use, commercial motives for enhancing marketability, an overemphasis on technology, and a lack of an explicit pedagogical base. Many current CAPT programs continue to emphasize achievement of a native accent rather than intelligibility (Levis, 2005, 2007). These issues with CAPT programs are not new. Multiple authors (e.g., Derwing & Munro, 2009; Levis, 2005, 2007; Neri et al., 2002; Thomson, 2012, 2013) have urged the design of innovative and pedagogically rigorous CAPT programs. In perception training, such an innovative model should be sensitive to learners’ L1s and diverse learner problems while reflecting research on L2 perception learning.

System Design

In this article, we describe the design and evaluation of a prototype perceptual phonetic training system. Our system was designed according to the considerations suggested by research outlined above, which were translated into the following design principles.

First, the segmentals in the system were selected with reference to the target participants' L1 background (Nilsen & Nilsen, 2010; Swan & Smith, 2002), as well as Brown's (1988) FL framework. Because a group of Russian EFL students had been selected as participants in the study prior to the design of the system, the segmental selections revolved around problematic sounds for native Russian speakers. Second, the words used for training were automatically controlled for frequency in general English data. Specifically, we ensured that our training stimuli were confined within the top 5,000 most frequent lemmata in the [Corpus of Contemporary American English \(COCA\)](#), a 520-million-word online database of real-world language use (Davies, 2008). Third, our system automatically adapted to errors specific to each learner. Corresponding training exercises were automatically generated in order to target the learner's problematic sounds. By allowing trainees to work on phonemic contrasts that they had difficulty perceiving, we hoped to ensure higher learning efficiency and promote learner autonomy and motivation.

Our prototype system took the form of a web-based learning tool accessible via computers and hand-held electronic devices. The layout of the tool's user interface (see [Figure 1](#)) was kept simple and self-explanatory, with a working area on the left and a timer on the right. All test and training items were displayed one by one in the working area of the screen. Instructions for completing each item were provided in the green bar on the top. A choice of *Don't Know* was made available for all the items through a  button in the upper-right corner.

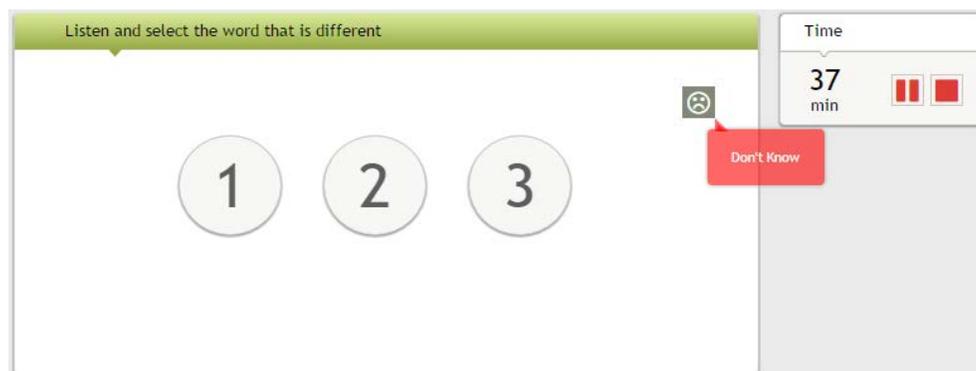


Figure 1. Interface of phonetic training system

The system was designed to support four item types shown in [Figure 2](#). All lexical and audio stimuli were presented as multiple-choice exercises created automatically by the system. The same types of items were utilized for both the training and the testing phases of the intervention, described below in more detail. The item types were (a) same–different discrimination (learners heard two words and needed to decide if they were the same or different), (b) oddity discrimination (learners heard three words and decided which one was the odd one out), (c) simple identification (learners heard a word and selected which one they heard from two options provided), and (d) yes/no identification (learners heard a word and needed to decide whether it matched the spelling provided). It should be noted that the latter two types of exercises suffered from a potential test design drawback in that they involved assessment of skills (recognition of lexical items and spellings) that were not explicitly trained in the study. To minimize the effect of this drawback, only high-frequency words were chosen as lexical stimuli in the study, assuming that learners tend to be more familiar with more frequently-occurring words.

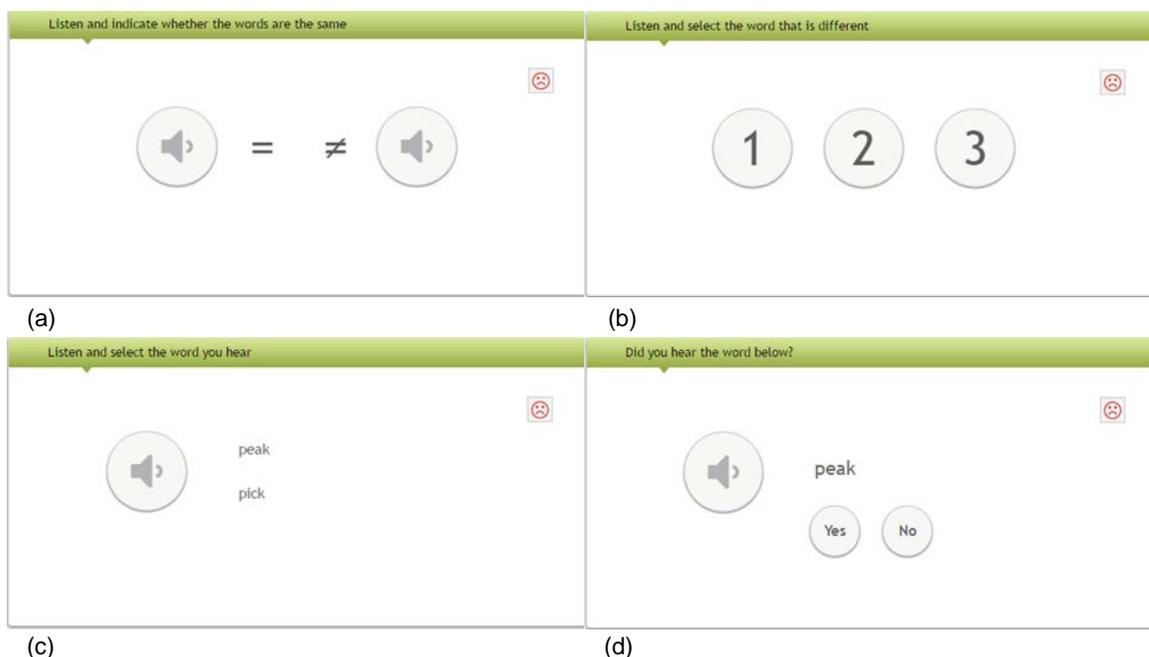


Figure 2. The four item types used in the study were (a) same–different discrimination, (b) oddity discrimination, (c) simple identification, and (d) yes/no identification.

Figure 3 illustrates the adaptive learning algorithm implemented in the system. When a learner first accessed the system, the training process began with the pre-test stage where the learner’s phonemic errors were diagnosed with a series of items. However, no evaluative feedback was provided at this time to limit learning at this stage. If the learner achieved 80% accuracy when completing the items on the pre-test, the system would proceed to the next phonemic contrast. If the learner did not meet the 80% accuracy threshold, the system would automatically present the learner with a series of training exercises. At the training stage, participants were given immediate feedback on their performance for each training item as well as an answer key for the items that they had gotten wrong. The training was repeated until the accuracy of 80% was achieved, in which case a post-test immediately followed.

Since the prototype of the system was designed for an EFL setting in Russia, 12 phonemic contrasts were selected for the study with two factors taken into consideration: (a) pronunciation errors typically shared by Russian speakers as documented in Nilsen and Nilsen (2010) and Swan and Smith (2002), and (b) the potential impact of a phonemic contrast on a speaker’s intelligibility as indicated in Brown’s (1988) table of FL values. Insights of two native-Russian teachers who were the English instructors of the participants in the study were also used as supplementary reference. The teachers were provided with a list of phonemic contrasts chosen as commonly conflated errors for Russian speakers and were asked which pairs they would recommend for inclusion in the training. They were also asked to suggest additional phonemic contrasts not on the given list. All the problem items mentioned in the literature were judged by the teachers as problematic, but several unlisted items were also proposed by the teachers such as /ε–ɪ/ and /ɔ–əʊ/, which indeed turned out to be difficult for some participants (for the number of participants who received training on each sound pair and the amount of training they received before passing training, see [Table 7](#) and [Table 8](#)). The use of these resources ensured the selection was principled and improved the opportunity for the target participants to benefit from training on the selected phonemic contrasts. [Appendix B](#) provides a list of the selected 12 sound pairs in a descending order based on the sounds’ FL values.

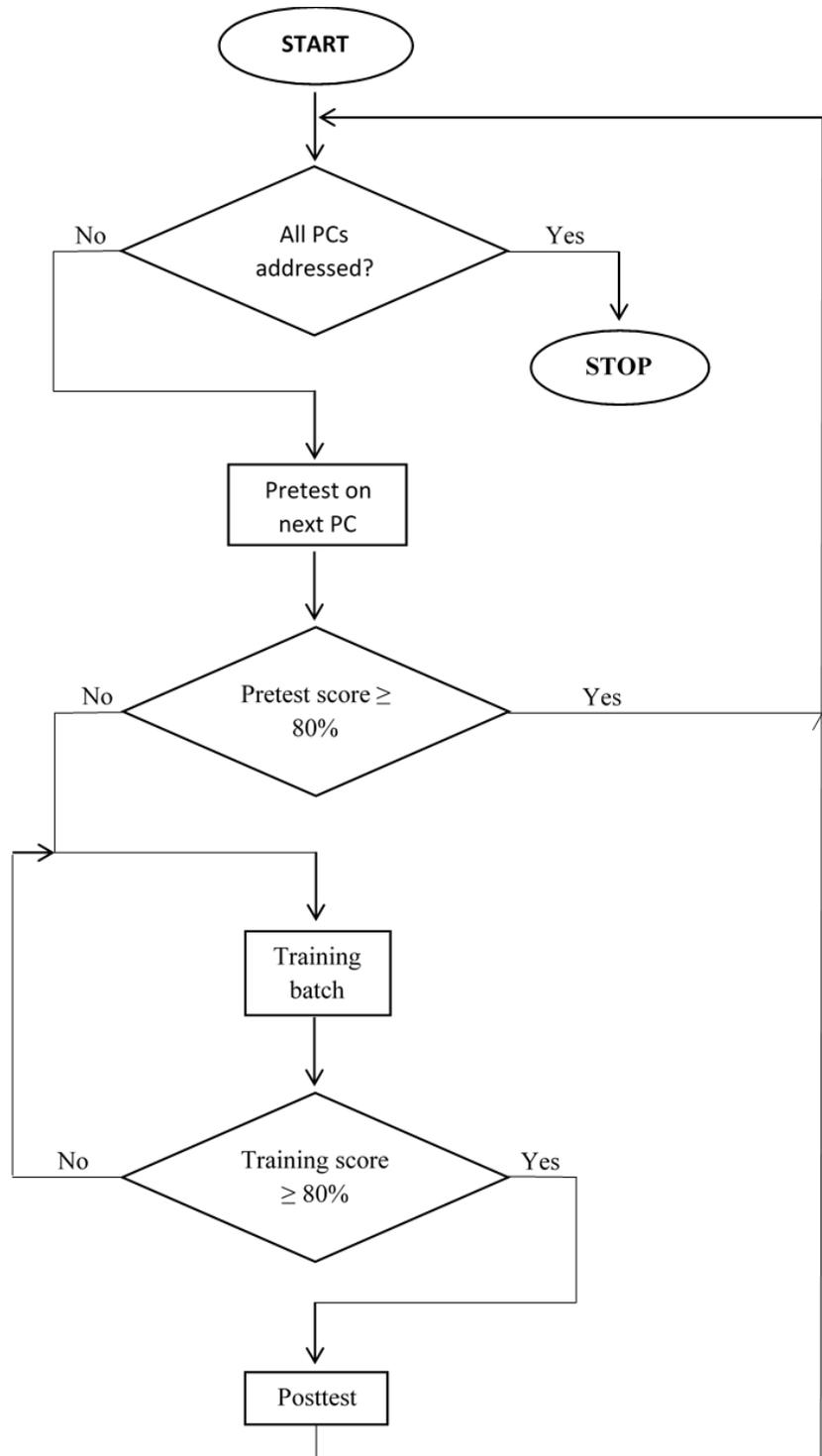


Figure 3. Adaptive learning algorithm (PC = phonemic contrast)

For each phonemic contrast, an array of seven minimal pairs was chosen for test and training stimuli. These minimal pairs were extracted from the [Illinois Speech and Language Engineering Dictionary](#) with word frequency and syllable environment controlled. We filtered the minimal-pair selection through the top 5,000 most frequently used lemmata in COCA to ensure that the selected minimal pairs were frequent and thus more likely to be familiar to learners in general. The decision to select seven minimal pairs for

each phonemic contrast was also based on the number of minimal pairs available in the COCA lemma list. The minimal pairs were randomly divided into three sets (see [Appendix C](#)); some were used during training and some not. Specifically, the participants were given training on the words in Set A and Set B, but not on the words in Set C, to examine whether the participants could transfer gains from trained to untrained words. In addition, different voices were used to create the audio for each word set: the voices for Set A and Set C were the same as the voices used for training, whereas Set B was recorded with voices that were new to the participants. The intention behind this setup was to examine whether the participants could extend changes in perceptual ability from familiar voices to new voices.

Audio stimuli included recordings of all the selected minimal pairs listed in [Appendix C](#). Four different voices were adopted for creating these stimuli, but only two voices were used for creating the stimuli used during training. The other two voices were reserved for test stimuli only. The two voices used for creating training stimuli included a female voice and a male voice and were generated using TTS, a technology that translates text automatically into soundwaves that represent speech (Delmonte, 2008). High-quality TTS was used in this study, provided by an industry partner indicated in the Acknowledgements section. The two natural voices used for recording Set B of the test stimuli were from two native speakers of standard North American English, one female and one male. The natural-voice stimuli were recorded with a digital voice recorder at a bit rate of 128 kbps and sampling frequency of 44.1kHz.

System Evaluation

The evaluation aspect of the study is situated within the framework proposed by Levis (2007), which highlights four criteria as guidance for assessing the effectiveness of CAPT applications. The criteria can be summarized as efficacy, transfer, retention, and spillover. This study focuses specifically on the first two criteria: (a) learner improvement on discriminating and identifying trained phonemes, and (b) learners' transfer of perceptual gains to untrained contexts. The examination of transfer, defined as the application of previously acquired knowledge in one setting to a different setting (Gagne, Yekovich, & Yekovich, 1993), is especially important because successful transfer is integral to robust learning (Logan & Pruitt, 1995) and to the end goal that education should pursue (MacKeough, Lupart, & Marini, 1995). In this study, we addressed two levels of transferring phonemic perceptual skills: (a) transfer from known voices to new voices, and (b) transfer from trained words to new words. The motive for assessing the first level of transfer stemmed from the notion that speech variability tends to interfere with learners' perception capacity; only when trainees are able to accurately perceive trained sounds articulated by unfamiliar voices can the training be regarded successful. The second level of transfer was inspected because the acoustic characteristics of a phoneme vary depending on the surrounding phonetic environments (Strange, Weber, Levy, Shafiro, & Nishi, 2002) and because the ability to perceive a phoneme in one phonetic context does not necessarily translate to other phonetic contexts (Thomson, 2012). In sum, the empirical research questions for this article are as follows:

RQ1: How effective is the training design in improving perception of trained phonemes?

RQ2: How well does the improvement generalize to untrained voices?

RQ3: How well does the improvement generalize to novel words?

Methodology

Participants

The study involved 32 native-Russian participants, 31 university-level students and one female English instructor aged 42. The 31 students included 9 females and 22 males between the ages of 20 and 23 ($M = 21.2$ years), majoring in either marine engineering or English translation from two EFL schools in Russia. The training was self-paced and completed outside of class time. The participants were also in control of

when and where to work on the phonetic exercises. Overall, the duration of the study for the participants ranged from 10 to 100 minutes, with the mean total time commitment per person being 70 minutes. The training content differed for each participant, since their phonemic problems varied in number.

Pre-Test

As illustrated in [Figure 3](#), the participants were first given a pre-test for each sound pair. The pre-test consisted of four types of test items presented to the participants in the following order: (1) same-different discrimination, (2) oddity discrimination, (3) simple identification, and (4) yes/no identification. As there were seven minimal pairs and four exercise types tied to each phonemic contrast, the total number of pre-test items per sound pair was 28. Among the items, eight audio files were by human speakers (for items in Set B) and the other 20 were by synthetic voices (for items in Sets A and C). Students scored 1 point for a correct response to a test item and 0 points for an incorrect or don't know response. With an arbitrarily decided accuracy rate of 80% or better, students would be subsequently directed to the pre-test on a different sound pair. Otherwise, they would be provided with training on the problematic sounds. As a result, each student received training on only those sound pairs with which they demonstrated perception difficulties.

Training

Training for each phonemic contrast was provided to students in batches of exercises. The number of training exercises in each batch was five, with each exercise targeting a single minimal pair. Two out of seven minimal pairs per phonemic contrast were withheld from training ([Appendix C](#)) to later test for transfer to untrained items, and only one type of exercise was offered to a specific student for each phonemic contrast. The assignment of exercise type to phonemic contrast was randomized by the computer. Synthetic voices were used for all the training exercises given to the participants. The gender of the voice was also randomly selected by the computer for each exercise item.

The quantity and length of training varied according to a student's training performance, which was evaluated automatically at the end of each set of training. If the cutoff score (80%) was reached, trainees exited training on the phonemic contrast; if not, they would be given another set of training exercises on the same phonemic contrast with the type of exercise previously determined by the computer program. (Exercise type was controlled to examine its effect on training, but it is not reported in this article.) That is to say, the training duration not only varied by student but also by phonemic contrast for each student. Such a setup, compared with many previous HVPT studies where training duration was either completely subject-dominated (e.g., Wang & Munro, 2004) or researcher-dominated (e.g., all other HVPT studies), was more likely to utilize trainee's time to the best advantage by focusing on the aspects that needed the most attention. On the other hand, as a by-product of this training model, there was a wide range of time on task, which added an extra independent variable that was not controlled for. The incorporation of immediate assessment also reduced the risk of having trainees rely on self-judgment to decide when to discontinue training.

Although the exercise types and the word bank for generating exercises were the same for each participant, the training content presented to students in each round was still unlikely to be the same because the system retrieved word entries for each exercise independently and randomly. For each lexical unit shown on the screen, the system would first randomly select a word from the minimal pair and then randomly designate a female or male voice to go with the word. For instance, to generate a same-different discrimination exercise for the word pair *beat* and *bit*, the system could produce any of the 16 possibilities listed in [Table 1](#). The motive for randomizing these selections was to make the training items diverse and less predictable as students looped through multiple training sets.

Table 1. Possible Presentations of Same–Different Discrimination Exercises for Beat–Bit

Word Combination	Voice Combination			
	F, F	M, M	F, M	M, F
<i>beat–beat</i>	Exercise 1	Exercise 2	Exercise 3	Exercise 4
<i>bit–bit</i>	Exercise 5	Exercise 6	Exercise 7	Exercise 8
<i>beat–bit</i>	Exercise 9	Exercise 10	Exercise 11	Exercise 12
<i>bit–beat</i>	Exercise 13	Exercise 14	Exercise 15	Exercise 16

Note. *F* = female voice, *M* = male voice

Post-Test

Students were given a post-test only on the phonemic contrasts they received training on. The post-test was the same as the pre-test and was administered by the training system automatically and immediately following a student's completion of training on a sound pair. Post-tests included trained words with trained voices, trained words with untrained voices, and untrained words with trained voices.

Data Analysis

The data to answer the research questions included the participants' scores on each pre- and post-test item. These scores were collected by the system in an automatic and de-identified manner along the participants' use of the system. As seen from the training algorithm (Figure 3), each participant received training that was tailored to his or her individual difficulties as assessed by the pre-test. As the participants received a pre-test score and a post-test score on each trained sound pair, the averages of the pre-test and post-test scores were used for data analysis. Table 2 shows the statistical approaches adopted for data analysis in response to each research question. All statistical analyses were conducted using SPSS. It should be noted that the scores used for analysis were normalized to the scale [0, 1].

Table 2. Statistical Approaches for Data Analysis

Research Question	Descriptive Statistics	Inferential Statistics
1 Training effectiveness	Differences in average pre- and post-test scores on Set A exercises	Matched pairs <i>t</i> -test
2 Transfer to new voices	Differences in average pre- and post-test scores on Set B exercises	Matched pairs <i>t</i> -test
3 Transfer to new words	Differences in average pre- and post-test scores on Set C exercises	Matched pairs <i>t</i> -test

Results

RQ1. Training Effectiveness

Participants' pre-test and post-test scores, averaged across sound pairs on the 12 items in Set A (both words and voices trained), were compared to examine how the training facilitated the participant's perceptual skills for discriminating and identifying trained words recorded with trained voices. Table 3 lists the difference in each participant's pre-post test scores. The difference was positive for a majority of the participants (19 out of 32), negative for 12 participants, and zero for one participant.

Table 3. *Performance on Trained Words Spoken by Trained Voices*

Student	Pre-Test Score	Post-Test Score	Gain
St10	0.92	0.67	-0.25
St17	0.84	0.63	-0.21
St5	0.71	0.59	-0.12
St24	0.67	0.58	-0.09
St32	0.67	0.58	-0.09
St30	0.75	0.67	-0.08
St29	0.83	0.78	-0.05
St31	0.80	0.75	-0.05
St16	0.67	0.64	-0.03
St11	0.80	0.78	-0.02
St13	0.75	0.73	-0.02
St26	0.49	0.47	-0.02
St27	0.71	0.71	0.00
St22	0.72	0.76	0.04
St7	0.69	0.75	0.06
St1	0.92	1.00	0.08
St14	0.68	0.77	0.09
St20	0.83	0.92	0.09
St21	0.71	0.80	0.09
St23	0.58	0.67	0.09
St4	0.67	0.77	0.10
St8	0.75	0.86	0.11
St9	0.61	0.72	0.11
St28	0.79	0.92	0.13
St3	0.64	0.78	0.14
St12	0.55	0.70	0.15
St18	0.75	0.92	0.17
St25	0.58	0.75	0.17
St2	0.59	0.79	0.20
St15	0.63	0.83	0.20
St6	0.67	0.92	0.25
St19	0.33	0.67	0.34

The participants' average pre-test scores (skewness = -0.69, $SE = 0.41$; kurtosis = 1.67, $SE = 0.81$) and post-test scores (skewness = -0.01, $SE = 0.41$; kurtosis = 0.19, $SE = 0.81$) were close to a normal distribution; a paired-samples t -test was conducted to statistically compare the participants' test performance on Set A before and after training. A significant difference was found in the scores for pre-test ($M = 0.70$, $SD = 0.12$) and post-test ($M = 0.75$, $SD = 0.12$), $t(31) = -2.13$, $p = .041$. The effect size of

the improvement was medium (Cohen's $d = .412$), which, according to Coe's (2002) effect-size-to-percentile-interpretation table, means that the average trainee in the study would score higher than 66% of students who were initially equivalent but not trained. These results show a statistically significant improvement in the participants' performance on all items in Set A from pre-test to post-test. This means that the training was effective in enhancing the participants' ability to perceptually discriminate and identify the target phonemic contrasts from word pairs they had received training on.

RQ2. Transfer to New Voices

To find out whether the training promoted transfer of perceptual discrimination and identification abilities from trained voices to untrained voices, participants' pre-test and post-test scores averaged across sound pairs on the eight items in Set B (trained words, new voices) were compared. Table 4 lists the difference between each participant's pre-test and post-test scores on Set B. The difference was positive for a vast majority of the participants (23 out of 32), negative for three participants, and zero for the six others.

Table 4. Performance on Trained Words Spoken by Untrained Voices

Student	Pre-Test Score	Post-Test Score	Gain
St5	0.69	0.51	-0.18
St3	1.00	0.92	-0.08
St14	0.70	0.63	-0.07
St10	0.88	0.88	0.00
St17	0.57	0.57	0.00
St19	0.63	0.63	0.00
St25	1.00	1.00	0.00
St31	0.63	0.63	0.00
St1	0.75	0.75	0.00
St13	0.66	0.69	0.03
St22	0.67	0.71	0.04
St26	0.54	0.59	0.05
St8	0.60	0.66	0.06
St15	0.63	0.69	0.06
St20	0.63	0.69	0.06
St28	0.69	0.75	0.06
St4	0.72	0.80	0.08
St12	0.70	0.78	0.08
St29	0.67	0.75	0.08
St2	0.78	0.88	0.10
St21	0.69	0.80	0.11
St9	0.67	0.79	0.12
St24	0.63	0.75	0.12
St27	0.57	0.69	0.12
St18	0.54	0.67	0.13
St30	0.75	0.88	0.13

St16	0.67	0.84	0.17
St7	0.67	0.88	0.21
St6	0.75	1.00	0.25
St11	0.63	0.88	0.25
St23	0.63	0.88	0.25
St32	0.38	0.63	0.25

The participants' average pre-test scores (skewness = 0.80, $SE = 0.41$; kurtosis = 2.59, $SE = 0.81$) and post-test scores (skewness = 0.13, $SE = 0.41$; kurtosis = -0.57, $SE = 0.81$) were approximate to normal distributions; a paired-samples t -test was conducted to statistically compare the participants' test performance on Set B before and after training. A significant difference was found in the scores for pre-test ($M = 0.68$, $SD = 0.12$) and post-test ($M = .76$, $SD = 0.12$), $t(31) = -4.29$, $p = .000$. The effect size of the score improvement ranged between medium and large (Cohen's $d = .624$), which, if converted to percentiles (Coe, 2002), means that the average trainee in the study would now score higher than 73% of students who were initially equivalent but not trained. The statistically significant improvement in the participants' performance suggests that the training successfully facilitated the participants to generalize perceptual gains from trained voices to untrained voices.

To investigate whether the difference in gain scores between the two sets was statistically significant, a paired-samples t -test was conducted to compare each participant's pre- and post-test improvement scores on Set A and Set B, both of which followed normal distribution. No significant difference was found in the gain scores between the two sets (Set A, $M = 0.05$, $SD = 0.13$; Set B, $M = 0.08$, $SD = 0.10$), $t(31) = -1.00$, $p = .326$, Cohen's $d = .242$. This suggests that the training facilitated similar perceptual acquisition of words spoken by trained voices and untrained voices.

RQ3. Transfer to New Words

To find out whether the training promoted transfer of perceptual discrimination and identification abilities from trained words to untrained words, participants' pre-test and post-test scores averaged across sound pairs on the eight items in Set C (trained voices, new words) were compared. The participants' average pre-test scores were close to normal distribution (skewness = 0.59, $SE = 0.41$; kurtosis = 0.61, $SE = 0.81$). The participants' average post-test scores were normally distributed (skewness = 0.02, $SE = 0.41$; kurtosis = -0.28, $SE = 0.81$). Table 5 lists that approximately half of the participants (14 out of 32) had a positive gain score. A decrease in score from pre-test to post-test was seen with 10 participants; the scores stayed unchanged for eight participants.

Table 5. Performance on Untrained Words Spoken by Trained Voices

Student	Pre-Test Score	Post-Test Score	Gain
St19	0.75	0.38	-0.37
St32	0.75	0.38	-0.37
St17	0.63	0.44	-0.19
St7	0.67	0.54	-0.13
St31	0.63	0.50	-0.13
St24	0.50	0.38	-0.12
St27	0.69	0.63	-0.06
St26	0.57	0.52	-0.05
St13	0.85	0.81	-0.04

St14	0.75	0.73	-0.02
St5	0.63	0.63	0.00
St10	0.50	0.50	0.00
St15	0.75	0.75	0.00
St21	0.78	0.78	0.00
St23	1.00	1.00	0.00
St25	0.63	0.63	0.00
St28	0.82	0.82	0.00
St30	0.50	0.50	0.00
St2	0.69	0.72	0.03
St8	0.66	0.69	0.03
St11	0.71	0.75	0.04
St22	0.73	0.80	0.07
St4	0.66	0.77	0.11
St9	0.63	0.75	0.12
St18	0.50	0.63	0.13
St29	0.54	0.67	0.13
St1	0.50	0.63	0.13
St16	0.50	0.63	0.13
St12	0.60	0.78	0.18
St20	0.50	0.69	0.19
St6	0.75	1.00	0.25
St3	0.63	0.92	0.29

A paired-samples *t*-test was conducted to statistically compare the participants' test performance on Set C before and after training. No significant difference was found in the scores for pre-test ($M = 0.66$, $SD = 0.12$) and post-test ($M = 0.67$, $SD = 0.16$), $t(31) = -4.23$, $p = .675$. The effect size was found to be very small (Cohen's $d = .076$), indicating no difference between the two sets of scores. These results indicate that the training failed to facilitate the transfer of perceptual gains from trained to untrained words.

Discussion

Analyses of pre- and post-training data showed the efficacy of the training model in enhancing the participants' perception of segmental contrasts. The participants were also able to generalize the perceptual improvement from words spoken with trained voices to the words spoken with untrained voices. Although the improvement was only medium level (Cohen's $d = .412$ for Set A, Cohen's $d = .624$ for Set B), the training effect was practically important, considering that the training efforts were commensurate with learner achievement. The trainees' entire time investment in the study, including involvement in the diagnostic and post-test, was only 70 minutes on average, shorter than two regular college-level class periods. In contrast, the pre- to post-test improvement, albeit moderate, can serve as evidence that the training was worthwhile.

Changes in learner performance through training varied across subjects as demonstrated in [Table 3](#), [Table 4](#), and [Table 5](#), suggesting that the patterns of perceptual acquisition may be learner-specific. Further analyses of the participants' performance revealed wide learner-level variation pertaining to training

effect (Set A), transfer to voices (Set B), and transfer to words (Set C). For example, Table 6 displays each participant's performance on test items in Set A (words trained, voices trained) on phonemic contrasts they received training on. Nine participants improved on all trained sound pairs, whereas nine others failed to improve on any of the trained sound pairs. Nevertheless, these patterns are not generalizable because the average number of trained phonemic contrasts was 3.13 ($SD = 2.50$), and 28% of the students (9 out of 32) received training on only one sound pair. Despite this, the drastic variations among trainees regarding their pre- to post-training gains might be evidence that perception learning is a unique process shaped by individual learner characteristics.

It should be noted that learners' major of study was not investigated in the study due to a lack of access to the necessary demographic information, but analysis on this variable could be meaningful by revealing potential patterns among the individual-level variations. Assumptions are that English translation majors were more likely to improve than marine engineering students due to different levels of motivation.

Table 6. *Each Participant's Pre- to Post-Training Gains on Test Items in Set A*

Subject	Number of Contrasts Trained	Contrasts Showing Positive Gain	Contrasts Showing No Gain	Contrasts Showing Negative Gain
St10	1	0%	0%	100%
St17	2	0%	0%	100%
St24	1	0%	0%	100%
St30	1	0%	0%	100%
St32	1	0%	0%	100%
St29	3	0%	33%	67%
St5	2	0%	50%	50%
St31	2	0%	50%	50%
St16	3	0%	67%	33%
St7	3	33%	67%	0%
St11	3	33%	33%	33%
St22	6	33%	50%	17%
St26	12	42%	17%	42%
St2	4	50%	25%	25%
St8	4	50%	25%	25%
St13	4	50%	0%	50%
St20	2	50%	50%	0%
St27	2	50%	0%	50%
St14	5	60%	20%	20%
St4	8	63%	25%	13%
St21	8	63%	25%	13%
St9	3	67%	33%	0%
St12	5	80%	0%	20%
St3	3	100%	0%	0%
St6	1	100%	0%	0%

St15	2	100%	0%	0%
St18	3	100%	0%	0%
St19	1	100%	0%	0%
St25	1	100%	0%	0%
St28	2	100%	0%	0%
St1	1	100%	0%	0%
St23	1	100%	0%	0%

In addition to inter-subject variation, the effect of training on the participants' performance on trained items (Set A), untrained voices (Set B), and untrained words (Set C) also differed based on phonemic contrasts. Table 7 shows trainees' performance on each phonemic contrast for test items in Set A (words trained, voices trained). The sound pairs exhibited different levels of acquisition difficulty. Trainees' average pre- to post-training gains on the phonemic contrasts ranged from -0.04 (/ɛ-ɜ/) to 0.19 (/ɛ-ɪ/). For certain phonemic contrasts (i.e., /əʊ-ɜ/, /d-t/, /ɛ-ɪ/, /g-k/), all or most trainees showed improvement with nobody showing negative gain. However, for some other contrasts (i.e., /ɛ-ɜ/, /i-ɪ/, /æ-ʌ/, /æ-ɛ/), only a minority of trainees improved through training and the proportion of trainees with decreased post-test scores was relatively high. This may be an indicator that some phonemic contrasts are easier to acquire than others. However, this hypothesis should be examined in future research, because the number of trainees in the study for many of the sound pairs (e.g., /ɛ-ɜ/, /æ-ʌ/, /əʊ-ɜ/, and /g-k/) was small.

Table 7. Participants' Pre- to Post-Training Gains on Each Phonemic Contrast for Set A Items

Phonemic Contrast	Number of Subjects	Mean Gain Score	Subjects With Positive Gain	Subjects With No Gain	Subjects With Negative Gain
/ɛ-ɜ/	2	-0.04	0%	50%	50%
/i-ɪ/	22	-0.01	32%	32%	36%
/æ-ʌ/	3	0.113	33%	33%	33%
/æ-ɛ/	28	0.023	43%	25%	32%
/ɑ-ʌ(ə)/	8	0.01	50%	13%	38%
/s-θ/	2	0.00	50%	0%	50%
/t-θ/	2	0.05	50%	0%	50%
/ɔ-əʊ/	9	0.06	56%	11%	33%
/əʊ-ɜ/	3	0.14	67%	33%	0%
/d-t/ final	11	0.18	82%	18%	0%
/ɛ-ɪ/	6	0.19	83%	17%	0%
/g-k/	4	0.13	100%	0%	0%

Putting together the analyses thus far, we can infer that phonetic acquisition is a process unique to each individual learner and shaped by the specific sounds being learned. Further evidence comes from an unclear correlation between the amount of training participants had received and the amount of gain they achieved through training. Table 8 lists the participant's holistic gain scores on trained items based on a descending order of the number of training batches. The data show no obvious correlation between the two variables. In general, the phonemic contrasts on which participants were trained more intensely (e.g., /i-ɪ/) were not necessarily associated with higher gains, and vice versa (e.g., /ɛ-ɪ/). Table 9 displays the training quantity for students who did not improve on any sound pairs in contrast with students who consistently improved on all trained sound pairs. Again, no clear patterns were discovered between

training amount and achievement. Students with one or multiple batches of training were seen in both categories. Learner achievement also exhibited variations even with the same amount of training on the same phonetic contrast, suggesting that phonetic acquisition is indeed learner-specific. For instance, one batch of training on /æ-ɛ/ led to improvement for St1, St3, St18, and St23, but not for St30. This was also a case with /i-ɪ/ for St10, St17, and St18. While the data do not show any generalizable pattern between training intensity and achievement, they do suggest that some students were faster than others in acquiring certain sounds. For example, for the contrast /æ-ɛ/, there were students (i.e., St1, St3, St18, St23, St25, St28) who improved after three or fewer training batches, but there were also students (i.e., St24, St32) who failed to improve after four or five batches.

Table 8. *Pre- to Post-Training Gain in a Descending Order by Training Quantity for Set A Test Items*

Phonemic Contrast	Average Training Batches per Student	Number of Trainees	Average Gain per Student
/ɛ-ɜ/	6.0	2	-0.04
/æ-ʌ/	5.3	3	0.12
/t-θ/	4.5	2	0.05
/əʊ-ɜ/	4.3	3	0.14
/i-ɪ/	3.3	22	-0.01
/ɔ-əʊ/	2.7	9	0.06
/ɑ-ʌ(ə)/	2.3	8	0.01
/æ-ɛ/	2.2	28	0.02
/g-k/	1.5	4	0.13
/d-t/ final	1.3	11	0.18
/ɛ-ɪ/	1.0	6	0.19
/s-θ/	1.0	2	0.00

Table 9. *Number of Training Batches for 100% Positive-Gain Students and 100% Negative-Gain Students*

Category	Student	Phonemic Contrast	Pre-post Training Gain (Set A Test Items)	Training Batches
Students showing no gains on any trained sound pair	St10	/i-ɪ/	-0.25	1
	St17	/i-ɪ/	-0.17	1
	St17	/æ-ɛ/	-0.25	2
	St24	/æ-ɛ/	-0.09	4
	St30	/æ-ɛ/	-0.08	1
	St32	/æ-ɛ/	-0.09	5
Students showing gains on every trained sound pair	St3	/d-t/ final	0.08	1
	St3	/g-k/	0.25	1
	St3	/æ-ɛ/	0.08	1
	St6	/d-t/ final	0.25	1
	St15	/d-t/ final	0.33	1
	St15	/i-ɪ/	0.08	2

St18	/i-ɪ/	0.17	1
St18	/æ-ɛ/	0.08	1
St18	/ɑ-ʌ(ə)/	0.25	1
St19	/d-t/ final	0.34	1
St25	/æ-ɛ/	0.17	3
St28	/æ-ɛ/	0.17	3
St28	/ɑ-ʌ(ə)/	0.08	5
St1	/æ-ɛ/	0.08	1
St23	/æ-ɛ/	0.09	1

No statistically significant effect was found in terms of the transfer of perceptual gains from trained words to untrained words, which echoes the widely accepted belief that L2 segmental acquisition can be highly sensitive to the linguistic context of a segment (Flege, 1995; Jamieson & Morosan, 1986; Munro et al., 2015; Thomson, 2011, 2012) because the acoustic characteristics of a phoneme can be affected by its surrounding phonetic and lexical environments (Munro & Derwing, 2008; Pisoni & Lively, 1995; Walley & Flege, 1999). Possibly, the effect of lexicon on phonological acquisition observed in the study could be accounted for by three important and interrelated linguistic theories that predict language acquisition behavior: the exemplar theory (Bybee, 2000), the analogical modeling theory (Skousen, 1989), and the TRACE model of auditory word recognition within the connectionist framework (Joanisse & McClelland, 2015). According to these theories, the categorization of novel linguistic stimuli occurs through comparisons of the stimuli with items already-stored in memory (as *exemplars*). While the first two theories relate to language learning in general, the TRACE model was proposed specifically for the decoding of speech input and is characterized by its dynamic and interactive nature in auditory language processing. The dynamism of the model posits that the acoustic input of a word is disassembled in a time-varying manner into units that are subsequently interpreted *in parallel* (as opposed to *serially*; see Joanisse & McClelland, 2015). The interactivity of the model suggests that linguistic input is processed from dual directions, bottom-up and top-down, between words and phonemes—within the top-down dimension, the spoken input of a word can be stored as an entity or several major sub-entities rather than independent units divided at the level of phoneme. Specific to the participants' experience in the study, their inability to generalize gains to novel words may be a result of the top-down effects in auditory input processing, since the trainees' accumulation of perceptual representations of trained words could potentially be established from the unsegmented speech stream at the lexical or sub-lexical level. Such knowledge, after being stored in the memory, became what the participants would later refer to upon receiving auditory input of untrained words. However, because the new stimuli were environmentally and lexically different from the stored exemplars, the learners encountered difficulties recognizing the new items.

The absence of phonological transfer at the word level may also be a sign that the training triggered only one of the two stages involved in language learning: *item learning* but not *system learning*. This two-stage distinction was initially made by Cruttenden (1981) to explain the developmental process of L1 learning and then extended to foreign language learning (Ellis, 1999; Ringbom, 1983). According to Cruttenden (1981), item learning, a prerequisite for system learning, “involves a form which is uniquely bonded with some other form or with a unique referent, whereas system-learning involves the possibility of the commutation of forms or referents while some (other) form is held constant” (p. 79). That is to say, most learners first learn items (e.g., words) as single entities by imitation and memory. Only later can they begin to realize that the items are in fact composed of discrete units which can be independently used with other units and form new items. Once learners build the capacity to decode the system of language, they are able to recognize (and perhaps produce) a linguistic unit despite potential apparent changes to its surroundings. In this study, the participants failed to discriminate and identify trained phonemic

distinctions embedded in new words, meaning that the participants did not yet exhibit recognition of the contrasts in novel linguistic contexts. In order to facilitate the transition to the second stage of learning, two changes can potentially be implemented to the training model. One is to intensify the training with longer training batches and higher training accuracy thresholds. The second is to enrich the training stimuli by exposing learners to a greater variety of phonemic variations through more lexical stimuli, including ones of lower frequency. An intensified, enriched training model may promote transition, since continual exposure to and accumulation of a target item used in different situations are catalysts for the transition from item learning to system learning (Cruttenden, 1981).

Perhaps some phonological guidance alongside phonetic training may also promote system learning by familiarizing learners with the certain predictable variability in the acoustic features of phonemes depending on contextual lexical variables. The provision of such guidance, ideally, should be based on L1 customization and adapted to error profiles of individual learners.

The generalization to new voices but not to new words raises the question of whether the acoustic properties of a phoneme are more easily modified by its linguistic environments than the inherent variability of multi-talker voices. The results also could be an indicator that the nature of the synthetic voices adopted in the study resembles human speech. While these inquiries can be pursued in the future, this study provides evidence that HVPT is effective for sharpening aural sensitivity to the variability in speaker voices and that TTS technology holds promise for being widely utilized for developing auditory materials for language learning. The efficacy of HVPT warns against the presently dominating pedagogical practices built on a single normative speech variety due to assumptions such as fear for learners' comprehension, while lending significance to materials and tools that incorporate diverse speaker models. The potential of TTS technology to facilitate research and application of HVPT is also worth exploiting, as TTS is more cost-effective compared with human speech and allows for more efficient manipulation of variables such as speech rate and voice model, yielding a larger quantity and variety of speech output (Delmonte, 2008; Handley, 2009; Sha, 2010).

Conclusion

In this study, we experimented with providing perceptual training on 12 phonemic contrasts to a group of Russian-speaking English learners using a novel, prototype HVPT system. Our findings demonstrate that the prototype was effective in enhancing the learners' perception of phonemic contrasts spoken by trained and untrained voices. The results lead us to believe that the prototype system, which is learner-oriented, flexible, efficient, varied in its input supply, and built to bridge research and practice, can be put forth as a perceptual training model for replication in language teaching and research settings. The ability of the system to allow individualization of high FL contrasts will improve learner efficiency in perceiving high-value sounds and also have practical implications. For the many language teaching centers where limited in-class time is available for pronunciation instruction, the training system can be a substitute instructor for students who struggle with the identification of vowels and consonants. The development of the system's capacities, an example of successful integration of technology into pronunciation instruction, also shows that pedagogy can be effectively enhanced by computational approaches.

The current study can be extended in a few dimensions. For instance, the study did not explore whether the participants would be able to generalize their perceptual improvement beyond the level of isolated words, which is a critical criterion for evaluating any CAPT system, since the ultimate goal of perceptual training is to increase the ability to comprehend utterances. Long-term retention, proposed by Levis (2007) as an important criterion for evaluating CAPT systems, was not investigated in the study. Another area worthy of exploration is the potential spillover effects of the training model. In light of the promoting effect of segmental perceptual training on articulation of trained sounds (Bradlow et al., 1997; Lambacher et al., 2005; Thomson, 2011), a potential spillover effect of the training model that can be examined is its capacity to bring about gains on untrained skills, such as production of the trained phonemes. Apart from the quantitative measures of training effect, learner experience and attitudes are worth exploring because

they can further unfold the potentials and limitations of the system and inform future upgrades of the system. An upgraded system should be able to adapt not only to learners within one L1 but also to multiple L1s. A good way to realize this would be to develop the system with phonemic lists encompassing errors for a variety of L1s and later direct learners to different diagnostics based on their L1s. This adaptability at a broader level would ultimately give the system more practical significance, particularly in ESL contexts where learners generally come from mixed L1s. Last but not least, it should be noted that context, as a component integral to the ultimate goal of language acquisition, was completely missing from the current training. Balancing form against context has been a longstanding problem for pronunciation materials development. Minimal sentences (Bowen, 1972) were proposed as a technique to mitigate the lack of meaningfulness and context of word-level drills, but the construction of such sentences (and potentially other forms of materials that are beyond the single-word level and that highlight specific phonemic forms) has been limited by the availability of phonemic contrasts belonging to the same parts-of-speech category (Levis & Cortes, 2008). Future developers are encouraged to try embedding minimal sentences as much as possible to the system and to seek for potential alternatives to addressing the issue.

Acknowledgements

The authors would like to thank the Andrey A. Hudyakov Center for Linguistic Research and Professor Tatiana A. Klepikova for assistance with software development and implementation and for the handling and anonymization of research data. We are also grateful to [iSpeech, Inc.](#) for providing the TTS technology for this study.

Notes

1. Phonetic context was not controlled for in testing and training materials development. We assumed the inclusion of multiple phonetic contexts was a characteristic for high-variability input treatment and thus a natural part of the treatment the participants received.

References

- Bahanshal, D. A. (2013). The effect of large classes on English teaching and learning in Saudi secondary schools. *English Language Teaching*, 6(11), 49–59.
- Baker, W., & Trofimovich, P. (2001). Does perception precede production? Evidence from Korean–English bilinguals. *The LACUS Forum*, 27, 273–284.
- Best, C. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Timonium, MD: York Press.
- Best, C., & MacRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech*, 46(2–3), 183–216.
- Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, 19, 447–465.
- Bowen, J. D. (1972). Contextualizing pronunciation practice in the ESOL classroom. *TESOL Quarterly*, 83–94.
- Bradlow, A. R. (2008). Training non-native language sound patterns: Lessons from training Japanese adults on the English /ɹ /-/l/ contrast. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology in Second Language Acquisition* (pp. 287–308). Amsterdam, Netherlands: John Benjamins.

- Bradlow, A. R., Pisoni, D. B., Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, 101(4), 2299–2310.
- Bradlow, A. R., Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception and Psychophysics*, 61(5), 977–985.
- Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly*, 22(4), 593–606.
- Bybee, J. (2000). Lexicalization of sound change and alternating environments. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology (vol. 5): Acquisition and the lexicon* (pp. 250–268). New York, NY: Cambridge University Press.
- Carlet, A., & Cebrian, J. (2014). Training Catalan speakers to identify L2 consonants and vowels: A short-term high variability training study. *Concordia Working Papers in Applied Linguistics*, 5, 85–98.
- Catford, J. C., & Pisoni, D. (1970). Auditory versus articulatory training in exotic sounds. *Modern Language Journal*, 54, 477–481.
- Chukharev-Hudilainen, E., & Klepikova, T. A. (2016). The effectiveness of computer-based spaced repetition in foreign language vocabulary instruction: A double-blind study. *CALICO Journal*, 33(3), 334–354.
- Chun, D. M., Jiang, Y., Meyr, J., & Yang, R. (2015). Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations. *Journal of Second Language Pronunciation*, 1(1), 86–114.
- Coe, R. (2002). *It's the effect size, stupid: What effect size is and why it is important*. Paper presented at The 2002 Annual Conference of the British Educational Research Association, Exeter, UK.
- Coniam, D. (2002). Technology as an awareness-raising tool for sensitising teachers to features of stress and rhythm in English. *Language Awareness*, 11(1), 30–42.
- Cruttenden, A. (1981). Item-learning and system-learning. *Journal of Psycholinguistic Research*, 10(1), 79–88.
- Davies, M. (2008). *The Corpus of Contemporary American English: 450 million words, 1990–2012*. Retrieved from <http://corpus.byu.edu/coca>
- Delmonte, R. (2008). Speech synthesis for language tutoring system. In M. Holland & P. Fisher (Eds.), *The path of speech technologies in computer-assisted language learning: From research towards practice* (pp. 123–150). New York, NY: Routledge.
- Denes, P., & Pinson, E. (1963). *The speech chain: The physics and biology of spoken language*. New York, NY: Bell Telephone Laboratories.
- Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42(4), 476–490.
- Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam, Netherlands: John Benjamins.
- Detey, S., & Racine, I. (2015). Does perception precede production in the initial stage of French nasal vowel quality acquisition by Japanese learners? A corpus-based discrimination experiment. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Retrieved from <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0894.pdf>

- Ducate, L., & Lomicka, L. (2009). Podcasting: An effective tool for honing language students' pronunciation? *Language Learning & Technology*, 13(3), 66–86. <https://dx.doi.org/10125/44192>
- Ellis, R. (1999). Item versus system learning: Explaining free variation. *Applied Linguistics*, 20(4), 460–480.
- Flege, J. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.
- Flege, J., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics*, 34(2), 153–175.
- Flege, J., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25(4), 437–470.
- Gagne, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). *The cognitive psychology of school learning* (2nd ed.). New York, NY: HarperCollins College.
- Gick, B., Bernhardt, B. M., Bacsfalvi, P., & Wilson, I. (2008). Ultrasound imaging applications in second language acquisition. In J. Hansen & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 309–322). Amsterdam, Netherlands: John Benjamins.
- Golestani, N., & Pallier, C. (2007). Anatomical correlates of foreign speech sound production. *Cerebral Cortex*, 17(4), 929–934.
- Handley, Z. (2009). Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication*, 51(10), 906–919.
- Hardison, D. M. (2003). Acquisition of second language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, 24, 495–522.
- Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, 8(1), 34–52. <https://dx.doi.org/10125/25228>
- Hardison, D. M. (2005). Second language spoken word identification: Effects of training, visual cues, and phonetic environment. *Applied Psycholinguistics*, 26, 579–596.
- Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL*, 15, 3–20.
- Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, 33(4), 575–591.
- 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.
- Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, 53, 298–310.
- Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *Journal of the Acoustical Society of America*, 126, 866–877.
- Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *Journal of the Acoustical Society of America*, 118, 3267–3278.
- Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/contrast by francophones. *Perception & Psychophysics*, 40(4), 205–215.

- Jamieson, D. G., & Morosan, D. E. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology*, 43(1), 88–96.
- Jenkins, J. (2000). *The phonology of English as an international language: New models, new norms, new goals*. Oxford, UK: Oxford University Press.
- Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, 119(2), 1118–1130.
- Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation, and processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(3), 235–247.
- Kawai, G., & Hirose, K. (2000). Teaching the pronunciation of Japanese double-mora phonemes using speech recognition technology. *Speech Communication*, 30, 131–143.
- Kiliçkaya, F. (2008). Improving pronunciation via accent reduction and text-to-speech software. In T. Koyama (Ed.), *Proceedings of the WorldCALL 2008 conference* (pp. 135–137). Nagoya, Japan: The Japan Association for Language Education and Teaching.
- Kim, M. S. (2005). Perception and production of Korean /l/ by L2 learners and implications for teaching refined pronunciation. *The Korean Language in America*, 10, 71–88.
- King, R. D. (1967). Functional load and sound change. *Language*, 43, 831–852.
- Lambacher, S. (1999). A CALL tool for improving second language acquisition of English consonants by Japanese learners. *Computer Assisted Language Learning*, 12(2), 137–156.
- Lambacher, S., Martens, W., Kakehi, K., Marasinghe, C., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, 26, 227–247.
- Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369–377.
- Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics*, 27, 184–202.
- Levis, J., & Cortes, V. (2008). Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching. In C. A. Chappelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 197–208). Ames, IA: Iowa State University.
- Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, 32, 505–524.
- Liang, Y. (2009). 大学英语大班教学的问题与对策 [Problems and approaches of teaching college English in large classes]. *科技信息 [Technology Information]*, 8, 481.
- Linebaugh, G., & Roche, T. B. (2013). Learning to hear by learning to speak: The effect of articulatory training on Arab learners' English phonemic discrimination. *Australian Review of Applied Linguistics*, 36(2), 146–159.
- Linebaugh, G., & Roche, T. B. (2015). Evidence that L2 production training can enhance perception. *Journal of Academic Language and Learning*, 9(1), A1–A17.
- Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/: 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.

- Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y. I., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, 96(4), 2076–2087.
- Logan, J. S., & Pruitt, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 351–377). Timonium, MD: York Press.
- 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.
- Lopez-Soto, T., & Kewley-Port, D. (2009). Relation of perception training to production of codas in English as a second language. *Proceedings of Meetings on Acoustics*, 6(1), 1–15.
- Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals*, 41(2), 364–379.
- MacKeough, A., Lupart, J. L., & Marini, A. (Eds.). (1995). *Teaching for transfer: Fostering generalization in learning*. London, UK: Psychology Press.
- McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System*, 57, 25–42.
- Motohashi-Saigo, M., & Hardison, D. M. (2009). Acquisition of L2 Japanese geminates: Training with waveform displays. *Language Learning & Technology*, 13(2), 29–47. <https://dx.doi.org/10125/44179>
- Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34, 520–531.
- Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning*, 58(3), 479–502.
- Munro, M. J., Derwing, T. M., & Thomson, R. I. (2015). Setting segmental priorities for English learners: Evidence from a longitudinal study. *International Review of Applied Linguistics in Language Teaching*, 53(1), 39–60.
- Neri, A., Cucchiari, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 15, 441–467.
- Neufeld, G. (1988). Phonological asymmetry in second-language learning and performance. *Language Learning*, 38(4), 531–559.
- Nilsen, D. L., & Nilsen, A. P. (2010). *Pronunciation contrasts in English* (2nd ed.). Long Grove, IL: Waveland Press.
- Nishi, K., & Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research*, 50, 1496–1509.
- Nishi, K., & Kewley-Port, D. (2008). Nonnative speech perception training using vowel subsets: Effects of vowels in sets and order of training. *Journal of Speech, Language, and Hearing Research*, 51, 1480–1493.
- O’Bryan, A., & Hegelheimer, V. (2007). Integrating CALL into the classroom: The role of podcasting in an ESL listening strategies course. *ReCALL*, 19(2), 162–180.
- Okuno, T., & Hardison, D. M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning & Technology*, 20(2), 61–80. <https://dx.doi.org/10125/44461>
- Olson, D. J. (2014). Benefits of visual feedback on segmental production in the L2 classroom. *Language Learning & Technology*, 18(3), 173–192. <https://dx.doi.org/10125/44389>

- Patten, I., & Edmonds, L. A. (2015). Effect of training Japanese L1 speakers in the production of American English /r/ using spectrographic visual feedback. *Computer Assisted Language Learning*, 28(3), 241–259.
- Pimsleur, P. (1963). Discrimination training in the teaching of French pronunciation. *Modern Language Journal*, 47, 190–203.
- Pisoni, D. B., & Lively, S. E. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 433–459). Timonium, MD: York Press.
- Richards, M. (2012). Helping Chinese learners distinguish English /l/ and /n/. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference, Sept. 2011*. (pp. 161–167). Ames, IA: Iowa State University.
- Ringbom, H. (1983). On the distinctions of item learning vs. system learning and receptive competence vs. productive competence in relation to the role of L1 in foreign language learning. In H. Ringbom (Ed.) *Psycholinguistics and foreign language learning: Papers from a conference* (pp. 162–173). Turku, Finland: Åbo Akademi.
- Rochet, B. L. (1995). Perception and production of second-language speech sounds by adults. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 379–410). Timonium, MD: York Press.
- Rvachew, S. (1994). Speech perception training can facilitate sound production learning. *Journal of Speech and Hearing Research*, 37, 347–357.
- Scovel, T. (1969). Foreign accents, language acquisition, and cerebral dominance. *Language Learning*, 19(3–4), 245–253.
- Sha, G. (2010). Using TTS voices to develop audio materials for listening comprehension: A digital approach. *British Journal of Educational Technology*, 41(4), 632–641.
- Skousen, R. (1989). *Analogical modeling of language*. Berlin, Germany: Springer Science + Business Media.
- Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics*, 36(2), 131–145.
- Strange, W., Weber, A., Levy, E., Shafiro, V., & Nishi, K. (2002). Within- and across-language acoustic variability of vowels spoken in different phonetic and prosodic contexts: American English, North German, and Parisian French. *The Journal of the Acoustical Society of America*, 112, 2384.
- Suemitsu, A., Dang, J., Ito, T., & Tiede, M. (2015). A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning. *The Journal of the Acoustical Society of America*, 138(4), 382–387.
- Swan, M., & Smith, B. (2002). *Learner English: A teacher's guide to interference and other problems*. Cambridge, UK: Cambridge University Press.
- Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal*, 28(3), 744–765.
- Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning*, 62(4), 1231–1258.
- Thomson, R. I. (2013). Accent reduction and pronunciation instruction are the same thing. In L. Grant (Ed.), *Pronunciation myths: Applying second language research to classroom teaching*. Ann Arbor: University of Michigan Press.

- Varden, J. K. (2006). Visualizing English speech reductions using the free phonetic software package WASP. In J. D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 127–165). Honolulu, HI: National Foreign Language Resource Center.
- Walden, M., L. (2014). Native Mandarin speakers' perception and production of English stop + liquid clusters in onset position. (Unpublished master's thesis). Syracuse University, Syracuse, NY.
- Walker, N. R., Trofimovich, P., Cedergren, H., & Gatbonton, E. (2011). Using ASR technology in language training for specific purposes: A perspective from Quebec, Canada. *CALICO Journal*, 28(3), 721–743.
- Walley, A. C., & Flege, J. E. (1999). Effect of lexical status on children's and adults' perception of native and non-native vowels. *Journal of Phonetics*, 27(3), 307–332.
- Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System*, 32(4), 539–552.
- Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, 113(2), 1033–1043.
- Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, 106(6), 3649–3658.

Appendix A. FL Values of Commonly Conflated Phoneme Pairs

FL Value	Vowels	Examples	FL Value	Consonants	Examples
10	/e-æ/	bet-bat	10	/p-b/	pat-bat
	/æ-ʌ/	bat-but		/p-f/	pool-fool
	/æ-ɒ/	cat-cot		/m-n/	meet- <u>neat</u>
	/ʌ-ɒ/	cut-cot		/n-l/	<u>n</u> ight-light
	/ɔ-əu/	ought-oat		/l-r/	<u>l</u> ow-row
9	/e-i/	bet-bit	9	/f-h/	fat-hat
	/e-ei/	bet-bait		/t-d/	tie-die
	/ɑ:-ai/	cart-kite		/k-g/	<u>c</u> old-gold
8	/ə-əu/	immersion-emotion	8	/w-v/	<u>w</u> ow-vow
	/i-i/	beat-bit		/s-z/	race-raise
7	---		7	/b-v/	<u>b</u> oat-vote
6	/ɔ:-ə/	form-firm	6	/f-v/	<u>f</u> an-yan
	/ɒ-əu/	cot-coat		/ð-z/	clo <u>th</u> ing-clo <u>s</u> ing
5	/ɑ:-ʌ/	bart-but	6	/s-ʃ/	<u>s</u> ea-she
	/ɔ-ɒ/	caught-cot		/v-ð/	<u>v</u> an-th <u>an</u>
4	/ə-ʌ/	bird-bud	5	/s-z/	per <u>s</u> on-Pers <u>ian</u>
	/e-eə/	shed-shared		/θ-ð/	<u>th</u> igh-th <u>y</u>
	/æ-ɑ:/	at-art		/θ-s/	<u>th</u> ink-s <u>ink</u>
	/ɑ:-ɒ/	cart-cot		/ð-d/	<u>th</u> ough-d <u>ough</u>
	/ɔ-u/	bought-boot		/z-dʒ/	<u>z</u> oo-J <u>ew</u>

	/ə-e/	<u>f</u> urther- <u>f</u> eather		/n-ŋ/	si <u>n</u> -si <u>ng</u>
3	/i-iə/	te <u>a</u> -te <u>a</u> r	4	/θ-t/	th <u>a</u> nk-th <u>a</u> nk
	/ɑ:-əʊ/	v <u>a</u> se-v <u>o</u> ws	3	/tʃ-dʒ/	ch <u>o</u> ke-j <u>o</u> ke
	/u-ʊ/	fo <u>o</u> l-f <u>u</u> ll	2	/tʃ-ʃ/	ch <u>a</u> ir-sh <u>a</u> re
2	/iə-eə/	be <u>e</u> r-b <u>a</u> re		/ʃ-ʒ/	Confuc <u>i</u> an-confus <u>i</u> on
1	/ɔ-ɔɪ/	s <u>a</u> w-s <u>o</u> y		/j-ʒ/	y <u>e</u> s-pl <u>e</u> as <u>u</u> re
	/u-ʊə/	t <u>w</u> o-t <u>o</u> ur	1	/f-θ/	de <u>a</u> f-d <u>e</u> ath
				/dʒ-j/	ju <u>i</u> ce-u <u>s</u> e

Note. Adapted from Brown (1988, p. 604)

Appendix B. Phonemic Contrasts Investigated in the Study

Contrast	Mention in the Literature	FL Value
/æ-ɛ/	Nilsen and Nilsen (2010), Swan and Smith (2002)	10
/ɔ-əʊ/	EFL teacher insights	10
/æ-ʌ/	EFL teacher insights	10
word-final /d-t/	Swan and Smith (2002)	9
/əʊ-ɜ:/	EFL teacher insights	9
/g-k/	Swan and Smith (2002)	9
/ɛ-ɪ/	EFL teacher insights	9
/i-ɪ/	Nilsen and Nilsen (2010), Swan and Smith (2002)	8
/ɑ-ʌ(ə)/	Nilsen and Nilsen (2010), Swan and Smith (2002)	5
/s-θ/	Nilsen and Nilsen (2010), Swan and Smith (2002)	5
/t-θ/	Nilsen and Nilsen (2010), Swan and Smith (2002)	4
/ɛ-ɜ:/	EFL teacher insights	4

Appendix C. Minimal Pairs Used in the Study

	Set A (Words Trained, Voices Trained)	Set B (Words Trained, Voices New)	Set C (Words New, Voices Trained)
/æ-ɛ/	latter-letter dad-dead flash-flesh	mansion-mention shall-shell	pat-pet gas-guess
/ɔ-əʊ/	cost-coast lawn-loan not-note	hall-hole pause-pose	road-rod soak-sock
/æ-ʌ/	staff-stuff dam-dumb lack-luck	ankle-uncle cap-cup	match-much drag-drug
/d-t/	add-at extend-extent slide-slight	fade-fate weed-wheat	coat-code kid-kit
/əʊ-ɜ:/	girl-goal learn-loan turn-tone	arrow-error birth-both	sir-so eager-ego
/g-k/	angle-ankle lock-log locking-logging	buck-bug dock-dog	back-bag pick-pig

/ɛ-ɪ/	medal–middle position–possession	lesson–listen set–sit	desk–disk left–lift
/i-ɪ/	lead–lid least–list reach–rich	scene–sin peak–pick	seat–sit sheep–ship
/ɑ- ʌ(ə)/	shot–shut body–buddy calm–come	dock–duck long–lung	boss–bus cop–cup
/s-θ/	gross–growth mouse–mouth seem–theme	sum–thumb face–faith	pass–path sink–think
/t-θ/	boot–booth eight–eighth thigh–tie	pat–path team–theme	death–debt tank–thank
/ɛ-ɜ/	debt–dirt beds–birds end–earned	bed–bird best–burst	edge–urge ten–turn

About the Authors

Manman Qian is a PhD Candidate in the Applied Linguistics and Technology program at Iowa State University.

E-mail: mqian@iastate.edu

Evgeny Chukharev-Hudilainen is an Assistant Professor in the Applied Linguistics and Technology program at Iowa State University. His research work includes using cognitive linguistics, computer science, and natural language processing to design, build, and evaluate technologies for second language learning and assessment.

E-mail: evgeny@iastate.edu

John Levis is a Professor in the Applied Linguistics and Technology program at Iowa State University. He specializes in second language pronunciation and speech intelligibility, with a focus on how second language pronunciation research affects the teaching of pronunciation.

E-mail: jlevis@iastate.edu