Annual Review of Applied Linguistics (2007) **27**, 184–202. Printed in the USA.
Copyright © 2008 Cambridge University Press 0267-1905/08 $12.00
doi: 10.1017/S0267190508070098

9. COMPUTER TECHNOLOGY IN TEACHING AND RESEARCHING PRONUNCIATION

John Levis

Pronunciation, long on the periphery of applied linguistics research and pedagogy, continues to grow in importance because of its central roles in speech recognition, speech perception, and speaker identity. Pronunciation-related issues such as comprehensibility, accent, and the mutual intelligibility of varieties of world Englishes are central to many questions in applied linguistics. This calls for a sophisticated understanding of how technological tools that have long been used to shed light on phonological categories can be applied to teaching. Research into computer-assisted pronunciation teaching (CAPT) suggests that both researchers and pronunciation teachers increasingly make use of technology to answer key questions, to ensure that claims are defensible, and to develop theories and practices that more closely match acoustic reality. This article reviews three key areas where computer technology and pronunciation intersect: (1) appropriate pedagogical goals and the measurement of improvement; (2) the ability of CAPT to give useful, automatic feedback; and (3) the use of technology in diagnosing pronunciation errors. This article concludes with recommendations for key technological competencies needed by any researcher or teacher who examines pronunciation-related issues.

Teachers and researchers have had high hopes for computer-assisted pronunciation teaching, or CAPT, for several decades (Molholt, 1988). Yet it remains in its infancy in many ways. Chun (2007), for example, did not classify pronunciation among the topics published in two prominent computer-assisted language learning (CALL) journals from 2001 to 2006, except perhaps as a subset of other topics such as speaking, listening, or computer-mediated communication (CMC). This is odd, because the use of computers is almost ideally suited to learning pronunciation skills. Computers can provide individualized instruction, frequent practice through listening discrimination and focused repetition exercises, and automatic visual support that demonstrates to learners how closely their own pronunciation approximates model utterances. In a foreign language teaching environment in which few teachers receive adequate training in teaching pronunciation (Breitkreutz, Derwing, & Rossiter, 2002;

Burgess & Spencer, 2000; Kawai & Hirose, 2000; MacDonald, 2002; Murphy, 1997), and in which little class time appears to be available for focused work on pronunciation because of emphasis given to other skills, CAPT seems tailor-made to meet a critical need. Despite this great promise, effective commercial CAPT applications are less innovative either in pedagogy or use of computer technology than one might expect.

A variety of experimental software applications and studies into how various features of pronunciation might best be taught have demonstrated the flexibility and value of CAPT. A wide variety of pronunciation-related features have been examined, including general pronunciation quality (Seferoğlu, 2005); speech rate, fluency, and liveliness (Hincks, 2005); vowels and consonants (Lambacher, 1999; Neri, Cucchiarini, & Strik, 2006a; Wang & Munro, 2004); vowel lengthening and pitch accents (Hirata, 2004; Kawai & Hirose, 2000); intonation (Cauldwell, 2002; Chun, 1998; Hardison, 2004a; Kaltenboeck, 2002; Levis & Pickering, 2004); and English stress timing (Coniam, 2002). CAPT seems to also work for children, if designed with that audience in mind (Mich, Neri, & Giuliani, 2006). The great majority of these studies demonstrate that CAPT, when constructed wisely, can be both effective and flexible in addressing pronunciation instruction.

Yet difficulties remain, and CAPT rarely makes its way into books that present state-of-the-art views of CALL. Some of the difficulties that CAPT has faced are pedagogical, some are technological, and some are related to teacher preparedness. Pedagogically, a significant gap often exists between CAPT applications and goals advocated by current pronunciation theory and pedagogy, such that CAPT applications can look suspiciously like traditional, drill-oriented pedagogy in new clothing. Although key researchers have strongly called for pedagogy to be based on empirical findings (Derwing & Munro, 2005), such a connection is all too infrequent in CAPT applications. Applications that are explicitly grounded in theory, like *Streaming Speech,* are rare. Because applications are often not grounded in appropriate descriptions of pronunciation (Pennington, 1999), they are not consistently able to measure improvement in meaningful ways. Technologically, CAPT systems often suffer from difficulties in giving learners adequate, accurate feedback and an inability to provide accurate and automatic diagnosis of pronunciation errors. Both of these areas relate to the use of automatic speech recognition (ASR) for accented speech. Finally, a substantial number of teachers are not able to make effective use of applications because of both a lack of training in pronunciation and in the use of technology.

This article begins by examining the importance of pedagogical goals. Next, I examine whether CAPT actually works and in what ways it works. In other words, do learners actually improve when using CAPT? Then I look at research on feedback and the use of ASR in CAPT before turning to the question of teacher preparation. In doing so, I offer suggestions of areas of technology that teachers and software developers must be able to understand to be more intelligent users of CAPT.

A writer's own experience and perspective always influences how a review such as this is written. Because I am far more expert on the pronunciation side of this topic and am primarily a consumer of CAPT rather than a developer, my review will necessarily take the perspective of a pronunciation practitioner who is interested in what CAPT can do for students as well as for teachers who will be unlikely to author their own software. There is a great deal more research available on critical areas such as speech recognition technology that I could have reported on, but did not, partly because much of it is more technical than I found useful, and partly because much of it asks questions (such as how spoken dialogue systems can be made more reliable) that are not directly relevant to CAPT. On the pronunciation side of this review, I try to avoid pronunciation issues that I do not see as directly relevant to CAPT design. I also have not specifically reviewed many individual applications. Detailed reviews of these can be regularly found in journals such as *Calico Journal, Language Learning and Technology,* and *Computer Assisted Language Learning*. A recent worthwhile review of CAPT that takes a wider scope than my review, including reviews of many individual applications, can be found in O'Brien (2006), described in the annotated references at the end of this article.

Appropriate Pedagogical Goals

The great psychologist Abraham Maslow was reported to have said, "If the only tool you have is a hammer, you tend to see every problem as a nail." Similarly, CAPT applications are tools to meet instructional goals, and the tool should be appropriate to the job. Before CAPT applications can be effective, the job they need to do should be specified through appropriate pedagogical goals, a requirement that is not always given much emphasis by designers (Neri, Cucchiarini, Strik, & Boves, 2002). CAPT, like CALL in general, should be based on an explicit theoretical approach to language teaching. Although general CALL principles may be useful, specific applications to pronunciation are still needed. Neri, Cucchiarini, Strik, and Boves (2002) suggested there is a critical need for specific CAPT principles for two reasons. First, many CAPT applications do not show any evidence of such principles, and second, general CALL principles often do not seem to have been written with pronunciation in mind. They argue that applications should do the following:

1. Provide a substantial amount of meaningful input, including the use of multiple models and accurate articulatory instruction.
2. Give learners a reason and desire to practice through rule-oriented practice and realistic materials.
3. Provide immediate, useful feedback, especially for those features that are most important for intelligibility.

Likewise, Pennington (1999, p. 434) wrote that most CAPT programs showed very little understanding of phonology, the range of acceptable variation in pronunciation, or how to apply this knowledge to teaching. Some of Pennington's principles for the design of CAPT materials are that CAPT systems should

1. Establish baseline, reference accents for instruction.
2. Set measurable goals and performance targets.
3. Be designed to build skills from easier to more challenging exercises.
4. Link pronunciation to other aspects of communication.
5. Raise users' awareness of how their L1 phonological systems differ from the system of the target language.

These recommendations call for pedagogical explicitness in CAPT design so that applications do not simply use a new medium to do the same old exercises or employ technology without pedagogical grounding. Neri, Cucchiarini, and Strik (2002a) reported that "many authors describe commercially available programs as fancy-looking systems that may at first impress student and teacher alike, but eventually fail to meet sound pedagogical requirements. . . . These systems . . . look more like the result of a technology push, rather than a demand pull" (p. 442). In other words, the programs exploit the possibilities of the computer interface without considering whether the tool is appropriate to the job.

Some studies explore new ways to teach that exploit the computer's strengths. Wang and Munro (2004) suggested expanding the range of pedagogical techniques to include those typically employed for research but not usually for teaching. In training learners to recognize difficult vowel contrasts, they used identification tasks (rather than discrimination of minimal pairs) and synthesized speech to expand learners' perceptual space, making the learners more sensitive to the range of phonetic variation in the pronunciation of vowel phonemes. Such an expansion of perceptual space usually requires the use of multiple models and voices, a practice advocated by many researchers. However, in the absence of multiple voices in CAPT applications, the researchers say that speech synthesis may be very useful.

Multiple speech models, another way to provide variation in input, is evident in some recent software for English, such as *Connected Speech* (Protea Textware, 2007), *Streaming Speech* (Cauldwell, 2002), and *Listening to American Accents* (http://www.speechinaction.com/). *Connected Speech,* a program originally published in Australia, is available with nine regionally varied speech models (both male and female voices) for three markets: the United Kingdom, North America, and Australia/New Zealand. *Streaming Speech* is available for British and American English.

Pronunciation teaching is subject to two overlapping and conflicting foci (Levis, 2005): a focus on accuracy (involving instruction on all possible aspects of pronunciation, with little attempt to prioritize attention) and a focus on intelligibility (involving instruction for only the elements of pronunciation that are critical for communicative success, while ignoring or deemphasizing those that are not). Traditionally, pronunciation instruction has been strongly accuracy oriented, assuming that all possible vowel and consonant sounds in the target language should be taught. Not only does this comprehensive coverage increase the teaching and learning burden, there is very little evidence that such accuracy is either possible for most adult learners or that it is desirable. All errors of pronunciation are not equal in their impact on listeners (Jenkins, 2000; Munro & Derwing, 2006).

Because some errors impact listener understanding more than others, it is now generally agreed among pronunciation researchers that the goal of instruction is not perfection but rather being good enough to achieve communicative goals. "Good enough" is usually referred to as intelligibility, that is, whether a speaker is understandable (e.g., Dalton & Seidlhofer, 1994; Morley, 1991). Intelligibility in the general sense, however, includes two main types of understanding, which researchers call intelligibility and comprehensibility. In the more specialized sense, *intelligibility* refers to whether listeners can adequately decode the words pronounced by a speaker in context. *Comprehensibility,* on the other hand, refers to whether listeners can understand (or believe they can understand) the message communicated by a speaker (see Munro & Derwing, 1999, for discussion of these terms). Applied to the teaching of pronunciation, both intelligibility and comprehensibility address listeners' ability to understand at the micro level of word identification and at a communicative level of meaning processing.

Intelligibility as a Principle for CAPT

The goals of accuracy and intelligibility are often incompatible in non-CAPT instruction but there is no reason for them to be so in CAPT applications. Although certain errors may hinder understanding no matter what the language background (such as word stress; see Field, 2005), many deviations that cause misunderstandings are language specific, such as /l/ and /r/ mispronunciations for Japanese speakers but not for speakers of French. In addition, some causes of misunderstanding may even be dependent on the listeners' perceptual backgrounds. For example, Jenkins (2000) discussed how, in communication in English between a Swiss German and Japanese speaker, the perceptual system of the German speaker and the pronunciation of the Japanese speaker both were critical in communication breakdowns.

Applied to CAPT design, these findings suggest that CAPT applications should, unless built for learners of a particular language group, be complete enough to allow teachers or learners to pick pronunciation features most likely to affect their pronunciation needs. Suprasegmental features such as word stress, length variations, and intonation should be included for all learners (Derwing & Rossiter, 2003). For segmentals, CAPT systems should guard against the tendency to do everything. The system should assist learners and teachers in prioritizing pronunciation topics by channeling learners toward typical vowel and consonant errors for their language backgrounds. For example, a Korean learner of English would, after setting up a user profile, be directed to pronunciation topics that are problematic for Korean learners. Such information about typical errors for a variety of language backgrounds can be found (for learners of English) in a variety of sources, such as Swan and Smith (2001) and Avery and Ehrlich (1992).

Even better than this rather crude channeling mechanism would be an error diagnostic informed by language specific filtering. A diagnostic component in a CAPT system should include perception elements in which learners identify and discriminate among problematic sounds. It should also include production elements in which learners have to produce potentially problematic elements. Their pronunciation of words and phrases could be scored by means of ASR technology, which works relatively well for very limited tasks. Finally, the diagnostic could also include a prediction component. Some pronunciation errors are a result of confusion about the connection between sound and spelling (e.g., the -ed endings, multisyllabic word stress, vowel pronunciations). A prediction component could identify the cause of these types of errors and provide appropriate remediation that addresses the cognitive basis of the problems.

Improvement and CAPT

A key question is whether CAPT works. In non-CAPT settings, this has often been asked about pronunciation instruction. Adult learners of foreign languages rarely achieve native-like accents. Because most nonnative speakers will have a noticeable accent with or without instruction, should pronunciation be taught? Asking about improvement is actually asking multiple questions. First, does instruction lead to improvement on the features trained in the CAPT system? Second, does learning transfer to novel contexts? Third, does learning last over time? And finally, does learning in one pronunciation topic lead to improvement in other areas? Is there a positive spillover?

Regarding the first question, CAPT seems to be effective in improving pronunciation accuracy. For a wide variety of pronunciation skills, learners improve through the use of well-designed CAPT instruction. Hirata (2004) taught learners Japanese pitch and duration contrasts. One experimental group learned the contrasts at the word level, while another learned the same words embedded in sentence-level practice. Both groups improved their production of these contrasts, with improvement being greater for those who learned the contrasts embedded in sentence-level practice. She stated that "developing abilities to produce and perceive L2 contrasts in sentences instead of only in an isolated word contest, is an important factor for L2 learners coping with real world situations" (p. 372). In another study of Japanese, Kawai and Hirose (2000) found that subjects more successfully made phonemic contrasts between short and long segments, with the greatest improvement on the short segments. Wang and Munro (2004) trained learners to pay attention to vowel quality rather than durational cues for the perception improvement of three vowel contrasts among learners of English, /i/-/ɪ/, /u/- /ʊ/, and /ɛ/- /æ/. They found that subjects increased their perception performance on all the contrasts. Hardison (2004a) found that learners who practiced French intonation, with NS models provided as feedback, improved their intonation production. Neri, Cucchiarini, and Strik (2006b) found that learners of Dutch improved their production of the trained Dutch segments.

The second and third questions, regarding transfer to untrained contexts and improvement over time, also point to the benefits of CAPT. Hirata (2004) found that pitch and duration training for learners of Japanese transferred to words that were not part of the training, that is, that "subjects who participated . . . acquired a generalized ability to produce and perceive Japanese words contrasting in pitch and duration" (p. 372). Kawai and Hirose (2000) found that phonemic lengthening, especially for vowels, transferred to novel minimal pairs. Wang and Munro (2004) found that the perceptual improvements in distinguishing vowel contrast generalized to new words. In contrast, Neri et al. (2006b) found that training on vowels and consonants did not cause any change in segments that were not part of the training. For vowels and consonants, at least, this suggests that changes follow training, and that without training, changes are less likely to occur. In regard to changes lasting over time, only one study of CAPT addressed this. Wang and Munro (2004) retested the subjects' perceptions 3 months later and found that the improvement had lasted.

The final question is the most intriguing. Given that CAPT instruction works for the trained items, is there any spillover to improvement in other areas? There is evidence that improvements in perception can lead to improvements in production (Bradlow, Pisoni, Akahana-Yamada, & Tohkura, 1997). Hirata (2004), however, found evidence that production improvements also can lead to improved ability to perceive contrasts. It is possible that improvement in one area spills over to improvements in the other, although the extent to which this is so needs to be tested for various pronunciation features. Hardison (2004a), in more evidence of spillover, found that not only did learners improve their intonation, but they also showed improvement in two areas that were peripheral to the treatment: lexical recall and segmental accuracy. She suggested that "as they [the subjects] became more confident with this aspect [prosody] of their language production, they were able to notice other elements such as liaison and their production of specific sounds" (p. 48).

Visual Feedback

Perhaps the central issue in CAPT is the provision of adequate feedback. Although this can be done in various ways, the most common methods are visualization and through automatic speech recognition. Visual feedback includes "the graphical display of a native speaker's face . . . [and] the vocal tract" both of which seem to improve learners' ability to identify words and produce more native-like speech timing (Ehsani & Knodt, 1998, p. 63). This kind of feedback is usually found in the study of phonetics, as in the phonetics page at the University of Iowa Web site (http://www.uiowa.edu/∼acadtech/phonetics/). However, the best-known CAPT visual displays are spectrograms, waveforms, and pitch tracings. These visual displays have been advocated for a long time (Anderson-Hsieh, 1994; Cranen, Weltens, de Bot, & van Rossum, 1984; Spaai & Hermes, 1993) and are still being used in research and teaching.

Applications of many phonetic displays to pedagogy have not always been successful because the displays are not transparently interpretable and may require more specialized training than a teacher is able to provide. Spectrograms and waveforms are widely argued to not be worth the pedagogical time and energy, at least when used as the sole source of feedback (Ehsani & Knodt, 1998; Neri, Cucchiarini, & Strik, 2002b). Pitch tracings, however, are relatively iconic, with rising, falling, and level lines on the display usually corresponding to rises, levels, and falls in a speaker's voice pitch. Although they require some training to interpret, they appear to be useful pedagogically (Chun, 2007; Hardison, 2004a) despite an ongoing need to define which intonational features are most important for intelligibility (Chun, 2007).

Nonetheless, spectrograms are still being advocated by some recent work. Coniam (2002) used spectrograms to raise language teachers' awareness of the relatively stress-timed nature of American English, and by extension, other inner circle Englishes. The teachers that Coniam described were native speakers of Cantonese in Hong Kong, and the outer-circle English variety spoken there, Hong Kong English, is relatively syllable-timed. Coniam used spoken dialogues from local TV shows involving well-known characters speaking both Hong Kong and American English. The utterances he used for analysis were analyzed through spectrograms, which demonstrated that Hong Kong English has a greater frequency of energy peaks than American English speech. The visual representation of the energy peaks in the spectrograms was successful in helping teachers understand this difference in the two varieties. The success of this awareness raising exercise appears to happen because Coniam did not ask much of the spectrogram, focusing only on the relatively iconic visual displays of syllable energy while also providing a significant amount of scaffolding that identified the location of the words associated with each peak. That his subjects were language teachers was also significant. His goals were that these expert users begin to understand the nature of rhythmic differences rather than produce those differences. This more restrictive goal contributed to the successful use of spectrograms.

Lambacher (1999) indicated success with teaching English consonants to Japanese learners using spectrograms. For difficult consonant features and contrasts (i.e., aspiration, r/l, nasals, s/ʃ, f/h, and s/ɵ), he illustrates how the contrasts look on visual displays built into the CAPT program. Lambacher offered no data on the effectiveness of his approach, but it is interesting to note that he did not advocate using spectrograms alone. For example, Lambacher taught aspiration with reference to a spectrogram but also through standard pronunciation teaching tricks like blowing away a piece of paper.

Pitch tracings, or models of the fundamental frequency of speech, while also requiring training to interpret, appear to be much easier to use successfully. Commercially available software such as Kay Elemetrics VisiPitch or its more technically oriented cousin, the Computerized Speech Laboratory (CSL), as well as free programs such as WASP and PRAAT, all provide the technology to represent intonation visually. Most also allow learners to compare their own production to a model utterance by overlaying their utterance's pitch tracing on that of the model. Pitch tracings are ubiquitous in CAPT studies and recommendations. Although pitch is most often modeled for sentence intonation (e.g., Hardison, 2004a, 2004b), pitch tracings have also been used to teach word-level pitch accents in Japanese (Hirata, 2004). Levis and Pickering (2004) recommended using pitch visualization to teach paratones, that is, initial extra-high pitch levels that mark discourse topic shifts in connected speech. *Streaming Speech* uses the discourse intonation model of David

Brazil to teach the intonation and pronunciation of natural connected speech, and discourse uses of pitch are advocated by Chun (1998; 2007).

ASR Feedback

Despite ongoing issues regarding the use of visual feedback tools, the central question in CAPT feedback is whether automatic speech recognition (ASR) can effectively provide immediate feedback that allows learners to know which parts of their pronunciation are correct and which are not. Conceptually, the problem is this: Do the words and phrases pronounced by nonnative speakers (NNSs) match the models of native speaker speech on which the ASR is based? If not, can the NNS be given global feedback on the mispronunciations (e.g., "Your score is 62%. This means that you often have words that are mispronounced.") or specific feedback on particular sounds or prosodic categories? Finally, how should feedback be given about the nature of the mispronunciation? Assuming these questions are successfully answered, another concern is whether ASR can be configured to provide accurate instruction to remedy errors. This last question, however, is a CAPT design issue and not unique to ASR. If automatic feedback is adequate, the provision of instructional remedies should be relatively straightforward.

Neri, Cucchiarini, Strik, and Boves (2002) wrote that "ideal systems should always include an option to provide feedback by means of ASR technology, so that the user can receive immediate information on his/her performance" (p. 458). Because ASR applications have been successful with native speakers of English, especially in the use of voice recognition for word processing, it might be assumed that they could be adjusted to provide feedback nonnative speakers. Some applications, such as recent versions of Dragon Naturally Speaking (ScanSoft, 2005), are 95% or more accurate for native-speaking English users (Ehsani & Knodt, 1998; Pogue, 2004). However, accuracy for these programs, when used by advanced proficiency but accented nonnative speakers of English, drops to near 70% (Coniam, 1999; Derwing, Munro, & Carbonaro, 2000). The reason for this drop is that commercial ASR dictation programs are not built for nonnative speakers. In fact, "much of the progress in the last 15 years in acoustic modeling [for ASR] is based on more detailed modeling, creating sharper and sharper distributions for narrower and narrower classes. This is diametrically opposite of the tolerance and robustness required for non-natives" (Van Compernolle, 2000, p. 76). Research on improving ASR effectiveness for nonnative speech continues to be of interest (e.g., Gorozny, Rapp, & Kompe, 2004; Oh, Yoon, & Kim, 2007), but much of this research is currently of little direct relevance to CAPT. The failure of ASR systems (developed for native speakers) to successfully handle nonnative speech does not mean that ASR is useless in pronunciation training. Rather, it means that new ASR systems with goals more appropriate to language learning contexts are needed.

Feedback for CAPT should be consistent with human feedback (Kim, 2006; Neri, Cucchiarini, & Strik, 2003), be immediate (Neri, Cucchiarini, Strik, & Boves, 2002), be pertinent and correct (Eskenazi, 1999), be given in a form that students can make use of (Kawai & Hirose, 2000), include information about when goals have been reached (Kawai & Hirose, 2000), and possibly suggest ways to address errors (Neri et al., 2002a).

These requirements reflect a real problem. It is difficult for ASR systems to pinpoint learner errors in read speech even where the system knows what words are intended (Truong, Neri, de Wet, Cucchiarini, & Strik, 2005). Precise feedback is even harder for suprasegmental errors because all languages make use of the same acoustic categories of pitch, duration, and intensity for differently organized prosodic systems. Also, disagreements continue about how the prosodic systems should be measured (for a still relevant discussion of this, see Ladd, 1980). ASR systems measure prosody without reference to linguistic organization, leaving a gap between human perceptions of correctness and what the systems can do.

The problem of wrong feedback is enormous in CAPT ASR systems. Neri, Cucchiarini, Strik, and Boves (2002) wrote that "erroneous feedback is a common problem. . . . Patently wrong error detection can be so frustrating for the student that [some] recommend using implicit rather than explicit judgemental feedback" (p. 458). In other words, because of the difficulty of ASR pinpointing particular errors, feedback that does not try to be too precise is likely to be more successful (e.g., "The word *system* was not pronounced correctly.") than feedback that singles out particular sounds for correction (e.g., "The first vowel in *system* was pronounced wrongly."). This is the difference between implicit and explicit feedback. By telling the truth at a sufficient level of generality, the feedback will at least not be wrong (Delmonte, 2000). ASR feedback is best when its strengths are used and its limitations are recognized (Ehsani & Knodt, 1998). Neri et al. (2003) reflected the current state of ASR-based feedback when they stated that to be effective, ASR-based feedback should "keep the recognition task as simple and as limited as possible, by carefully designing the learning activities" (p. 1158).

Does ASR-based feedback work, given all its limitations? Feedback appears to be effective, but the picture may be complicated by the complex relationship between segmental accuracy and global ratings of improvement (cf. Derwing & Rossiter, 2003). Neri et al. (2006b) did an experimental study of CAPT effectiveness. Two experimental groups, one using CAPT and receiving feedback and one using CAPT without feedback (and one control group) were tested using a CAPT system focused on Dutch segmental contrasts that are difficult for L2 learners of Dutch. Global pronunciation quality after training was rated by six expert raters. All three groups improved after the treatment period, but there was no significant difference in the global ratings. In further analysis, the researchers found that the accuracy of the target sounds was significantly better for the group receiving feedback over the other groups, and that the group simply using the CAPT training was significantly better than the control group. They concluded that "the automatic feedback . . . was effective in improving the quality of the targeted phonemes" (p. 1985). The mismatch between the global ratings and specific improvement occurred because the targeted phonemes were much less frequent than other sounds, and presumably their impact on the global score was not great enough to be noticed in the global ratings.

Automatic Diagnosis of Pronunciation Errors

Automatic diagnosis of pronunciation, both globally and related to specific errors, is a hopeful application of ASR technology. One study says that "speech recognition technology is key to the automatic evaluation of pronunciation quality" (Neumeyer, Franco, Digalakis, & Weintraub, 2000, p. 83). The trick is to know which door is opened by the key, a door to general pronunciation scores or a door that specifies the errors made by the nonnative speaker. Research indicates that the first door can be opened, while the second door remains tightly shut in search of the right key.

The standard against which ASR systems should be measured when diagnosing pronunciation quality is whether they sufficiently correlate with human judgments (Franco, Neumeyer, Digalakis, & Ronen, 2000). From this perspective, the accuracy of automatic scoring varies greatly depending on the type of evaluation task. At one end, Neri, Cucchiarini, Strik, and Boves (2002) found in examining one ASR-based system that only 25% of the actual errors were detected, and that a small number of correct pronunciations were called errors. These two types of errors, where errors are not detected and where correct pronunciations are labeled as errors, seem to be common to all ASR systems. At the more optimistic end of the evaluation spectrum, the use of certain automatic scores such as duration, log-posterior, and speech rate allow some ASR evaluation algorithms to approach the correlations found to exist between human raters, at least when there is enough speech material being tested (Cucchiarini, Strik, & Boves, 2000; Neumeyer et al., 2000; Rypa & Price, 1999). Usually, performance is better when a combination of parameters is measured rather than only one (Cucchiarini et al., 2000; Franco et al., 2000; Witt & Young, 2000). Although it appears possible to predict human ratings with ASR, the combination of automatic measures used and the quality and representativeness of the speech database to which the nonnative speech is compared cause the performance of different ASR studies to vary greatly. Cucchiarini et al. (2000), for example, found that two of their measures correlated well to human ratings of fluency and speech rate but not to ratings of overall pronunciation. Witt and Young (2000) found that their Goodness of Pronunciation score had to be modified by several other measures and models of the learners' L1 to get within shouting distance of human raters. These problems with predicting human ratings reflect the fact that machine scores attempt to correlate with humans, but machines do not, and perhaps cannot, hear like humans. Ladd, in discussing rhythmic organization in language, quoted a 1950s era researcher, Fred Householder:

> We can't hear noises repeated with fair regularity at more than a certain average frequency without grouping them rhythmically (as every subway rider can testify), and once a given pattern is established we will hear it over and over till some new irregularity breaks the rhythm and starts another pattern. In this domain variation of amplitude is, of course, important, but is easily dispensed with (by organ players and harpsichordist, for example). Pitch variation is less dispensable, but timing or interval variation is

> the most important factor of all. *Machines don't hear like people because people hear things that aren't there,* but the machines do hear very well all the factors which induce us to hear what isn't there. (Ladd, 1980, p. 26; my emphasis)

Progress on the disconnect between how humans perceive speech and how ASR models perception needs to be made before ASR can make more than incremental changes in accuracy. Scharenborg (2007), in a comparison of the goals and methods of research in human and automatic speech recognition, stated that "human listeners are able to use all information that is present in the acoustic signal to differentiate between phones and thus words, while ASR systems can only use the information that is encoded in the acoustic features. . . . ASR systems thus do not have available all information that is available to human listeners" (p. 341). If Householder is correct, human listeners also have other information available, the "things that aren't there." All of this makes human perception robust, allowing humans to be "far better at dealing with accents, noisy environments, differences in speaking style, speaking rate, etc." (Scharenborg, 2007, p. 344). ASR, on the other hand, is not robust, and ASR systems for nonnative speech can currently, at best, only approach human raters. That closeness, however, may be enough for many uses of automatic pronunciation evaluation.

Pinpointing Specific Errors

The key has not yet been found for the door that opens to automatically evaluating particular pronunciation errors, although this would be far more useful to language teachers than global measures of pronunciation quality. Telling a learner that their pronunciation is worth a grade of "C" is almost guaranteed to bring the question, "What should I work on?" Global measures offer no answer to this kind of question. Neri et al. (2003) stated that "recent research on ASR-based CAPT has . . . shown that this technology is not yet mature enough to provide relatable detailed diagnoses of pronunciation errors" (p. 1159).

Cucchiarini et al. (2000) correlated human ratings of Segmental Quality with ASR measures (which worked well for predicting human ratings of fluency and speech rate). They found that "segmental quality . . . is predicted most poorly on the basis of automatic scores, which is not a positive finding if we consider that segmental quality is the best predictor [for human raters] of overall pronunciation" (p. 116). Truong et al., (2005) constructed an ASR system that automatically measured the accuracy of three Dutch phonemes commonly mispronounced by learners of Dutch, /ɑ/, /ɣ/ and /x/. They found that errors in the pronunciation of /x/ were automatically detected most successfully, but that the ASR system was less successful identifying difficulties with the vowels. Nonetheless, the approach seems on target for automatic evaluation. Human raters don't need to hear every instance of a phonemic error to make the judgment that there is a systematic problem. In addition, even though learners do not always mispronounce sounds that are difficult, a system that can identify frequent errors could be very useful for learners and teachers. What is clear is that a lot of work needs to be done defining the parameters of the segmentals in a language and including them in ASR systems before precise feedback on individual segments is possible (for one example of this process, see Neri, Cucchiarini, & Strik, 2006a).

The success of an ASR system depends on the uses to which it is put. Global measures of pronunciation may be useful for some contexts, such as testing (e.g., Versant Testing, the test previously called PhonePass and Set-10). For the classroom teacher and the self-directed learner, however, there is a strong need for ASR systems that can pinpoint specific errors and provide help in addressing the errors (Neri et al., 2003).

Teacher Preparation and CAPT

What do teachers need to know to get the best use of CAPT applications? It seems clear to me that all teachers need to have a basic understanding of technology that is used in phonetics research. Teachers should be able to understand, at the very least, spectrograms, waveforms, and fundamental frequency contours, and they should be familiar with terms like low pass filtering, if only to be intelligent consumers. Freely available programs like PRAAT and WASP, or more costly options like the Computerized Speech Lab (CSL), should supplement any pronunciation training course. Teachers should also understand the strengths and limitations of ASR technology and be able to critique the use of ASR in a variety of language learning applications, perhaps through their own or their students' use of the applications. Third, teachers should become familiar with exercises that appear to be effective or not effective with a computer interface, and they should have the opportunity to develop and test their own computer-based pronunciation exercises through the use of basic CALL authoring tools.

Conclusion

For any teacher who thinks that pronunciation is essential, CAPT is immensely promising. We know that pronunciation is taught infrequently and unsystematically; this occurs for several reasons. A lack of teacher training leads to untrained and unconfident teachers; the varied needs of learners make group instruction irrelevant for some, especially in ESL contexts where classes have learners of many L1s and varied pronunciation needs; and pronunciation, in competition with other language skills, is simply a lesser priority in many classrooms and is often not taught. CAPT promises a way out of this bind, allowing teachers to have access to pronunciation teaching that hopefully goes beyond their own skills, providing individualized instruction and offering additional instructional time in a language laboratory or outside of class. Hardison, describing why the effectiveness of computer-assisted pronunciation training in her research study (on acquisition of prosody) cannot be directly compared to teacher-led instruction, gives a hint of the potential benefits of CAPT:

> Consider the following elements of the present computer-assisted training program that would need to be duplicated in an instructor-led approach. Feedback involved 30 sentences spoken by

> each of three talkers. We know that talker and stimulus variability contribute significantly to successful L2 speech training . . . therefore, in a non computer-assisted approach, three different instructors would be needed to provide feedback throughout data collection that, in the present study, required blocks of several hours set aside each day throughout the week for 3 weeks, and in total, spanned several months. Moreover, what feedback (also a significant factor in successful training) would be given by the instructors? Recall that segmental accuracy in this study was not a focus of training for the participants but part of the investigation of generalization; therefore, an instructor would have to restrict feedback to prosody only. Several more questions then arise. Could all instructors do **exactly** the same thing? If not, another variable enters the picture. In addition, how would feedback be provided by instructors on prosody only? The computer program provides a visual display in real-time and the opportunity to overlay the NS version on the learner's—a salient form of feedback that drew many positive comments from participants. . .While not an exhaustive list, the above points serve to emphasize that comparison of approaches, in general, is highly problematic. (Hardison, 2004a, p. 48)

Hardison argued, in effect, that CAPT has certain research (and teaching) benefits over traditional classroom instruction. First, CAPT is tireless. Teachers simply cannot provide the level of practice and feedback needed for many students to improve. Second, CAPT is consistent. It is always the same in its presentation of stimulus material and in the kind of feedback given. Teachers often are not. Third, CAPT provides variety, both in the numbers of voices used as models and in opportunities for visual feedback, especially in areas like pitch movement. Finally, CAPT offers the chance to meet varied individual needs more easily than any teacher can. It promotes learner autonomy in working on pronunciation, a critical factor in success. Hardison is not arguing that teacher-led instruction is unimportant. There is sufficient empirical evidence that both instructor-led teaching and CAPT both lead to pronunciation improvement. There is also evidence that a lot less pronunciation teaching is going on in classrooms than one would hope and that most teachers feel unprepared to teach pronunciation. CAPT is an opportunity to address these problems, not a denigration of the teacher's role. Rather than a false fight over an already too small piece of the language teaching pie, CAPT is a way to expand the pie so that more teachers and learners can enjoy their own use of spoken language.

ANNOTATED REFERENCES

Hardison, D. (2004a). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning and Technology, 8*(1), 34–52.

This article examines whether computer assisted prosody training leads to improved prosody, and as a bonus, to improved segmental accuracy and lexical recall. The treatment involved intensive practice of sets of 30 French sentences with varied intonational contours. Rather than being models for the subjects, native French speaker sentences were used as feedback to the subjects' initial production. The most provocative result of the study is the spillover effect of the prosody training. The intensive practice led subjects to notice segmental features that were not in focus. The prosody also stimulated substantial recall of the lexical content of the test sentences, suggesting that the subjects' lexical memory was improved by their prosodic memory built through the training. The study offers some thought-provoking implications for the value of suprasegmentals in pronunciation teaching.

Hirata, Y. (2004). Computer-assisted pronunciation training for native English speakers learning Japanese pitch and duration contrasts. *Computer Assisted Language Learning, 17*(3–4), 357–376.

This study, which examined how learners of Japanese acquired pitch and duration contrasts, is an excellent example of how CAPT can be used to suggest answers for research issues and directions for teaching. Japanese is a pitch accent language, with words distinguished by their patterns of H and L pitches. It is also a language with phonemic lengthening. The interaction of these two prosodic features allowed the researcher to examine prosody acquisition more fully than is usually the case in research studies. Training included some subjects receiving word-level and others sentence-level practice. Although the research questions lent themselves to minimal pair practice, the researcher did not include this as a treatment. Training was effective, and training of words in sentences was more effective than word-level practice.

Neri, A., Cucchiarini, C., & Strik, H. (2003). Automatic speech recognition for second language learning: How and why it actually works. *Proceedings of 15th International Conference of Phonetic Sciences*, 1157–2260. Retrieved July 7, 2007, from http://lands.let.kun.nl/literature/neri.2003.1.pdf

This article is an accessible, general treatment of the pluses and minuses of ASR. It highlights four key issues in the use of ASR. (1) Can ASR recognize nonnative speech successfully? (2) Can ASR provide a reliable assessment of pronunciation? (3) Can ASR be used to identify specific errors? (4) Can ASR provide remediation? These questions occur in one form or another in many of the articles addressing the use and effectiveness of ASR. For someone who is unfamiliar with the topic, this is a good gateway to further study.

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication, 30*, 83–93.

This article is a more complex yet still accessible examination of some of the important steps to verify ASR's accuracy in a CAPT program. It shows how machine scores and human scores are correlated in evaluating pronunciation, and it addresses the key to any ASR system, the corpus of speech (native and nonnative) on which recognition is modeled. There are some mathematical formulas, but the gist of the article is not dependent on them, as the authors explain clearly what the formulas represent.

O'Brien, M. (2006). Teaching pronunciation and intonation with computer technology. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching* (CALICO Monograph Series, Vol. 5, pp. 127–148). San Marcos, TX: CALICO.

This recent review of CAPT is a complement to my review. It offers principles to evaluate CAPT systems, and it reviews CAPT systems for a wide variety of languages. It focuses on three areas: basic pronunciation software, software with ASR, and software using visualization techniques. The article provides charts comparing visualization software and the kind of feedback used in different software titles. The review also compares the typical pedagogy, input, assessment techniques, and assumptions about learner autonomy for different uses of technology. This flyover view of commercial CAPT capabilities will be helpful to those who need to identify applications that may be of use for their own situations.

OTHER REFERENCES

Anderson-Hsieh, J. (1994). Interpreting visual feedback on suprasegmentals in computer assisted pronunciation instruction. *Calico Journal, 11*(4), 5–21.

Avery, P., & Ehrlich, S. (1992). *Teaching American English pronunciation*. Oxford: Oxford University Press.

Bradlow, A., Pisoni, D., Akahana-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101*, 2299–2310.

Breitkreutz, J., Derwing, T., & Rossiter, M. (2002). Pronunciation teaching practices in Canada. *TESL Canada Journal, 19*, 51–61.

Burgess, J., & Spencer, S. (2000). Phonology and pronunciation in integrated language teaching and teacher education, *System, 28*, 191–215.

Cauldwell, R. (2002). Streaming Speech: Listening and advanced pronunciation for advanced learners of English. *Talking Computers, Proceedings of the IATEFL Pronunciation and Computer Special Interest Groups*, pp. 18–22.

Chun, D. (1998). Signal analysis software for teaching discourse intonation. *Language Learning and Technology, 2*(1), 74–93.

Chun, D. (2007). Come ride the wave: But where is it taking us? *Calico Journal, 24*(2), 239–252.

Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English. *System, 27*, 49–64.

Coniam, D. (2002). Technology as an awareness raising tool for sensitising teachers to features of stress and rhythm in English. *Language Awareness, 11*(1), 30–42.

Cranen, B., Weltens, B., de Bot, K., & van Rossum, N. (1984). An aid in language teaching: The visualization of pitch. *System, 12*(1), 25–29.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Different aspects of expert pronunciation quality rating and their relation to scores produced by speech recognition algorithms. *Speech Communication, 30*, 109–119.

Dalton, C., & Seidlhofer, B. (1994). *Pronunciation*. Oxford: Oxford University Press.

Delmonte, R. (2000). SLIM prosodic tools for self-learning instruction. *Speech Communication, 30*, 145–166.

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397.

Derwing, T., & Rossiter, M. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning, 13*, 1–17.

Derwing, T., Munro, M., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly, 34*(3), 592–603.

Ehsani, F., & Knodt, E. (1998). Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning and Technology, 2*(1), 54–73.

Eskenazi, M. (1999). Using Automatic Speech Processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning and Technology, 2*(2), 62–76.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399–423.

Franco, H., Neumeyer, L., Digalakis, V., & Ronen, O. (2000). Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication, 30*, 121–130.

Gorozny, S., Rapp, S., & Kompe, R. (2004). Generating non-native pronunciation variants for lexicon adaptation. *Speech Communication, 42*, 109–123.

Hardison, D. (2004b). Contextualized computer-based L2 prosody training: Evaluating the effects of discourse context and video input. *Calico Journal, 22*(2), 175–190.

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System, 33*, 575–591.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

Kaltenboeck, G. (2002). Computer-based intonation teaching: Problems and potential. *Talking Computers, Proceedings of the IATEFL Pronunciation and Computer Special Interest Groups*, 11–17.

Kawai, G., & Hirose, K. (2000). Teaching the pronunciation of Japanese double-mora phonemes using speech recognition technology. *Speech Communication, 30*, 131–143.

Kim, I. S. (2006). Automatic speech recognition: Reliability and pedagogical implications for teaching pronunciation. *Educational Technology and Society, 9*(1), 322–344.

Ladd, D. R. (1980). *The structure of intonational meaning*. Bloomington: Indiana University Linguistics Club.

Lambacher, S. (1999). A CALL tool for improving second language acquisition of English consonants by Japanese learners. *Computer Assisted Language Learning, 12*(2), 137–156.

Levis, J. (2005). changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369–378.

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System, 32*(4), 505–524.

MacDonald, S. (2002). Pronunciation–views and practices of reluctant teachers. *Prospect, 17*(3). Retrieved April 10, 2007, from http://nceltr.edu.au/prospect/17/pros17_3smac.asp

Mich, O., Neri, A., & Giuliani, D. (2006). The effectiveness of a computer assisted pronunciation training system for young foreign language learners. *Proceedings of CALL 2006* (pp. 135–143). Antwerp, Belgium. Retrieved July 7, 2007, from http://lands.let.kun.nl/literature/neri.2006.4.pdf

Molholt, G. (1988). Computer-assisted instruction in pronunciation for Chinese speakers of American English. *TESOL Quarterly, 22*(1), 91–111.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL Quarterly, 25*(3), 481–520.

Murphy, J. (1997). Phonology courses offered by MATESOL programs in the United States. *TESOL Quarterly, 31*(4), 741–764.

Munro, M., & Derwing, T. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*(Supp.1), 285–310.

Munro, M., & Derwing, T. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*, 520–531.

Neri, A., Cucchiarini, C., & Strik, H. (2002a). Feedback in computer assisted pronunciation training: Technology push or demand pull? *Proceedings of International Conference on Spoken Language Processing 2002* (pp. 1209–1212). Denver, CO. Retrieved July 7, 2007, from http://lands.let.kun.nl/literature/neri.2002.2.pdf

Neri, A., Cucchiarini, C., & Strik, H. (2002b). Feedback in computer assisted pronunciation training: When technology meets pedagogy. *Proceedings of "CALL professionals and the future of CALL research" Conference* (pp. 179–188). Antwerp, Belgium. Retrieved July 7, 2007, from http://lands.let.kun.nl/literature/neri.2002.1.pdf

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning, 15*(5), 441–467.

Neri, A., Cucchiarini, C., & Strik, H. (2006a). Selecting segmental errors in L2 Dutch for optimal pronunciation training. *International Review of Applied Linguistics, 44*, 357–404.

Neri, A., Cucchiarini, C., & Strik, H. (2006b). ASR-based corrective feedback on pronunciation: Does it really work? *Proceedings of International Conference on Spoken Language Processing 2006* (pp. 1982–1985). Pittsburgh, PA. Retrieved July 7, 2007, from http://lands.let.kun.nl/literature/neri.2006.2.pdf

Oh, Y. R., Yoon, J. S., & Kim, H. K. (2007). Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. *Speech Communication, 49*, 59–70.

Pennington, M. (1999). Computer-aided pronunciation pedagogy: Promise, limitations, directions. *Computer Assisted Language Learning, 12*(5), 427–440.

Pogue, D. (2004, December 2). *Speaking naturally, anew*. Retrieved July 7, 2007, from http://www.nytimes.com/2004/12/02/technology/circuits/02POGUE-EMAIL.html?ex=1259730000&en=339f1ddcdf5fab69&ei=5088&partner=rssnyt

Protea Textware. (2007). Retrieved July 7, 2007, from http://www.proteatextware.com.au/

Rypa, M., & Price, P. (1999). VILTS: A tale of two technologies. *Calico Journal, 16*(3), 385–404.

ScanSoft. (2005). Dragon Naturally Speaking (8th ed.) [Computer software] Nuance Communications, Inc., www.nuance.com.

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication, 49*, 336–347.

Seferoğlu, G. (2005). Improving students' pronunciation through accent reduction software. *British Journal of Educational Technology, 36*(2), 303–316.

Spaai, G., & Hermes, D. (1993). A visual display for the teaching of intonation. *Calico Journal, 10*(3), 19–30.

Swan, M., & Smith, B. (2001). *Learner English* (2nd ed.). Cambridge: Cambridge University Press.

Truong, K., Neri, A., de Wet, F., Cucchiarini, C., & Strik, H. 2005. Automatic detection of frequent pronunciation errors made by L2 learners. *Proceedings of InterSpeech* (pp. 1345–1348). Lisbon, Portugal.

Van Compernolle, D. (2001). Recognizing speech of goats, wolves, sheep and . . . non-natives. *Speech Communication, 35*, 71–79.

Wang, X., & Munro, M. (2004). Computer-based training for learning English vowel contrasts. *System, 32*, 539–552.

Witt, S., & Young, S. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication, 30*, 95–108.