Computer Assisted Pronunciation Training: Targeting Second Language Vowel Perception Improves Pronunciation

RON I. THOMSON
Brock University, Canada

ABSTRACT
This paper first provides an overview of factors that constrain ultimate attainment in adult second language (L2) pronunciation, finding that first language influence and the quantity and quality of L2 phonetic input account for much of the variation in the degree of foreign accent found across adult L2 learners. The author then evaluates current approaches to computer assisted pronunciation training (CAPT), concluding that they are not well grounded in a current understanding of L2 accent. Finally, the author reports on a study in which twenty-two Mandarin speakers were trained to better discriminate ten Canadian English vowels. Using a specially designed computer application, learners were randomly presented with recordings of the target vowels in monosyllabic frames, produced by twenty native speakers. The learners responded by clicking on one of ten salient graphical images representing each vowel category and were given both visual and auditory feedback as to the accuracy of their selections. Pre- and post-tests of the learners’ English vowel pronunciation indicated that their vowel intelligibility significantly improved as a result of training, not only in the training context, but also in an untrained context. In a third context, vowel intelligibility did not improve.

KEYWORDS
Computer Assisted Pronunciation Training, L2 Speech Perception and Production, High Variability Phonetic Training

INTRODUCTION
While achieving native-like pronunciation in a second language (L2) is rarely a necessary or realistic goal for adult learners, even developing adequately intelligible L2 speech is no small task, requiring vast amounts of experience with the target language. In most cases, pronunciation inaccuracies betray underlying perceptual inaccuracies, suggesting that many adult L2 learners lack efficient L2 bottom-up listening skills as well, although they often suffer in silence. It is quite striking, then, that despite its importance in terms of both the intelligibility of speech and its relation to listening, pronunciation remains a largely neglected component of L2 instruction.

Unfortunately, experimental studies investigating the causes of L2 accent have not often translated into improved pronunciation training, particularly with respect to the teaching of segmentals (i.e., vowels and consonants). The failure of research to impact practice is not only due to the often-noted chasm between researchers and practitioners (see Derwing & Munro, 2005 for an overview), but also because the context of learning and the resources available in language classrooms do not typically allow implementation of some research-motivated techniques. This paper provides a brief overview of research perspectives on L2 accent, and argues that computer-mediated instruction uniquely affords the possibility for
learners to access the type of training that can lead to significant improvement in L2 pronunciation. This claim is then investigated in a computer-assisted pronunciation teaching (CAPT) study, which incorporates research findings into the design of the software used, and provides training in a language program’s computer lab.

BACKGROUND

The Development of L2 Pronunciation

Flege’s (1995) Speech Learning Model (SLM) has been particularly influential in research investigating factors that constrain ultimate attainment in L2 pronunciation. The SLM starts with the assumption, held by others (e.g., Birdsong, 2007; Bongaerts, van Summeren, Planken, & Schils, 1997) that the cognitive mechanisms used to learn an L1 are available to adult L2 learners, but that accessing them is difficult. Furthermore, while the context of L2 learning is recognized as being fundamentally different from L1 learning, there is an implicit belief that to the extent L2 learning can be shaped to better imitate L1 acquisition, ultimate attainment in L2 pronunciation can be pushed much further than has usually been the case. The factors most commonly identified as constraining L2 pronunciation accuracy include interactions between learners’ L1 and L2 phonological systems, age of learning, and the amount of experience learners have with the L2 (Flege, Bohn, & Jang, 1997; Flege, Frieda, & Nozawa, 1997; Gottried and Suiter, 1997; Piske, MacKay, & Flege, 2001).

Flege’s SLM (1995) expands on earlier attempts (e.g., Corder, 1971; Lado, 1957; Selinker, 1972) at explaining the role of learners’ L1 systems in L2 pronunciation. A key contribution in this regard is the recognition that the difficulty adult learners face in acquiring L2 sound categories can be predicted on the basis of how similar sounds are in the learners’ L1s and L2s. In the case of identical L2 sounds, L1 transfer is predicted to be immediate, while in the case of highly dissimilar L2 sounds, new categories will develop, although the timeframe required will vary across categories and learners. For those L2 sounds that are similar, but not identical to L1 sounds, learning will be much more difficult. In such cases, learners might wrongly associate a sound in the L2 with a similar L1 category, when in fact a categorical difference exists. This is most problematic when learners perceive two or more L2 sounds as belonging to a single L1 category. In such cases two-category merger often occurs (e.g., Spanish speakers who perceive “sit” and “seat” to contain the same Spanish /i/-like vowel).

With respect to age of learning and accent, the belief in a putative biological critical period (e.g., Lenneberg, 1967) has lost support in favor of a more complex and less dogmatic view. While age is still seen as a major factor, it is increasingly unclear that the relationship between age and accent is biological in nature. Rather, age effects may at least partially stem from differences in learning environments and the accompanying experience normally available to learners of different ages. Flege, Munro, & MacKay (2005) examined the L2 pronunciation of over 200 Italian immigrants to Canada. Participants had arrived in Canada across a broad range of ages (2 to 23 years) and had resided there for an average of 32 years. Judgments of their speech by native English speakers revealed that rather than a precipitous increase in accentedness associated with a particular age of learning (e.g., around adolescence), as the Critical Period Hypothesis would predict, a gradual linear increase in accentedness occurred as age of learning increased.

Further research has led to the conclusion that the degree of experience learners have with the L2 is as a strong or stronger a predictor of ultimate attainment than is biological age (Yamada, 1995; Flege, Bohn, and Jang, 1997). There is also a relationship between strength of L2 accent and the extent to which learners continue to use their L1 vis-à-vis their L2 (Flege, Frieda, & Nozawa, 1997). While the effect of L2 experience on accent is in-
creasingly well attested, there is a growing awareness that the quality and content of L2 experience should also be considered (Moyer, 2009). For example, learning to perceive and produce an L2 vowel in one consonantal context may not transfer to different consonantal contexts (see Flege, 1995; Strange, 2007; and Thomson, Nearey, & Derwing, 2009). Similarly, differences in the way sound categories are produced by different native speakers (NSs) of the target language can also impact learning. That is, learners sometimes perceive two NS productions of the same target category as belonging to two separate categories. This means individual productions of target categories will vary with respect to their usefulness as input for L2 learners. For example, Mandarin L1 learners of English in Thomson et al.’s (2009) study misperceived some vowels produced by one NS speaker, but not those produced by another NS. Evaluation of the misperceived tokens indicated that they were acoustically more similar to competing Mandarin vowel categories than were the correctly perceived tokens, despite both being equally good exemplars of the intended English category. Since tokens that were more Mandarin-like resulted in misperception of the target vowel as equivalent to a competing Mandarin category, these learners would likely have some difficulty establishing the intended English category if they were trained using only the first speaker’s voice.

Lexical context can also affect the pronunciation of speech sounds (Walley & Flege, 1999). In a study investigating the development of Canadian English vowel categories by groups of Mandarin and Slavic immigrants to Canada, Munro and Derwing (2008) concluded that variation in the pronunciation of the same vowel in different words might be related to each token’s lexical frequency in the ambient language. In general, L2 productions of more frequent English words contained more intelligible vowels. Somewhat related to frequency, learners’ familiarity with lexical items also plays a role. Thomson and Isaacs (2009) demonstrated that Mandarin learners’ familiarity with English words predicted intelligibility scores for vowels found within those words.

Finally, apart from differences in the nature of input, the extent to which learners can direct their attention to phonetic form is also important. If learner attention is oriented toward phonetic information, more of the input can be incorporated into emerging L2 categories. Researchers argue that because cognitive resources are finite, simultaneously processing both form and meaning is often difficult for L2 learners (Lee, Cadierno, Glass, & VanPatten, 1997; Schmidt, 2001). Since meaning is of greater importance to communication, it is the first feature to receive attention, often coming at the expense of attention to form. Flege (1995) argues that these limits on attention are implicated in the development of L2 pronunciation. Consequently, he suggests that for learners to fully benefit from phonetic input, it should be presented in contexts where competing demands for attention are kept to a minimum. Borden, Gerber, and Milsark (1983) provide evidence that meaning interacts with phonetic form in an experiment showing that L2 speakers’ pronunciation is more intelligible when they imitate nonsense syllables than when they imitate real words. This suggests that once words are recognized, inaccurate phonetic representations stored in long-term memory are activated, resulting in pronunciation that diverges from the stimulus. In the case of nonsense words, learners seem to be better able to attend to the phonetic form, since there is no competition from previously stored phonetic representations associated with meaning. In an experimental study, Guion and Pederson (2007) explicitly investigated the effect of attention on phonetic learning. Two groups of English speakers were asked to learn a set of Hindi words that contained pairs of sounds known to be difficult for English speakers to discriminate. One group was instructed to pay particular attention to the initial consonants in each word and was told that many words differed by sounds that were likely to be difficult to discriminate. A second group of learners was simply asked to learn the words, without any attention drawn toward their phonetic properties. Pre- and post-tests measuring the
participants’ ability to discriminate the Hindi sound contrasts found that the sound-attending group improved significantly more than the meaning-attending group.

FROM RESEARCH TO PRACTICE

It was argued in the preceding section that interactions between L1 and L2 categories, and differences in the quantity and quality of input available to L2 learners, accounts for much of the variation in strength of foreign accent. Therefore, the focus of pedagogical intervention should be to maximize the impact of experience, while insuring that learners are exposed to sufficient variation that they become aware of how L2 sounds differ from L1 categories. If adult L2 learners are able to obtain large quantities of high quality L2 input, significant gains in pronunciation might be achieved over a shorter period of time. Unfortunately, the type of input that has just been shown to be required is not readily available in most naturalistic settings, nor is it realistic to provide it in traditional classroom-based settings. Technology, however, affords the opportunity for progress in this area.

Computer Assisted Pronunciation Training

Although recent research has increased our understanding of L2 pronunciation learning, most computer assisted pronunciation training (CAPT) applications available today merely replicate classroom instruction, but on a computer monitor (e.g., minimal pair practice, descriptions and diagrams of articulatory gestures, etc.). This limits the applications’ usefulness to offering more accessible, self-paced instruction. Some CAPT applications have been criticized as being nothing more than flashy packaging of pedagogically unsound approaches to pronunciation learning that may actually confuse rather than help learners (Hansen, 2006; Neri, Cucchiarini, Strik, & Boves, 2002; Pennington, 1999). For example, one very popular program provides a pronunciation meter that purportedly tells learners how native-like their pronunciation is. It does not tell them what is wrong with their pronunciation, however, and has the unfortunate tendency of indicating that some highly accented productions are native-like, while native speakers are sometimes judged as accented. This suggests that the underlying speech recognition algorithm either does not work, or the pronunciation meter was never intended to be anything other than a marketing gimmick.

Progress has been made toward using CAPT in ways that complement rather than replace classroom pronunciation instruction. For example, some applications provide visual feedback concerning the phonetic properties of the learner’s recorded speech and allow comparison to the phonetic properties of a native speaker model. Visual displays tend to be most effective for teaching suprasegmentals (Anderson-Hsieh, 1994; Eskenazi, 1999; Neri, et al., 2002). Intonation, rhythm and stress can be indicated using pitch contours, relative duration of segments, and amplitude of segments. This information is displayed visually in ways that are simple for learners to interpret. While visual feedback on suprasegmental features of speech is reported to lead to improved pronunciation that also generalizes to new contexts (e.g., Hardison, 2004, 2005), it is uncertain whether the resulting improvement is any greater than that achieved through traditional classroom-based suprasegmental instruction (e.g., Stenson, Downing, Smith, & Smith, 1992). When it comes to providing displays of the acoustic characteristics of individual segments (i.e., spectrograms), there is seemingly no benefit to learners. Spectrograms are uninterpretable to non-experts, and do not convey any information that can be readily used to improve pronunciation. Nevertheless, some CAPT programs do display spectrograms, although the designers’ motivations for doing so are unknown.

For the reasons just provided, Neri et al. (2002) argue against the use of spectrograms for segmental feedback, instead advocating the use of automatic speech recognition (ASR) techniques. While ASR is incapable of telling learners how to produce sounds, it is able to
indicate when sounds are not being produced in a way that the machine recognizes as intelligible. It is this ability to provide learners with feedback that is a major appeal of using ASR in CAPT (Hansen, 2006; Neri et al., 2002). A number of popular language software programs reviewed by Neri et al. (2002) rely on ASR to simulate interactive conversations. In these programs, the learner uses prefabricated responses, which he/she has learned or must read off the monitor, to respond to conversational turns made by the computer. If the ASR system fails to recognize the learner’s utterance, they must repeat it. Unfortunately, it is unclear to what extent such ASR systems correlate with human judgments of intelligibility. It may be the case that the user is perfectly intelligible, but that the system simply fails to recognize the utterance, or that the user’s speech is not sufficiently similar to the dialect used to train the ASR system. As a result of these sorts of errors, there is some risk that speakers who are adequately intelligible may be made to feel as though they are not. Some evidence of this type of mismatch between ASR responses and human listeners is provided by Derwing, Munro, and Carbonaro (2000), who found that one ASR system recognized approximately 24-26% fewer words produced by nonnative speakers than did human listeners.

Since current ASR technology is known to often provide erroneous feedback to learners (Neri, Cucchiarini, & Strik, 2008) its usefulness as a pronunciation teaching tool is limited. In fact, very little is known about the efficacy of ASR tools in CAPT (Neri et al., 2008). A few studies have investigated whether ASR results in measurable improvements in pronunciation, but with mixed results. Hinks (2003) concludes that ASR is only of benefit to those who have the strongest foreign accent to begin with, but not of much benefit to those who begin training with intelligible, albeit accented, pronunciation. In cases where ASR techniques have resulted in improved pronunciation, it is uncertain that the same improvements could not have been achieved through traditional classroom-based pronunciation instruction. Mich, Neri, and Giuliani (2006) found comparable amounts of improvement across two groups of young learners, one in a traditional class and one using ASR feedback, suggesting that ASR does not add to, but simply replaces, the teacher. Similar results were found by Neri, Cucchiarini, and Strik (2006a), who compared adult learners in traditional vs. ASR-based classes. Finally, studies of ASR’s effectiveness in CAPT have not shown that improvement in the ASR environment results in more intelligible pronunciation in novel contexts. For example, there is no convincing evidence that ASR-based pronunciation training is generalizable to new words, nor whether it impacts the learners’ perceptual system.

Considering current trends in CAPT in light of what we know about the underlying sources of accent, it is obvious that most CAPT is far from cutting edge. First, it does not attempt to address interactions between L1 and L2 categories. CAPT programs typically utilize a single speaker to provide input for a given lexical item or sentence. No consideration is given to the possibility that a single speaker’s voice may not be ideally suited for presenting input to a range of speakers who speak a variety of L1s. Second, CAPT typically fails to increase the impact of input by improving its quantity and quality. Admittedly, because many CAPT programs capture learners’ attention, input is more likely to get noticed than is the case in traditional classrooms. However, there is very little variation in the input provided. Furthermore, most CAPT applications present sounds in the context of words, with no consideration for the relationship between word frequency, word familiarity and pronunciation. Finally, while feedback in CAPT may direct the learner’s attention to phonetic properties, when feedback is erroneous, it may be counterproductive.
**Experimental Techniques**

One research-based technique that could be extended to CAPT is high variability phonetic training (HVPT). The defining feature of HVPT is that learners are exposed to multiple voices producing the target sounds, rather than a single voice as is often the case in a classroom environment, or even in many CAPT applications. In HVPT, after hearing sounds presented randomly via the computer, learners must click on labels indicating which sound they perceived, and are then given feedback on the accuracy of their response. Logan, Lively, and Pisoni (1991) and Lively, Logan, and Pisoni (1993) were the first to show that increasing the variability of the input results in greater and more generalizable gains in L2 speech perception. Since then, many more studies have shown similar results. Unfortunately, a disproportionate number of such studies have focused on training Japanese listeners to perceive the English /l/-/r/ contrast (Bradlow, 2008; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Iverson, Hazan, & Banister, 2005; Lively et al., 1993; Lively, Pisoni, Yamada, Tokhura, & Yamada, 1994; Logan et al., 1991; McClelland, Fiez, & McCandliss, 2002). Several recent HVPT studies, however, have begun examining the technique’s usefulness in training listeners to better perceive L2 English vowels. For example, Lambacher, Martens, Kakehi, Marasinghe, and Molholt (2005) found HVPT helped Japanese speakers improve their perception of five American English vowels. Nishi and Kewley-Port (2007) found that Japanese speakers’ perception of nine American English vowels rapidly improved as a result of HVPT, later finding the same effects for Korean speakers’ perception of the same American English vowels (Nishi & Kewley-Port, 2008).

Incorporating the HVPT technique into CAPT programs can address many of the constraints on L2 pronunciation learning outlined in the first section of this article. For example, in HVPT, L2 learners are trained to identify sounds produced by multiple talkers, in multiple phonetic contexts. This strategy is based on the belief that exposing learners to high variability input will allow them to begin recognizing the difference between meaningful phonetic cues associated with particular sound categories, and cues that are irrelevant to category identity. It is believed that this knowledge will then transfer to the perception of novel speech tokens. In contrast, when trained on a single voice, learners may not attend to the relevant cues at all. Instead, they will rely on any cues that help them discriminate between training stimuli, whether those cues are generalizable or not. For example, a learner might rely on durational differences to discriminate between English /i/ and /ɪ/ tokens produced by a particular voice. Unfortunately, durational differences for this contrast are only an unreliable secondary cue to vowel identity, and are often unavailable in natural speech.

Because HVPT is a computer-mediated technique, it is also able to orient learners’ attention to the phonetic learning task, by presenting sounds in contexts where the demand for attention to meaning is minimized. It is also reasonable to conclude that the interactive nature of the training also helps focus the learner’s attention, and corrective feedback allows learners to readjust errors in perception the moment they occur.

**PURPOSE OF THE CURRENT STUDY**

The purpose of this study is not to critique and extend previous versions of HVPT, which is done in many of the studies cited, but to show how principles from HVPT can be extended to language lab environments. Despite its proven effectiveness, the HVPT technique has not found its way into CAPT applications. Rather, this technique has largely been restricted to use in speech laboratory experiments. Furthermore, most HVPT studies have been concerned with the technique’s effect on L2 speech perception, and rarely investigate whether perceptual training improves pronunciation. Since pronunciation and perception are inextricably linked (Baker & Trofimovich, 2006; Flege, 1995; Thomson, 2008), improvement in perception should allow learners to more effectively monitor their own productions. Theo-
retically, having a more native-like perceptual system should promote gains in pronunciation accuracy.

Only one study is known to have demonstrated that HVPT training can result in significant improvement in the pronunciation of consonants. In that study, Bradlow et al. (1997) trained Japanese speakers in the perception of English /l/ and /r/. Although the researchers did not explicitly provide pronunciation instruction, the learners’ productions of /l/ and /r/ were found to have improved as a result of training. In addition, a single study has demonstrated that the pronunciation of L2 vowels can improve in response to HVPT. Lambacher et al. (2005) report that Japanese speakers trained in the perception of five American English vowels significantly improved their pronunciation scores for those vowels. A limitation of Lambacher et al.’s study is that it trained learners on a small subset of American English vowels. Nishi and Kewley-Port (2008) present evidence that for L2 vowel perception, training on a subset is not as effective as training on a larger set covering more of the vowel space.

In the study reported here, using high variability training, twenty-two Mandarin learners were trained to identify ten Canadian English vowels produced by twenty-one native speakers. The decision to train vowels rather than consonants was made because there is less research involving vowels, and it was felt that vowel instruction would be of greater benefit to learners. L2 vowels are more difficult to learn than are consonants, regardless of the learners’ L1 background. For example, Neri, Cucchiarini, and Strik (2006b) found that speakers’ intelligibility of Dutch L2 consonants improved over time without direct intervention, while Dutch L2 vowels did not improve to the same extent. Munro and Derwing (2008) found that Mandarin and Slavic speakers who had immigrated to Canada continued to experience great difficulty with English vowels, despite spending a year in the country. Finally, Bent, Bradlow, and Smith (2007) found that vowels contribute more to the intelligibility of words than do consonants, suggesting they should be a pedagogical priority.

Research questions

1. Can computer-mediated training in the perception of L2 English vowels improve speech intelligibility without explicit pronunciation practice?
2. Does perceptual training generalize to L2 productions elicited by an unfamiliar voice?
3. Can perceptual training in one phonetic environment improve speech intelligibility in new phonetic environments?
4. Do differences in the quality of training stimuli lead to differences in learner outcomes?

Method

Native English speakers

Twenty-one native English speakers of a standard variety of Canadian English (M age = 29.9; range, 18-55) provided the speech samples used as training and testing stimuli.

L2 English learners

Twenty-two well-educated adult Mandarin L1 learners of English participated in the study (M age = 36.3; range, 27-50, 14 female and 8 male). Their average length of residence in Canada was 11.5 months (range 4-48). All were enrolled in ESL classes at the time of the study, and had received an average of 4.3 months (range 1-13) of instruction. Those few participants who had been in Canada longer than a year reported having had few interac-
tions in English prior to beginning their ESL classes. Upon arrival in Canada, participants’ English listening and speaking ability had been assessed as being between Canadian Benchmarks 1 and 3, indicating that they were beginners. Some were able to read and write at slightly higher levels, having received prior English instruction in China. The ESL program assessed them as between beginner and lower intermediate at the time of the study. Students received very little explicit pronunciation instruction as part of their ESL program. All reported having normal hearing.

Training stimuli preparation
Twenty of the native English speakers produced ten Canadian English vowels, /i, ɪ, e, ɛ, æ, ɒ, ʌ, ʊ, u/, in /bV/ and /pV/ frames, and were recorded in a quiet room using a Marantz PMD 660 digital recorder, with a sampling rate of 44.1 KHz. Limiting training stimuli to these two contexts was intended to allow for some phonetic variation, while not overwhelming learners with the sort of variation that can lead to confusion. As pointed out earlier, the perception of speech sounds is known to be context dependent (Flege, 1995; Strange, 2007; Thomson & Isaacs, 2009), meaning learning in one phonetic context does not immediately lead to improvements in different contexts. The choice to use open rather than closed syllables was made in order to increase the likelihood of learning by the L1 Mandarin speakers in this study. In earlier studies (Munro & Derwing, 2008; Thomson, 2005), there was an indication that Mandarin learners of English had difficulty producing the same ten English vowels, not only because of differences between the Mandarin and English vowel inventories, but also because the test stimuli included coda consonants that are prohibited by Mandarin phonotactic constraints. Although the lax vowels used in this study are not found in isolated open syllables in English, the native speakers who provided these samples had no difficulty producing these syllables. These vowels are found in open English syllables in words (e.g., the first syllables in ‘bitter’, ‘better’, ‘batter’, ‘butter’, ‘booker’, etc.). All training speech samples were elicited twice using an auditory prompt, where speakers heard the target syllables produced in the carrier phrase, “The next word is ____” and responded by saying, “Now I say ____”, repeating the target syllable they just heard. After recording each speaker, Sound Studio 3 was used to extract the target syllables, normalize them to peak amplitude and save them as individual sound files. Only the first repetition of each item was used as training stimuli, except where the researcher considered the first repetition to be of poor quality. For only seven of the 400 items was the first repetition replaced by the same speaker’s second repetition of the same item. In total four /ɛ/ productions and three /ʌ/ productions were replaced with second repetitions. In each case, although the second repetition was determined to be superior, it was still perceived as a somewhat ambiguous exemplar of the target category. This fact was not seen as deleterious to the overall training program. First, these tokens represent a small proportion of the forty tokens of each vowel category. Secondly, it was decided that such ambiguity might actually be reflective of the sort of variation present in naturalistic input.

The resulting sound files were then used to create two training sets. One set comprised modified versions of the original 400 recordings whereby, using Praat’s Pitch Synchronous Overlap Add method, the vowel portion of each recording was doubled in length. Using this technique, the resulting recordings sound perfectly natural, although longer than those found in normal speech. It was hypothesized that providing learners with lengthened versions of the vowels would allow them to better attend to the quality of the vowel than is possible in the shorter original versions. The second training set comprised a 200 item subset of the original 400 recordings, representing the 20 items from each vowel category that were deemed to be least like any Mandarin counterpart. Items were selected using linear discriminant analysis scores derived in a previous study (Thomson et al., 2009) from acous-
tic properties known to be highly correlated with human vowel perception (i.e., formants and pitch).

**Test stimuli preparation**

Test stimuli included 80 recorded elicitation prompts. For twenty items, the ten vowels were produced in /bV/ and /pV/ contexts by a familiar speaker who had also provided training stimuli (Voice 1). For the remaining sixty items, the ten vowels were produced in the same /bV/ and /pV/ contexts in addition to four new contexts, /gV/ and /kV/ and /zV/ and /sV/ by a speaker who had not provided training stimuli (Voice 2). Both speakers spoke a standard Canadian variety of English, and both were phonetically trained. The author and another colleague evaluated the resulting recordings and determined that they were sufficiently clear productions of the target sounds.

**Training procedure**

The 22 L2 English learners were randomly assigned to one of two training groups (11 to each group). One group was trained on the set of stimuli containing the vowel-lengthened versions of the HVPT stimuli (Lengthened Vowel Group). The other group was trained on the 200-item subset of stimuli determined to be acoustically least similar to any Mandarin L1 category (Select Vowel Group). Given limits on the total number of participants available for this study, a traditional control group was not used. However, no previous research has ever found short-term improvement in L2 vowel pronunciation without pedagogical intervention (e.g., Lambacher et al., 2005; Nishi & Kewley-Port, 2007). Even over longer periods of time, improvement is very gradual, if it happens at all. For example, in a longitudinal study of English vowel development by Munro and Derwing (2008), although Mandarin and Slavic speakers’ English vowels became significantly more intelligible over time, this took four months, and the extent of improvement was very limited. Furthermore, the significant improvement that was found only occurred in the four months immediately preceding their arrival in Canada. Over the next eight months, no further improvement was detected in either group. The Mandarin learners in the present study had already been in Canada for over a year, indicating that they were well past the period during which pronunciation naturally improves, at least in a short timeframe.

Training was implemented using a stand-alone application created with MATLAB® and easily installable on Windows PCs. In the program, learners were taught to associate a distinctive nautical flag image with each of the ten English vowel categories (see Figure 1). The choice of non-orthographic symbols for use in training was made to avoid any confounding effects from potentially inaccurate past associations between sounds and English orthography. Throughout training, the flags’ location on the screen remained static.

During an initial training session, pairs of flags and their associated sounds were presented in a step-wise fashion, where the learner heard the target vowels spoken by a single voice, after which the corresponding image flashed on the screen. In each of the training sessions that followed, 200 items from each training set were randomly presented. After hearing each item, the learner was asked to respond by clicking on the flag that they felt matched the vowel they had just heard. After correct responses, learners heard a chirp, while after incorrect responses, learners heard a beep, and saw the correct response flashed on the screen. They then had to click on the correct response in order to continue to the next item. Participants completed a total of eight training sessions over a period of three weeks. Each session was self-paced, and typically took between fifteen and twenty minutes to complete. Learners did not evidence any difficulty learning to associate sounds with flags. By the end of the second training session, learners were identifying several known English vowel cate-
gories at near ceiling accuracy rates. Several learners did report ignoring the visual information available on the flag symbols and instead relying on their locations on the screen.

Figure 1
Screenshot of training screen showing nautical flag images associated with each English vowel category

Phonetic symbols were not visible in the training program and are only provided here for illustrative purposes.

Testing procedure
Before any perceptual training sessions had occurred (Time 1), and again after training was complete (Time 2), the L2 learners were recorded producing the target English vowels in the same elicited imitation task used to collect the training stimuli from the native speakers. That is, participants heard the target syllable presented in the carrier phrase, “The next word is ____” and repeated the target syllable in the carrier phrase “Now I say ____.” The testing phase included elicited imitation of the ten target vowels produced by the two test speakers in the /bV/ and /pV/ contexts, and imitation of the vowels produced by the second test speaker in /gV/, /kV/ and /zV/, /sV/ contexts. Participants completed two repetitions of the production test at Time 1 and Time 2, with two orders of stimuli presentation, blocked by voice and context, counter-balanced across participants. Recordings of L2 productions were made in a quiet room using the same equipment and procedure as that outlined for recording the native speaker productions of training stimuli.

Native speaker judges
The intelligibility of the resulting vowel productions was assessed by five native speakers of a standard variety of Canadian English (M age = 30.6; range = 21-46), four females and one male. One judge was the author. All were phonetically trained and accustomed to identifying English vowels in terms of the International Phonetic Alphabet. All reported having normal hearing.

Vowel tokens were judged using Praat’s multiple forced choice identification tool. All 7040 L2 English productions as well as the 80 test stimuli items were fully randomized and divided into 36 separate blocks for presentation to judges. Multiple listening sessions were necessary in order to limit fatigue. The listening task was self-paced and took place over several weeks. Judging all items required between 12 and 15 hours of listening, depending on the judge. During the task, judges heard each item and then clicked on the IPA symbol.
that was closest to the vowel they heard. Judges were also provided with a “replay” button, which allowed them to listen to each item up to five times. However, judges were told that using the replay button should be reserved for those rare tokens that were particularly ambiguous. An “oops” button was also provided, which allowed them to reassess an item if they accidently made a mistake. Following Munro and Derwing (2008), intelligibility scores for each token were defined as the proportion of judges who identified the vowel as being correctly produced. This provides a gradient measure of intelligibility, whereby tokens that are unambiguously correct receive the highest score (i.e., 1), while those that are unambiguously incorrect receive the lowest score (i.e., 0). Those that are ambiguous, but identified by some judges, receive an intermediate score (i.e., .2, .4, .6, or .8).

RESULTS

Interjudge agreement

Of the 80 elicitation stimuli used in the pre and post-tests, 79 were correctly identified by at least half of the five judges (M intelligibility score = .96; range .4-1). Voice 2’s /zɛ/ production was correctly identified by only two of the five judges. With respect to L2 productions, three out of five judges agreed on the identity of 97% of the items, four out of five on 80% of the items, and five out of five on 60% of the items. In no case was there complete disagreement on the identity of a particular item. That is, for every item, at least two of the five judges agreed with each other on its identity.

Training Set, Test Voice and Time

Table 1 provides descriptive statistics for the L2 speakers’ mean vowel intelligibility in response to each test Voice and Context at each time.

| Table 1 |
| Mean intelligibility scores in response to each test stimuli voice and context |

<table>
<thead>
<tr>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 1 x Time 2 Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice 1 /b,pV/</td>
<td>0.76</td>
<td>0.08</td>
</tr>
<tr>
<td>Voice 2 /b,pV/</td>
<td>0.75</td>
<td>0.09</td>
</tr>
<tr>
<td>Voice 2 /g,kV/</td>
<td>0.72</td>
<td>0.12</td>
</tr>
<tr>
<td>Voice 2 /z,sV/</td>
<td>0.67</td>
<td>0.12</td>
</tr>
</tbody>
</table>

**Significant at the .001 level

To determine whether there was any significant improvement in the L2 speakers’ English vowel intelligibility scores after training, in response to Voice 1 and Voice 2 elicitation stimuli, and with respect to the two training conditions, a three-way ANOVA was conducted with Time (2 levels), Voice (2 levels) and Vowel (10 levels) as within-group factors. Training condition (2 levels) was used as a between-group factor. No significant effect was detected for Training condition, that is, the Lengthened Vowel group versus the Select Vowel group. However, a significant main effect was found for Time, F(1, 20) = 18.841, p < .01, partial η² = .485. Figure 2 illustrates improvement over time for each vowel produced in response to Voice 1 and Voice 2 stimuli respectively.
Figure 2
Learner responses to /b,p,V/ elicitation stimuli at Time 1 and Time 2 in response to Voice 1 and Voice 2 stimuli

A significant main effect for Vowel was also found, $F(9, 180) = 33.148$, $p < .01$, partial $\eta^2 = .624$, indicating that some vowels were more intelligibly produced than were others. There was no significant main effect for test Voice. However, a significant Voice x Vowel interaction was found, $F(9, 180) = 4.492$, $p < .01$, partial $\eta^2 = .183$.

While it is unsurprising to find that some English vowels were easier for Mandarin speakers to produce than were others, the significant Voice x Vowel interaction was less expected. This indicates that learners responded differently to each voice prompt in the elicited imitation test. Post-hoc Bonferroni adjusted $t$-tests indicated that the only significant difference in response to the two elicitation voices was for the vowel /ɪ/, $t(21) = -3.984$, $p = .001$, which was more intelligible when produced in response to Voice 1. Although Bonferroni adjusted $t$-tests can sometimes lead to Type II errors, in this case, even without a correction for ten pair-wise comparisons, one for each vowel, differences in responses to the two test Voices would not have reached significance for any other vowel category.

**Training set, Test Context and Time**

To determine whether there was any significant improvement in the L2 speakers’ English vowel intelligibility scores after training, in response to Voice 2 elicitation stimuli in untrained phonetic contexts, and with respect to the two training conditions, a three-way ANOVA was conducted with Time (2 levels), phonetic Context (3 levels) and Vowel (10 levels) as within-group factors. Training condition (2 levels) was a between-group factor. Again, no significant main effect for Training condition was detected. Significant main effects were found for Time, $F(7, 180) = 7.507$, $p = .013$, partial $\eta^2 = .273$, Vowel, $F(9, 180) = 6.394$, $p < .01$, partial $\eta^2 = .708$, and phonetic Context, $F(2, 40) = 22.189$, $p < .01$, partial $\eta^2 = .526$. Significant Context x Vowel, $F(18, 360) = 6.906$, $p < .01$, partial $\eta^2 = .257$, and Context x Time, $F(4, 40) = 4.366$, $p = .019$, partial $\eta^2 = .179$, interactions were also found. Figures 3 and 4 illustrate mean intelligibility scores for L2 productions in response to /g,k,V/ and /z,s,V/ stimuli produced by Voice 2.
To examine the Context x Vowel and Context x Time interactions, a series of post-hoc Bonferroni adjusted t-tests were conducted. In terms of the effect of Context on Vowel, tests indicated that tokens of the English vowels /e/, /ɛ/ and /æ/ were produced significantly more intelligibly in /b,pV/ contexts than in /z,sV/ contexts, $t(21) = 4.874, p < .000$ and $t(21) = 3.702, p = .001$; $t(21) = 4.004, p = .001$ respectively. L2 productions of English /ʊ/ were significantly more intelligible in /b,pV/ and /g,kV/ contexts than in /z,sV/ contexts, $t(21) = 4.969, p < .000$ and $t(21) = 4.503, p < .000$ respectively. In terms of the Time x Context interaction, intelligibility improved from Time 1 to Time 2 in the /b,pV/ training context, $t(21) = -3.317, p = .003$, and in the /z,sV/ context, $t(21) = -2.585, p = .017$, but not in the /g,kV/ context, $t(21) = 0.203, p = .841$.

Figures 5 and 6 provide a visual illustration of each learner’s individual degree of improvement in English vowel pronunciation in the /b,pV/ context from Time 1 to Time 2, in response to Voice 1 and Voice 2 respectively. As can be seen, all but four of the twenty-two
learners improved in their pronunciation in response to Voice 1 stimuli, while all but three improved in response to Voice 2 stimuli. No correlation was found between age and vowel intelligibility at Time 1 or Time 2, nor was there any significant correlation between age and the extent of improvement. Interestingly, the oldest learner, who was 50 years old, experienced some of the greatest improvement in her mean vowel intelligibility scores in response to the training.

Figure 5
Individual learners’ average L2 English vowel intelligibility scores before and after training in response to Voice 1 /b,pV/ stimuli. Broken lines indicate those participants who regressed.

Figure 6
Individual learners’ average L2 English vowel intelligibility scores before and after training in response to Voice 2 /b,pV/ stimuli. Broken lines indicate those participants who regressed.
DISCUSSION

The goal of this study was to develop a CAPT application that incorporated training features reflecting a current understanding of how L2 pronunciation develops. In particular, it attempted to address constraints stemming from interactions between L1 and L2 categories, while also increasing the quantity and quality of phonetic experience beyond what is typically available to adult learners.

To limit the effect of L1 transfer, English vowel-training stimuli were created using CV syllable frames. This meant that learners could hear the target vowels in a familiar context. Using real English words with final consonants would have introduced a confounding variable, since Mandarin speakers might be distracted by the presence of a coda consonant, which is prohibited in Mandarin syllables with similar initial consonants. The effect of the learners’ L1 was also addressed by including twenty native speakers in the training stimuli set. This meant that while some tokens may have provided learners with evidence that certain English vowels were comparable to Mandarin categories, the probability that there would be at least some evidence for new category formation where relevant was greatly increased.

To increase the quantity and quality of phonetic experience available to learners, several strategies were used. First a HVPT technique was employed, providing variation in terms of the number of native speakers providing input, but also providing manageable variation in the context in which vowels were learned (i.e., both bV and pV). This type of within-category variation provides learners with many opportunities to begin noticing perceptual cues necessary for the discrimination of English vowels. Second, the training program was designed to help learners direct their attention to the relevant phonetic information by enhancing the information available (i.e., modified training sets) and limiting other distracting information (i.e., avoiding the use of orthography by using nautical flags as symbols to represent sounds, and using items embedded in highly controlled phonetic contexts, rather than random words that might orient learner attention away from form and towards meaning).

The impact of the application on pronunciation accuracy was tested both before and after training and assessed by native speaker judges. The results indicate that when designed using a principled, research-based approach, computer-mediated training in the perception of L2 English vowels can improve speech intelligibility without explicit practice in production. The partial eta squared values indicate a medium effect of training on pronunciation.

The CAPT program developed for this study resulted in improved intelligibility scores not only in response to English vowel productions elicited using a voice that had previously been heard in training, but also in response to productions elicited using a novel voice. These results suggest that the program helped learners isolate relevant phonetic cues to vowel identity that were then generalizable to new speakers.

Somewhat unexpectedly, this study indicates that perceptual training in one phonetic environment can also lead to improved speech intelligibility in a different phonetic environment. While training learners to perceive Canadian English vowels in /bV/ and /pV/ syllables resulted in significant improvement in the training context, significant improvement was also found in the novel /zV/ and /sV/ contexts. These contexts differ from the training context along two phonetic dimensions, place and manner of articulation. In contrast, perceptual training in /bV/ and /pV/ contexts did not result in significantly improved intelligibility scores in the /gV/ and /kV/ contexts, despite the latter contexts only differing from the training context along one dimension, place of articulation. One possible explanation is that unlike
/g-k/, both the /b-p/ and /z-s/ pairs are produced near the front of the mouth. Consequently, transitions from these anterior consonants into following vowels would have a similar effect on the vowels, while transitions from the /g-k/ pair into following vowels would be quite different, and lead to a difference in perception.

A somewhat disappointing finding was that differences in the quality of the training stimuli used did not lead to differences in learning outcomes across the two training groups. In one sense, however, it might be reasonable to interpret this as a positive result. It suggests that even if unmodified stimuli are used, when the stimuli are highly variable, and when the training program promotes the orientation of learners’ attention toward the phonetic properties of the stimuli, further enhancement of the stimuli is unnecessary. That is, since learners are so focused on the task, even less than ideal training items will provide the sort of input that can promote learning.

The results can also be interpreted with reference to Flege’s (1995) Speech Learning Model (SLM). As outlined at the beginning of the article, the SLM predicts that learners will have the most difficulty learning sounds that are either similar but not identical to Mandarin categories, or those sounds that are unlike any Mandarin categories. Based on Thomson et al.’s (2009) comparison of Mandarin and English vowel inventories, the English vowels /i/, /e/, /a/, /o/, /ʊ/ should be relatively easy for Mandarin speakers to transfer from Mandarin. In contrast, /ɪ/, /ɛ/, /æ/, /ʌ/, /u/ are expected to be difficult. Figures 3 and 4 indicate that with the exception of /u/, most improvement as a result of the training provided in this study occurred in the vowels expected to be the most difficult, while for the vowels that were deemed to be easily transferrable, performance was already near ceiling prior to training. This indicates that the training affected precisely those vowels that learners find the most difficult to learn in the absence of training (see Munro & Derwing, 2008). At the same time, the results of the training provided in the current study indicate that some learners did not successfully acquire these difficult vowels. Longer training might reveal some learners require greater exposure. Conversely, a longer training period might reveal that some learners are unlikely to successfully acquire these vowels, despite training.

CONCLUSION

While some argue that a language teacher is indispensible, and can never be replaced by a computer (Secan, 1990), it might also be argued that CALL, and CAPT in particular, should become indispensible, never to be replaced by a teacher. Computers can and should be more than tools that allow learners access to language knowledge and practice beyond the classroom – a role to which they have often been limited (e.g., Chapelle, 2001; Seferoğlu, 2005; Wang, 2004). Current technology has the potential to offer more than classroom-based instruction on a monitor. Instead, CAPT should offer targeted teaching of language features in ways that traditional classrooms are unable to provide. The potential for CAPT to incorporate innovative research-based techniques is enormous, and still in its infancy. By incorporating a research-based technique into a CAPT application, the study in this paper provides preliminary evidence of what such techniques can achieve.

At the same time, there are limitations to the study reported here. Although it has been demonstrated that under highly controlled conditions, this type of training can lead to rapid gains in the intelligibility of L2 pronunciation, it is uncertain precisely how far such training will generalize. The fact that learning transferred from /bV/ and /pV/ contexts to /zV/ and /sV/ contexts provides promising evidence that transfer is possible, despite Flege’s (1995) claim to the contrary. However, transfer did not occur in /gV/ and /kV/ contexts. It is also uncertain whether learning will transfer to real words, and to the perception and production
of continuous speech. Further research is needed to determine how far such learning can extend, and to develop techniques that promote transfer to as many contexts as possible.

Although found to be successful, some elements of the program’s design may still have constrained its efficacy. For example, the use of nautical flags was intended to prevent the association of training stimuli with previously learned orthographic labels, for which learners may have had an inaccurate sound-symbol match. In contrast, it is possible that using non-orthographic labels may have introduced a layer of complexity that actually detracted from the goal of focusing attention on phonetic information, while using known orthographic labels might have facilitated learning. Thomson and Isaacs (2009) found evidence that in some cases, reference to orthographic information may actually promote more intelligible L2 pronunciation.

Another limitation of the technique demonstrated here is that as with any CAPT approach, it requires access to both computer hardware and software. Many language programs and learners do not have such access. One possibility is to create web-based versions of such programs, allowing access from any machine, and across platforms. Although this still requires computer hardware, the program itself would be more readily accessible to users, at home, at school, or in an internet cafe. A web-based application would also allow endless research possibilities, as teachers and researchers could collaborate remotely, monitoring the effect of perceptual training and its impact on pronunciation, in order to improve future iterations of the software. Mobile technology is another possible context for implementation. Uther, Uther, Athanasopoulos, Singh and Akahane-Yamada (2007) report that employing a Java-based version of perceptual training on mobile devices such as cell phones can result in perceptual improvement comparable to that achieved using computers. Their study was limited to training the /l/-/r/ contrast, however, making it difficult to generalize their results to a larger context. Also, while mobile technology is appealing, Joseph and Uther (2009) note that limitations are imposed by the memory capacity of many mobile devices. For fully developed applications of the sort described in the current study, the demand on memory can become quite substantial.

Designing a program such as the one used in this study may, at first glance, seem beyond the capability of most language teachers. In fact, the basic principles behind the training demonstrated here can be implemented using platforms that many ESL teachers and computer lab instructors already use. For example, the functionality needed for listening to sound files and selecting from among response alternatives is available in many popular Learning Management Systems, such as Moodle or Sakai. Using such systems, instructors could design instructional resources for pronunciation instruction that incorporate high variability input, and immediate feedback – the two principle features of the technique described in this study. Furthermore, the database of speech recordings need not be as large as the twenty speakers used in the study reported here. Most research investigating HVPT has relied on only three to six talkers, yet still yields promising results.

Finally, it is hoped that this study will provide language instructors with knowledge that can help them critically evaluate other CAPT applications. With a basic understanding of how L2 speech perception and production develops, more reliable predictions about the potential of commercially available programs can be made. This knowledge can also lead instructors to develop creative means for enhancing the efficiency of CAPT programs. For example, they will be better prepared to encourage learners to use those components of CAPT applications that are most likely to lead to success, while avoiding features that may be nothing more than unproven gimmicks. Ideally, however, researchers, teachers, and programmers should work together to design effective CAPT tools that can be made more widely available.
NOTES
1 These were a subset of participants from a larger study examining L2 English vowel development (Thomson, 2007). Only these participants were assessed for their English vowel pronunciation.
2 Praat is software designed for phonetic analysis, manipulation, and experiment creation, and is freely downloadable at www.praat.org.
3 As part of a larger study, participants were also exposed to the same stimuli on two additional occasions, but were not given feedback as to the accuracy of their responses.

REFERENCES


ACKNOWLEDGEMENTS

The author is grateful to Terrance Nearey for programming the computer application used in this research, and to Tracey Derwing for her helpful feedback. He is also grateful for the participation of staff and students from NorQuest College. Finally, he is thankful to the anonymous reviewers for their constructive criticism. The Social Sciences and Humanities Research Council of Canada provided funding for this study.

AUTHOR’S BIODATA

Ron Thomson teaches undergraduate and graduate courses in linguistics and English language teaching in the Department of Applied Linguistics at Brock University. His research interests include the development of L2 oral fluency and pronunciation, and the assessment of L2 accent. He is also involved in the ongoing development of a web-based English pronunciation training application, using a High Variability Phonetic Training technique.

AUTHOR’S ADDRESS

Dr. Ron I. Thomson
Department of Applied Linguistics
Brock University
St. Catharines, Ontario
L2S 3A1, Canada
Email: rthomson@brocku.ca