Directions for the future of technology in pronunciation research and teaching

Mary Grantham O’Brien, Tracey M. Derwing, Catia Cucchiarini, Debra M. Hardison, Hansjörg Mixdorff, Ron I. Thomson, Helmer Strik, John M. Levis, Murray J. Munro, Jennifer A. Foote, and Greta Muller Levis

1 University of Calgary | 2 University of Alberta | 3 Simon Fraser University | 4 Radboud University | 5 Michigan State University | 6 Beuth University Berlin | 7 Brock University | 8 Iowa State University

This paper reports on the role of technology in state-of-the-art pronunciation research and instruction, and makes concrete suggestions for future developments. The point of departure for this contribution is that the goal of second language (L2) pronunciation research and teaching should be enhanced comprehensibility and intelligibility as opposed to native-likeness. Three main areas are covered here. We begin with a presentation of advanced uses of pronunciation technology in research with a special focus on the expertise required to carry out even small-scale investigations. Next, we discuss the nature of data in pronunciation research, pointing to ways in which future work can build on advances in corpus research and crowdsourcing. Finally, we consider how these insights pave the way for researchers and developers working to create research-informed, computer-assisted pronunciation teaching resources. We conclude with predictions for future developments.

Keywords: pronunciation, technology, data, computer-assisted pronunciation teaching, research, future directions, automatic speech recognition, intelligibility, comprehensibility
1. Introduction

In August 2016, the Pronunciation in Second Language Learning and Teaching (PSLLT) conference featured a Roundtable, generously sponsored by Language Learning, on the role of technology in pronunciation research and instruction. Several eminent scholars described their own work and engaged in a discussion of how best to advance second language (L2) pronunciation research and teaching with technology. The organizers chose this theme because although technology is used extensively in pronunciation research, and is a growing component of language classrooms, some researchers and teachers express discomfort with new technological innovations: some implementations are difficult to use, and others are seen as unwelcome replacements for instructors. The goals of the Roundtable were to highlight some exceptional L2 research that makes use of technology, and to explore ways in which the pronunciation field can be enhanced by technology for researchers, teachers, and language learners. This article is based on the Roundtable discussions, which addressed the nature of pronunciation data; the user-friendliness of pronunciation technology; computer-assisted pronunciation teaching (CAPT); software development; and involvement of teachers in the appropriate use of technology to improve their students’ perception and production. The goals of this paper are to summarize the current state of technology in pronunciation research and teaching and to provide directions for the future. As such, it serves as an introduction to this special issue that focuses on cutting-edge research employing technology.

In recent decades, some researchers have advocated a move away from prioritizing native-like speech as a goal for L2 learners and toward striving for intelligible, comprehensible speech (Levis, 2005). As such, native-like accuracy is no longer viewed as a realistic goal for most adult L2 speakers, but clear speech is. Technology is one tool that may help both teachers and learners achieve this goal. Despite many calls for increased use of technology in pronunciation teaching and research (e.g., Levis, 2007), only limited progress has been made. A key requirement is cross-disciplinary skills, an issue which cannot be readily addressed without the collaboration of experts from diverse areas.

2. Current uses of technology in pronunciation research

Technology has been used in pronunciation research for decades, from phonograph records to sophisticated digital tools. Current uses range from simple recording and playback of speech to Automatic Speech Recognition (ASR) for assessment and feedback on learner productions. Some software enables
researchers to directly observe the acoustic features of speech (through waveforms, spectrograms, and pitch tracks) and to analyze, synthesize, and manipulate speech. Other applications facilitate computer annotation of speech corpora (e.g., Phon, <https://www.phon.ca>; Strik & Cucchiarini, 2014). A range of experimental platforms enables researchers to gather data from participants. Developments in hardware such as ultrasound technology permit visualization of learners’ articulatory movements (Abel et al., 2015). Many types of speech technology were originally developed for people with communicative disabilities, but have been extended to research in L2 speech; thus, the two fields can inform each other (Strik, 2012). Current technology holds great promise for both research and teaching. Below we focus on the benefits of these technologies and the hurdles faced by researchers and teachers wishing to implement them.

2.1 Freeware

Two examples of freeware that can be effectively applied in pronunciation research are Praat (Boersma & Weenink, 2017), for phonetic analysis, and Anvil (Kipp, 2001, 2014) for video annotation research. Together they provide a time-aligned display of the acoustic and visual (e.g., gestural) elements of a speech event. However, they both require considerable expertise. Employing these tools, Hardison (2016) examined the relationship between the visual and acoustic beats in the natural speech of native and advanced nonnative teachers of English. The study investigated the teachers’ polyrhythmic sequences (i.e., those with different rhythms for speech and gesture), which are important because they highlight focused information. Speech segmentation, annotation, and analysis were performed in Praat using waveforms, pitch tracks, and spectrograms. The annotations were then imported into Anvil, where they were temporally aligned with the video. (See Figure 1)

As seen in Figure 1, annotations are placed in a window (the annotation board), the configuration of which requires a coding scheme. Although several examples of coding schemes are available on the Anvil website, researchers may need to write their own to meet specific needs, as was the case in Hardison’s study.

Shown in the figure are annotations of the speech and gestures of an advanced L2 English (L1 Korean) instructor telling a class about the ways that novice language teachers at another institution had prepared an EFL lesson. A student mentioned some activities that would have been helpful for the EFL class and asked why they had not been included. The instructor responded quickly with the statement “they can’t invest more time,” with emphasis on “can’t” and “more.” “They” referred to the teachers, and the investment of time referred to lesson preparation.
In a post-study interview, the instructor indicated that she had interpreted the student's comment as an unfair criticism of the novice teachers.

The first track shows a waveform, with intermediate phrase and words annotated in the next two tracks. To analyze the speech-gesture relationship, features of the pitch-accented syllables, such as vowel onset and pitch peak, were marked, though some limitations exist on this capability. On the *words* track, the vertical boundary immediately to the left of "an" (in "can't") and "ore" (in "more") marks the vowel onsets; the boundary to the right of each of these elements denotes the pitch peak. The accompanying beat gesture sequence involved the up-down movement of one of the instructor’s hands. The highlighted *apex* track shows the apex of each beat (the downbeat, in this case). Comments on the *head* track include mention of when the instructor moved her head from side to side, starting before the word "can’t," and when she tilted it forward with "more." The onset of such movements often precedes their lexical affiliates, alerting the perceiver to anticipate important information. Brow raise also accompanied the emphatic information.

The comparison of native speaker (NS) and nonnative speaker (NNS) gestural-speech alignment provides insights into ways in which NNSs may be helped to produce utterances that are easier for students to process. Although the outcomes were illuminating, there were challenges in utilizing these tools.

Because a great deal of pronunciation research is carried out in laboratories, researchers use a range of platforms to conduct experiments and gather data from participants. These include freeware like *PsychoPy* <http://www.psychopy.org/>
and OpenSesame <http://osdoc.cogsci.nl/> or commercial software, such as E-Prime <https://www.pstnet.com/eprime.cfm> or Superlab <http://www.cedrus.com/superlab/>. Commercial tools tend to be more user-friendly, but expensive. However, even commercial programs require some technical understanding to conduct training experiments.

### 2.2 Automatic speech recognition (ASR)

Other technologies cannot be used off the shelf at all, but are developed specifically for use in L2 pronunciation research. Considerable research has addressed the usability of ASR technology in L2 pronunciation work. (For an overview, see Cucchiarini and Strik, 2018). Early studies primarily dealt with pronunciation assessment and showed relatively strong correlations between human pronunciation ratings and machine scores (Cucchiarini, Strik, & Boves, 2000a, 2000b, 2002; Neumeyer, Franco, Digalakis, & Weintraub, 2000; Witt & Young, 2000). While such high correlations may be sufficient for pronunciation assessment, use of ASR technology in the context of L2 pronunciation teaching and learning requires more detailed measures to localize and identify pronunciation errors and provide focused corrective feedback immediately after an utterance. For this reason, more recent speech technology research has developed and investigated ASR-based measures for pronunciation error detection (Hu, Qian, Soong, & Wang, 2015; Qian, Meng, & Soong, 2012; Strik, Truong, de Wet, & Cucchiarini, 2009; Van Doremalen, Cucchiarini, & Strik, 2013). The aim of this research is to determine which ASR-based measures best reflect human judgments of pronunciation quality. This is usually done by evaluating the performance of the algorithms on corpora of learners’ speech to determine whether the automatically detected pronunciation errors resemble those annotated by human raters. In general, speech technology research on pronunciation error detection is not well known among L2 pronunciation researchers, who may consider it too technical and inaccessible. Moreover, pronunciation researchers are primarily interested in improving L2 learners’ intelligibility and comprehensibility, but they have not yet collected sufficient amounts of representative and reliable data (speech recordings with corresponding annotations and judgments) indicating which errors affect these speech dimensions and which do not. These data are essential to train ASR algorithms to assess L2 learners’ intelligibility and comprehensibility. (See also the section on speech corpora.) It is clear that ASR research can profit from collaboration among speech technologists and pronunciation researchers, to ensure that advanced technological approaches are employed to address appropriate issues in L2 pronunciation and that systems are developed to best support language learners.
2.3 Text to speech

Text-to-speech (TTS) systems, developed for individuals with severe speech impairment, the visually impaired, or those with reading disabilities, artificially convert written texts into spoken language. In L2 contexts, TTS systems may be used to model the pronunciation of words, sentences, and longer texts for L2 learners. The work by Mixdorff and colleagues in the area of L2 prosody (e.g., Hilbert, Mixdorff, Ding, Pfizinger, & Jokisch, 2010) is largely based on methodology developed and employed for predicting and evaluating speech prosody in TTS systems. As much as natural sounding prosody is a prerequisite for natural sounding, intelligible speech synthesis, language learners must master certain aspects of L2 prosody to communicate effectively. In speech synthesis, the aim is to develop computational prosodic models that predict prosodic factors such as duration and pitch contours from text (Hilbert & Mixdorff, 2011). Also necessary are analytic methods for evaluating the goodness of prosodic models. Although such evaluation still draws heavily on human judges, the goal is to create quantitative descriptors based on the speech signal that are reliably correlated with perceptual judgments. The gold standard for speech synthesis in these measurements is the natural speech signal.

Work on Australian English spoken by Vietnamese learners (Ingram, Mixdorff, & Kwon, 2009; Mixdorff & Ingram, 2009) and Canadian English by speakers of Cantonese (Mixdorff & Munro, 2013) employed the Fujisaki Model (Fujisaki & Hirose, 1984), which produces a continuous pitch contour (i.e., fundamental frequency or F0) from a parsimonious representation. First, syllabic duration analyses were carried out on sentences, with a group of native English speakers as the comparison group. Subsequent modeling revealed that listeners’ judgments of fluency and prosodic goodness were strongly correlated with the rhythmic similarity between individual learners and the reference group. In the realm of F0, English L2 learners employed a larger F0 range than the native speakers and did not differentiate as much between primary and secondary accents. These findings suggest that modeling of the rhythmic properties of L2 speech can predict listeners’ perceptions of fluency and overall prosodic accuracy.

To examine the relative impact of prosodic and segmental deviations in the comprehensibility of L2 speech, voice morphing and prosodic transformation on utterances of learners at various levels of proficiency were conducted, followed by perceptual evaluations from phonetically untrained listeners. Voice morphing is a speech signal processing technique that produces an acoustic continuum between two existing recordings, whereas prosodic transformation transplants the prosodic parameters from one utterance (the native speaker’s in this case) to another (the learner’s). Similar to the results on speech synthesis, accuracy
of syllable durations had a stronger influence on the perceived prosodic quality than pitch. As could be expected, utterances from learners whose unmodified speech was rated poorly benefited the most from transplanted prosody (Mixdorff & Munro, 2013). What these findings suggest is that getting the timing or the rhythm right is more important than pitch.

Like ASR, TTS is still not at a stage that most L2 pronunciation practitioners and researchers are comfortable with. However, in the next few years, we predict that this technology will become more user-friendly, and will be incorporated into programs that individualize instruction for learners.

2.4 The cloud

Another option for making technology more accessible to researchers is the “cloud.” Rather than existing locally on the end-users’ computers, cloud-based applications are stored on Internet servers. This not only makes them more readily accessible, but also allows for any updates to be immediately implemented for all end-users. Another advantage of cloud-based applications is that they allow for large-scale data collection (data harvesting). While the benefits of cloud-based programs are compelling, creating such applications is effortful, and can be costly. To be done properly, it requires that someone with coding and design experience be a part of a research team that develops a given product, especially if it is to be user-friendly. Thomson’s (2018) English Accent Coach is an example of one cloud-based pronunciation teaching and research tool. Its researcher mode, available from the author upon request, provides a user-friendly graphical interface that allows researchers to design highly controlled experiments in which learners proceed through a set of predetermined tasks to test their perception of specific English segmental contrasts and to conduct controlled training (with or without feedback) on English segmental contrasts. For the participants (learners), resulting experiments have the identical look and feel as the publicly accessible and user-friendly version of the web application, with which many learners may already be familiar. A limitation of this niche web-based application is that it allows testing and training only with its built-in database of stimuli. While this database is extensive, including over 3000 possible words and syllables, spoken by 30 distinct talkers, it does not allow researchers to add their own tokens to the existing training sets. Presumably additional databases and experiments could be designed in the future focusing on learners’ perceptions of suprasegmental contrasts such as lexical stress assignment.

The increasing use of many research tools has often been bottom-up and organic, such as videos explaining specific uses of Praat and other tools, sharing of scripts, etc. There is a clear need for extensive on-going training demonstrating
and documenting uses of technology for research purposes. For example, the 2017 and 2018 PSLLT Conferences featured two half-day workshops focused on the use of technology (analysis of corpora using *Praat* and an introduction to *JsPsych* and *MTurk*). Much more is needed, perhaps using technology tools such as Webinars or wikis to make training broadly accessible.

3. **Data collection**

Data collection in L2 pronunciation research could be improved. Much research on second language pronunciation instruction focuses on the acquisition of a relatively small number of pronunciation features, most often individual speech segments (Thomson & Derwing, 2015; see Chun, Jiang, Meyr, & Yang, 2015; Trouvain & Gut, 2007, among others, for important exceptions). This is not to say that segments are not important, but a wider range of speech phenomena should be examined. Researchers tend to record learners’ productions of highly-monitored (often very short) read-aloud tasks as opposed to more spontaneous tasks, which are likely to reflect normal speech to a greater degree (Thomson & Derwing, 2015). Furthermore, there is considerable homogeneity among the participants who produce speech samples as well as those who rate them: They tend to be relatively similar in terms of age, education level, and first languages (English is predominantly the language taught or the learners are native English speakers) (Lee, Jang, & Plonsky, 2015).

L2 pronunciation research has been criticized for the use of relatively small sample sizes and consequent low statistical power (Lee et al., 2015). Furthermore, the majority of extant studies consider native-like pronunciation (i.e., the lack of a foreign accent) to be the most appropriate target for L2 learners, in spite of more than a century of appeals from L2 specialists to emphasize intelligible speech (Abercrombie, 1949; Derwing & Munro, 2005; Sweet, 1900).

Based on recent reviews of the literature, future pronunciation research in general would benefit from the following:

- a wide range of both segmental and suprasegmental foci;
- larger subject pools;
- participants at various ages and stages of the L2 learning process, from several demographic backgrounds (i.e., less focus on university students);
- participants from a range of L1 backgrounds;
- data on participants’ language learning history and proficiency in the L2;
- a range of target languages (in addition to English);
- a variety of contextualized speaking tasks;
– more emphasis on longitudinal rather than cross-sectional data;
– comprehensibility ratings (as opposed to accent ratings) on the collected data;
– an examination of interaction effects of multiple interventions.

3.1 Spoken learner corpora

Darcy, Ewert and Lidster (2012) discuss the lack of systematicity in deciding which pronunciation features to teach or when and how to teach them. However, the development of CAPT resources targeting intelligibility and comprehensibility entails pinpointing aspects of L2 speech that actually lead to problems in understanding. Catford (1987) proposed, and Munro and Derwing (2006) provided preliminary empirical support for, the importance of functional load for English segments. Functional load is a measure of how many minimal pairs two segments distinguish in otherwise homophonous words in a language. For example, there are many more /k/~ /h/ contrasts that discriminate among words (e.g., <cat> ~ <hat>), than the number of minimal pairs distinguished by /d/ and /z/ (e.g., <ding> ~ <zing>). Prosodic factors must also be considered in any determination of intelligibility; for instance, both lexical stress assignment (Field, 2005; Zielinski, 2008) and primary sentence stress assignment (Hahn, 2004) play important roles in English comprehensibility.

Each of these studies is exploratory and small. To gain a greater understanding of intelligibility and comprehensibility issues for English as an L2 alone, far larger studies are necessary, employing many more listeners and speakers in a wide range of contexts. Thus, although research investigating pronunciation among L2 learners of English is common, there is no overall big picture bringing the component parts together. Moreover, we know even less as a research community about what makes other, non-English L2 speech difficult to comprehend (Lee et al., 2015; Thomson & Derwing, 2015).

To address these lacunae, we propose the creation of accessible, fully annotated spoken learner corpora. Like other corpora, such as the Child Language Data Exchange System (CHILDES) database (MacWhinney, 2000), spoken learner corpora are “electronic collections of natural or near-natural data […] assembled according to explicit design criteria” (Granger, Gilquin, & Meunier, 2016, p.1). Most spoken corpora are associated with corresponding sound files and they include little markup of pronunciation features, partly because of the time-consuming nature of transcription and the lack of automated prosodic analysis (Staples, 2015). If the sound files exist, they are often not made public. Despite moves to address this lack of availability (Staples, 2015), little has changed yet (for examples, see French learner corpora at <http://www.flloc.soton.ac.uk/> and Spanish learner corpora at <http://www.splloc.soton.ac.uk/>.)
Ballier and Martin (2016) differentiate among three types of spoken learner corpora: mute corpora, which are transcriptions of spoken data; speaking corpora, which associate sound files with transcriptions of speech, such as the Speech Accent Archive, (Weinberger, 2017); and phonetic corpora, which align annotations with recorded speech signals. Often phonetic corpora include annotations at the levels of phrases, words, segments, and intonation, among others. Tools such as Phon (Rose & MacWhinney, 2014) are available to aid researchers in the annotation of spoken data. Access to many existing learner corpora is restricted, and, due to the transient nature of internet resources, many of the links to corpora are broken. For these reasons, it may be prudent to create new corpora in English and other target languages according to detailed, standardized criteria. A model of such a corpus is JASMIN-CGN (Cucchiarini, Driesen, Van Hamme, & Sanders, 2008), which is a repository of contemporary Dutch as spoken by children of different age groups, elderly people, and L2 learners of Dutch with different L1s in the Netherlands (NL) and Flanders (FL). The L2 learner sub-corpus of the main corpus contains 12h 34m (NL) and 9h 15m (FL) of speech recordings from children between 7 and 16, as well as 15h 01m (NL), 8h 02m (FL) of adult speech. Orthographic transcriptions are available for all speech recordings as well as automatically generated broad phonetic transcriptions and part-of-speech tagging. The speech material consists of equal proportions of read speech (phonetically rich sentences and short stories) and extemporaneous speech in the form of human-machine dialogues. Although read speech does not meet the requirement that a corpus contain “natural or near-natural data,” this may be a necessary adjustment for the development of L2 pronunciation corpora. For example, the Speech Accent Archive (Weinberger, 2017) could be considered a corpus with such an adjustment. However, even with this adjustment, pronunciation corpora must not be built entirely with read speech. The JASMIN-CGN has been used for data-driven studies of pronunciation errors made by Dutch L2 learners, for adapting and testing ASR technology for Dutch L2 learners (Van Doremalen, Cucchiarini, & Strik, 2010) and for developing and testing ASR-based pronunciation error detection algorithms (Van Doremalen, Cucchiarini, & Strik, 2013; Van Doremalen, 2014).

Based on the insights of learner corpus research, we make the following recommendations to researchers who wish to create spoken learner corpora:

- Obtain permission from participants to include the data in a speech corpus at the data collection stage (Mackey & Gass, 2005);
- Include, at the very least, basic metadata about participants (age, L1, age of learning the L2, target language immersion experience, a commonly under-
stood measure of proficiency (e.g., CEFR level), and the recording context (date, location, and language of the interview)). Researchers creating their own corpora should refer to The Oxford Handbook of Corpus Phonology (Durand, Gut & Kristoffersen, 2014);

- Take a multi-site approach, with several colleagues building the corpus according to agreed-upon criteria. This cannot be said too strongly. Gilquin (2015) argues for strict criteria in the design of a spoken corpus, saying that, “in the case of learner corpora, design criteria are even more crucial given the highly heterogeneous nature of interlanguage” (p. 16).

Analyses of phonetic corpora can serve as the basis for a robust description of the pronunciation of a given language and its intra- and inter-variety and register variations, especially for perception training. Teaching and researching intelligibility must involve attention to both production and perception (Levis, 2005), even though perception is often neglected in favor of the learner’s role in producing intelligible speech (Thomson & Derwing, 2015).

3.2 Collecting data through crowdsourcing

One approach to enhancing sample sizes and broadening the range of listeners from across the globe for pronunciation research is crowdsourcing, using platforms such as Amazon Mechanical Turk (MTurk) <https://www.mturk.com> or CrowdFlower <https://www.crowdflower.com>. These enable researchers to tap the intuitions of non-expert listeners, who can transcribe what they hear (for intelligibility measures), rate their ease of understanding of speech samples (comprehensibility), and answer questions about other impressions of the speech samples. Although it requires relinquishing the control inherent in laboratory-based studies, crowdsourcing enables researchers to obtain substantial numbers of listeners who are more representative of the population at large (Cooke, Barker, & Lecumberri, 2013, p. 140) and to address the previously-discussed concerns about small sample sizes.

Eskenazi (2013) recommends that researchers take the following into consideration when implementing crowdsourced perceptual studies:

- Ensure that the listeners’ audio works well, for example, through the transcription of a small set of audio data as a pretest;
- Require listeners to complete a prequalification task. A brief language proficiency test would establish that listeners are indeed speakers of the target language;
- Keep tasks simple, ensuring that listeners understand them.
Because outliers could have a significant impact on the results of crowdsourced studies, we also suggest that researchers consider both making use of large sample sizes (i.e., over 100 listeners) and employing outlier detection statistics to discard data from listeners who have not completed the task as instructed.

A combination of big data in the form of both fully annotated phonetic learner corpora and intelligibility transcriptions and comprehensibility ratings from hundreds of listeners will enable researchers to determine which aspects of pronunciation from a range of L2s – as produced by speakers from a variety of L1s – affect both intelligibility and comprehensibility. This information could then be used to create pronunciation technology training materials targetting aspects of L2 speech that interfere with understanding. Given the tremendous variation across learners from the same L1 (Smith & Hayes-Harb, 2011), there will always be a need for instructors to help learners identify where the focus of their attention should be. However, by conducting large-scale investigations of comprehensibility and intelligibility, instructors and learners will have a much clearer starting point for choosing which aspects of pronunciation are most important.

4. Computer-assisted pronunciation teaching (CAPT)

CAPT, typically the use of specially-designed technology to train pronunciation features, offers a number of benefits to supplement classroom-based pronunciation training: CAPT systems are tireless and consistent, they facilitate learner autonomy and individualized instruction, and they can expose learners to a variety of speech models (Hardison, 2004; Levis, 2007; O’Brien, 2011). CAPT users have more access to pronunciation practice focusing on their specific needs than do learners who participate in more traditional programs, particularly if teachers direct them to appropriate software and monitor their progress (Thomson & Derwing, 2015). CAPT can also utilize simple technology not originally designed for teaching or researching pronunciation, but widely available to many learners and easy to repurpose. For instance, Bueno Alastuey (2010) had her students interact on Skype with other learners and native speakers of the target language. Many L2 learners opt to practice pronunciation on their mobile phones – another readily available technology. However, in a comprehensive examination of smartphone applications for pronunciation, Foote and Smith (2013) found that most focused on a single aspect of pronunciation (largely individual segments), which, in many instances, may not have had an effect on intelligibility or comprehensibility. The task of selecting or recommending appropriate CAPT resources may thus be difficult for teachers, many of whom lack adequate training in how to identify learner
problems or to effectively evaluate pronunciation apps or other technology (e.g., Baker, 2014; Foote, Holty, & Derwing, 2011; Murphy, 1997).

Central to the development of effective CAPT resources are the following key components: A range of speakers and meaningful input types; the inclusion of perception tasks; the ability to identify errors affecting intelligibility and/or comprehensibility; clear, concise, real-time feedback (including visual feedback) consistent with human feedback; information on how a learner’s pronunciation differs from the target; guidance on how to improve comprehensibility; and the ability to set measurable goals and track improvement for individual learners (Chun, 2013; Neri, Cucchiarini, Strik, & Boves, 2002; Pennington, 1999).

4.1 Carrying out research on CAPT

Although research has shown that computer pronunciation training can be effective, many investigations cover only a short training period and address a small number of features that are tested in planned and/or trained utterances (Thomson & Derwing, 2015). A meta-analysis on the effectiveness of CAPT resources has indicated smaller effects than face-to-face instruction (Lee et al., 2015). Moreover, it is often difficult to attribute changes in pronunciation abilities to CAPT training, as some studies do not include control or comparison groups (Thomson & Derwing, 2015). Despite the potential benefits of CAPT, to fully understand its potential effectiveness, additional training studies with sounder designs are necessary, as outlined below.

4.2 Examples of effective CAPT and research

CAPT offers researchers an excellent opportunity for gathering data while at the same time training students. In addition, it is possible to equip CAPT systems with logging capabilities so that interactions between users and the system are stored for tracking learner trajectories over long periods of time. This makes it possible to analyze not only the outcome of practice, but also the learning process. For instance, English Accent Coach (Thomson, 2018) (described earlier) trains learners to perceive English vowel and consonant contrasts with the assumption that perceptual learning will result in improved pronunciation (Thomson, 2011). The program employs High Variability Phonetic Training (HVPT), which exposes learners to a range of productions of individual speech segments, thus encouraging the establishment of robust perceptual representations (Lively, Logan, & Pisoni, 1993). Users hear recorded stimuli containing target segments and click on the phonetic symbol representing the sound they believe they heard. They then receive immediate feedback on the accuracy of their responses. The training stimuli are highly
variable because they incorporate the same target sounds produced by multiple talkers in multiple phonetic contexts. Researchers can both harvest information about every click made by learners and also observe their accuracy scores on particular items over time. In concert with production recordings made before and after a course of training, these data can be used to assess the efficacy of the training, and, given their scale, can shed light on the developmental process itself. For instance, in one study in which 15 L2 learners of English completed 40 perceptual training sessions over a period of two months, Thomson (2016) observed that improved perception of English vowels in one phonetic context (e.g., after an /h/) did not transfer to perceptual knowledge of the same vowel categories after other consonants (e.g., after /b/, /d/, or /g/). In another study, Thomson and Derwing (2016) determined that training learners to perceive English vowels in isolated syllables (many of which were non-words) promoted significant improvement in the pronunciation of real words containing those vowels. However, perceptual training on the vowels in the target real words did not lead to improved pronunciation of those words, at least not over the course of the study. This research was possible only because of the scale of data that could be collected using a cloud-based application. Obtaining the same results using more traditional methods would entail a much longer, piece-meal process.

4.2.1 Visual tools

Hardison (2004), among others, has shown the effectiveness of visual tools such as pitch displays which permit learners to visualize similarities and differences between their pre- and post-training productions, and between their L2 productions and those of native speakers.

Mixdorff, Külls, Hussein, Shu, Guoping, and Si (2009) engaged in three years of development and evaluation of a CAPT system for German learners of Mandarin. The team included a Chinese Studies expert charged with testing the current state of the system in the classroom. Learning materials followed the widely used Practical Chinese Reader (Liu et al., 1981). The system is based on speech signal analysis using phone and tone recognizers that were optimized during the project. Both the speech signal envelope and the pitch contour are visually represented in real-time, in contrast to many common pronunciation training systems. Training is performed in a loop through imitation of examples, typically single words or short phrases. As soon as the learner stops talking, the imitation is evaluated in terms of the correctness of initials, finals, and tones of each syllable. During the course of the project, several perception and production studies were performed to determine the most relevant problems for German learners of Mandarin (Hussein et al., 2011). Tone and segmental correctness were assessed on readings and imitations of tokens of varying complexity. This assessment was
always carried out by an expert teacher, a group of native listeners and by the ASR system. Students were happy with a tool they could relate to when practicing their pronunciation despite many problems involved with lab software such as installation issues on certain computer systems, crashes, and false hits. In a final controlled evaluation with defined training units, the researchers noted improvement, especially in the area of tonal correctness, which is most problematic for Western learners of Mandarin (Do et al., 2012).

Waveforms and spectrograms can also be sources of informative feedback. Learners vary in terms of the amount of instruction and guidance they require to use these displays. For target languages such as Japanese, in which consonant and vowel duration is contrastive, training with waveforms compared to auditory-only input or no input (controls) facilitated improvement in perceptual accuracy by L1 English learners, whether administered as web-based training (Motohashi-Saigo & Hardison, 2009) or lab-based training (Okuno & Hardison, 2016). In addition, generalization of performance was found for novel stimuli with transfer to production improvement.

Although visual displays are still viewed as somewhat arcane for many teachers, Chun et al. (2015) used Praat visual representations of pitch curves in a regular class of learners of Mandarin to teach tones over the course of nine weeks. Human ratings of pre- and post-test learner productions indicated a significant improvement; furthermore, the majority of students reported that seeing model pitch contours and comparing them with their own helped them to produce a more target-like utterance.

For contrasts involving spectral (vs. temporal) differences, L1 English learners of L2 Spanish found that training with spectrograms helped them to distinguish their own stop consonant productions [b, d, g] in intervocalic position from the target forms [β, ð, ɣ] (Olson, 2014). Visual displays are well received by many learners, who often ask why they are not a common supplement to in-class communicative activities. A probable reason is that teachers often lack the training to use such technology.

4.2.2 ASR tools

Cucchiarini, Neri and Strik (2009) developed and evaluated the ASR-based Dutch CAPT system addressing 11 frequent and persistent pronunciation errors in L2 Dutch. To measure identification accuracy, relevant metrics such as precision and recall were calculated. Precision indicates the extent to which errors detected by the system are real errors, while recall measures the proportion of errors detected by the system among all errors made. Precision varied between 79% and 82% and recall between 75% and 86%. Thus, pronunciation error detection was relatively accurate; furthermore, the students appreciated working with the system. Practice did not
significantly improve overall pronunciation quality, probably because it focused on only 11 errors and was relatively short (two sessions of 30–60 minutes); however, the reduction in the number of mispronunciations was significantly larger for the experimental group using the ASR-based system than for a control group.

In the ‘Development and Integration of Speech Technology into Courseware for Language Learning’ (DISCO) system (Strik, Colpaert, Van Doremalen, & Cucchiarini, 2012), ASR technology was developed to provide practice and feedback on pronunciation, morphology, and syntax in L2 Dutch. (See Figure 2.) User evaluations were positive (Van Doremalen, Boves, Colpaert, Cucchiarini, & Strik, 2016) and accuracy in mispronunciation detection was high (Van Doremalen, 2014). These ASR systems were designed to detect and provide feedback on errors in L2 learners’ speech that had been previously selected by Dutch L2 teachers based on criteria such as frequency, persistence and salience. Thus the feedback was not based on accuracy alone (as many other ASR systems are).

![Figure 2. Screenshot of a pronunciation exercise in DISCO. The feedback indicates pronunciation errors for /y/ and /ei/, represented by the graphemes ‘uu’ and ‘ij’, respectively](image)

While such systems are effective, their development requires a team with the right mix of expertise to design and develop a user-friendly application that does
not crash regularly, operates smoothly, and does not require a high degree of technical knowledge on the part of the user. It is possible to use cloud-based applications, which do not require local updating when computer hardware or operating systems change, which can render locally installed applications obsolete. The downside is that cloud-based applications require regular updates to ensure compatibility with new versions of web-browsers as well as a stable, high-speed internet connection, which although increasingly commonplace, is not always assured.

Lee et al. (2015) and Thomson and Derwing (2015) call for more research on the effectiveness of CAPT resources. We recommend an emphasis on the following:

– the development of protocols for technology-based research;
– the inclusion of both control (no training) and comparison (different training) groups;
– a focus on both segmental and suprasegmental features that affect comprehensibility (as we continue to pinpoint what those are);
– longitudinal studies lasting a minimum of three months, but preferably longer;
– delayed post-tests to determine whether training has a lasting effect;
– investigations of effects of training that extend beyond the trained materials;
– an operationalization of improvement focusing on the ability of human listeners to understand the speech samples, rather than native-like accuracy.

5. Encouraging collaboration

The creation of large research projects to develop effective CAPT resources requires long-term collaboration among a range of specialists from applied linguistics, pedagogy, phonetics, computer science, and engineering. One prerequisite for such work is the opportunity for researchers to come together to establish a common understanding of the issues. A second is long-term, stable funding. With regard to the first prerequisite, not only is it important to establish a research agenda, but it is also critical to learn what teachers and learners know about pronunciation and what kinds of support they desire and require to succeed. At a minimum, this means a more detailed understanding of the needs of native and nonnative teachers, and of teachers who are well-trained and those who are less well-trained. As for the second prerequisite, projects like these are not without barriers. Not only is carrying out large-scale, collaborative research expensive, but a major stumbling block to such collaboration is inherent in some university structures where interdisciplinary research is encouraged, but individual grants and publications are still valued more highly.
We make the following recommendations for researchers hoping to move forward with large-scale CAPT projects:

- consistent engagement with classroom teachers and their students, who should receive some benefits for having participated in the research endeavor;
- interdisciplinary courses involving undergraduate and graduate students from diverse fields in all aspects of the research project. At Iowa State University, for example, externally funded interdisciplinary projects in applied linguistics involve computer scientists, pronunciation researchers, computational linguists, and graduate and undergraduate students. Creating an interdisciplinary team in which graduate students work together with and mentor undergraduates leads to a richer involvement across disciplines.
- open, modular software architecture that makes components of computer applications reusable by the same or other programmers in the creation of new applications. This could be accomplished, for example, through releasing source code through open-source licenses or through open and well-documented application programming interfaces. Another important benefit of a modular software architecture would be the de-coupling of the underlying learner models and instructional logic from the graphical user interface (GUI). For example, an application created to teach English sounds to adults could be reworked with a different GUI for a more gamified application aimed at children.

6. Future directions

Looking to the future, we predict a number of exciting breakthroughs for pronunciation research generally, and for CAPT specifically. These fall into three main areas: software advances, individualization of instruction, and ASR systems that are better aligned with the goals of communicative language instruction. First, we anticipate that TTS systems will gain in functionality and that a wider variety of software will be developed to appeal to a range of learning styles. Second, we expect to see more modular forms of training. The traditional approach to pronunciation instruction has involved the entire phonological system. Research has clearly shown that language learners do not need to become experts in phonetics and phonology and that many potential learning targets do not require training at all, either because they are quickly learned without intervention, or are unlikely to impede intelligibility or comprehensibility. Moreover, some targets that are initially difficult show substantial improvement without training (Derwing & Munro, 2015; Munro, Derwing, & Thomson, 2015), and many expected
L1-specific difficulties are not difficult for all learners (Munro et al., 2015; Qian, Chukharev-Hudalainen, & Levis, 2018). Still other difficulties show particular resistance to learning (Hahn, 2002; Pennington & Ellis, 2000). In other words, pronunciation training must address difficulties that are specific to individual learners and are likely to make a difference in communicative success. This calls for flexible training modules that can be successfully implemented using technology. Finally, advances in ASR should result in better analysis of errors, more precise feedback, and enhanced opportunities for the recognition of more spontaneous speech samples.

In terms of the field at large, we propose the establishment of a professional body that could speak with authority about L2 pronunciation development and combat the wealth of misinformation that has been spread, such as the confusion between accent reduction and pronunciation instruction. (See Derwing & Munro, 2015 for a detailed examination of accent reduction practitioners who have limited or no knowledge of research in the field or have no sense of effective teaching.) An interdisciplinary entity is necessary to ensure that the appropriate technologies are employed for the right purposes, that these technologies are adapted to specific goals by using appropriate data and insights, and that technology-based approaches are pedagogically sound. Ironically, anybody with access to Skype and Paypal can open a business as an accent reduction coach, offering lessons, often at exorbitant rates, and often with no demonstrable outcomes. Lippi-Green (2012) compares such accent coaches to individuals who “have developed a miracle diet and charge money for it” (p. 229). At a more local level, universities have a responsibility to offer adequate training for language teachers to ensure that they can comfortably teach pronunciation, using technology judiciously to individualize instruction to learner needs (Foote et al., 2011).

We are at a crossroads now (where we have been stuck for some time). Technology offers outstanding possibilities, but only if implemented while taking into account which research questions should be addressed, and how best to incorporate what we already know about pronunciation learning into pedagogical aids for learners.

Acknowledgements

The authors thank Scott Jarvis, Executive Director of Language Learning, and the sub-committee of the Language Learning Board of Directors, for funding the Roundtable event at PSLLT 2016. We are grateful to the Social Sciences and Humanities Research Council of Canada for a grant awarded to the first two authors in support of PSLLT 2016 and open access to this paper.
References


Qian, X., Meng, H., Soong, F. (2012). The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training. *Proceedings of Interspeech 2012* (pp. 775–778), Portland, OR.


**Address for correspondence**

Mary Grantham O’Brien
School of Languages, Linguistics, Literatures and Cultures
University of Calgary
C216 Craigie Hall
2500 University Dr. NW, Calgary, AB T2N1N4
Canada
mgobrien@ucalgary.ca
Co-author information

Tracey M. Derwing
Dept of Educational Psychology
University of Alberta
tderwing@ualberta.ca

Catia Cucchiarini
C.cucchiarini@let.ru.nl

Debra M. Hardison
Michigan State University
hardiso2@msu.edu

Hansjörg Mixdorff
Computer Science and Media
Beuth University Berlin
mixdorff@beuth-hochschule.de

Ron I. Thomson
Dept of Applied Linguistics
Brock University
rthomson@brocku.ca

Helmer Strik
w.strik@let.ru.nl

John M. Levis
English Department
Iowa State University
jlevis@iastate.edu

Murray J. Munro
Department of Linguistics
Simon Fraser University
mjmunro@sfu.ca

Jennifer A. Foote
English Language School, Faculty of
Extension
University of Alberta
jfoote@ualberta.ca

Greta Muller Levis
English Department
Iowa State University
gmilevis@iastate.edu