Viewpoint

RELC

RELC Journal
2021, Vol. 52(1) 189–205
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0033688220977406
journals.sagepub.com/home/rel
SAGE

Computer-Assisted Pronunciation Training (CAPT): Current Issues and Future Directions

Pamela M Rogerson-Revell
University of Leicester, UK

Abstract

This viewpoint essay considers the current status of computer-assisted pronunciation training (CAPT) before examining some of the current issues and future directions in the field. The underlying premise is the pedagogic potential of CAPT systems and resources for teaching and learning, and the need for greater synergy between technological design and functionality on the one hand, and pedagogic purpose on the other. Some of the key issues examined include providing accurate and individualised automated feedback for pronunciation, for both learning and assessment, and evaluating the effectiveness of CAPT tools and systems. When considering future directions, the discussion focuses on what aspects of pedagogy are likely to be at the forefront of developments, including ubiquitous learning; intelligent tutoring and authentic interaction; and goal-oriented, task-based learning.

Keywords

Pronunciation, technology, computer-assisted pronunciation teaching, future directions, automatic speech recognition, ubiquitous learning

Introduction

It is many years now since Stephen Bax applied the concept of ‘normalisation’ in relation to the seamless integration of technology into learning and teaching. Bax (2003: 23) states that normalisation will have been achieved:

Corresponding author:
Pamela M Rogerson-Revell, School of Education, University of Leicester, LE17RH, UK.
Email: pmrr1@le.ac.uk

> when computers . . . are used every day by language students and teachers as an integral part of every lesson, like a pen or a book . . . without fear or inhibition, and equally without an exaggerated respect for what they can do. They will not be the centre of any lesson, but they will play a part in almost all. They will be completely integrated into all other aspects of classroom life, alongside coursebooks, teachers and notepads. They will go almost unnoticed.

Bax was referring to language learning generally rather than pronunciation specifically, but it is interesting to consider how far we have come towards the goal of integrating technology in relation to pronunciation teaching and learning.

It is fair to say that for many of us, digital technologies such as smartphones, computers, televisions, and tablets are ubiquitous in our everyday lives. For quite a few teachers and students, but by no means all, they are also commonly used in language classrooms or for self-study. The field of computer-assisted language learning (CALL) has developed massively in recent decades, and interest specifically in computer-assisted pronunciation training (CAPT) has grown similarly, with a recent proliferation of web-based and mobile apps and resources. Some of these are becoming increasingly technologically sophisticated, in some cases incorporating technologies such as automated speech recognition (ASR) and artificial intelligence (AI), further enabling opportunities for language production and individualised feedback. The increasing attractiveness and availability of such tools does not of course ensure their pedagogic value or effectiveness, and, in fact, on closer inspection many CAPT resources appear to be technology-driven rather than pedagogy-led, lacking the solid pedagogic underpinnings required to ensure effective language learning.

Nevertheless, the affordances of technology for pronunciation teaching and learning are undeniable and have been well documented elsewhere (see, e.g., Rogerson-Revell, 2011; Fouz-González, 2015; Levis, 2007; Pennington and Rogerson-Revell, 2019). In essence, CAPT resources have the potential to provide an individualised, stress-free, self-paced learning environment with limitless access to a wide range of multimodal material as well as opportunities for immediate, customised feedback.

CAPT can add value to traditional learning methods by maximising opportunities for exposure to a broader variety of spoken language, including different first language (L1) and second language (L2) accents and different speech genres and styles. Related to this is the potential to incorporate a variety of audiovisual input using an increasing range of delivery platforms, from educational websites (e.g. *British Council Learn English*, https://learnenglish.britishcouncil.org/) to mobile apps (e.g. *Duolingo*, https://www.duolingo.com/; *Mondly,* https://www.mondly.com/) and social media platforms such as Twitter, Facebook, and WhatsApp. One of the main affordances of such input is the ability, for example, to incorporate visual animations to help learners see and hear how sounds are articulated in real time. Such resources can also increase opportunities for language production, or output, from controlled pronunciation practice, such as the repetition of sounds or structured responses, to authentic interactions in meaningful contexts, such as when using smartphone apps or social media platforms. As well as allowing the learner to practise speaking, one of the main benefits of CAPT is being able to receive automated, immediate feedback on their pronunciation. Furthermore, developments in ASR mean that feedback can also be customised and targeted to the individual.

One of the main affordances of digital technologies, in general, is their capacity to motivate and engage learners, particularly the young and/or digitally literate. At the same time, it is important that tools and resources are technically intuitive and robust, so as not to exclude less experienced users. Equally, the novelty value of the 'wow' factor can soon wear off if not supported by solid pedagogic foundations. It is essential for teachers, researchers, and developers to consider the affordances of CAPT resources when evaluating their usefulness, in order to understand what elements of technology can add value or enhance pronunciation teaching, learning, or assessment. In the following sections, I will discuss further the pedagogic potential of CAPT resources alongside some of the current issues and future directions in the field.

Current Issues

Pedagogy Versus Technology

Despite growing interest in pronunciation from researchers and teachers, there is still wide variation in how much priority is given to pronunciation in classrooms, curriculum, exams, and textbooks (Henderson et al., 2012). Despite this, many students understand how important pronunciation is for successful communication and are, therefore, keen to benefit from CAPT apps and systems. Many of these are targeted and used by learners as self-study resources, with little in the way of quality control to evaluate how rigorous or effective they are in terms of pronunciation learning. Technological novelty tends to take centre stage and may temporarily disguise lack of pedagogic rigour but is unlikely to maintain motivation in the long term.

The tension between pedagogy and technology is a key issue, in both CALL and CAPT, and has been well documented (Rogerson-Revell, 2011; Levis, 2018; Pennington, 1999; Pennington and Rogerson-Revell, 2019). Pennington and Rogerson-Revell (2019), for instance, have called for greater collaboration between pedagogic and technical experts when designing products:

> Such collaboration could result in products in which functionalities are selected on pedagogic value as well as technical capabilities. Collaborations could also help ensure the quality of resources in terms of the validity and accuracy of the learning content and the reliability of feedback. (2019: 273)

As Pennington and Rogerson-Revell point out, 'one of the difficulties is that there is no obvious fit between language learning pedagogies and the affordances of digital technologies' (2019: 238). As a result, many CAPT resources are less innovative pedagogically than one would expect. Indeed, in some cases, as technology progresses, pedagogy appears to regress, returning to audiolingual approaches of repetition, mimicry, and drilling. While such methods still have their place, they are not sufficient to help develop communicative or phonological competence in a language.

Traditionally, CAPT resources tend to be quite limited in terms of pedagogic structure and content, with little evidence of an underlying phonological syllabus or clear learning aims or outcomes. Learning units typically focus on phonemes and employ a relatively narrow range of activity types, often involving minimal pair discrimination practice and giving generalised feedback, such as ‘yes/no’ or ‘correct/incorrect’. However, there are examples of more sophisticated, pedagogically informed resources, often those which have been developed through technological/academic collaborations, such as Cauldwell’s (2012) *Cool Speech* app, which focuses on the features of fluent natural speech and is based on extensive academic research, and the *Sounds of Speech* app (https://soundsofspeech.uiowa.edu/), developed by the University of Iowa.

A common criticism of CAPT tools is that many of them adopt a ‘one size fits all’ approach, providing generalised content and feedback, rather than individualised support for learners (Derwing and Munro, 2015; Levis, 2018). However, technological advances are increasingly enabling such tools to help diagnose individual student pronunciation needs accurately and quickly and to provide more customised, targeted learning materials. For example, with some commercial apps such as *ELSA* (https://elsaspeak.com/en/) and *Pronunciation Power* (http://englishlearning.com), learners can take an initial diagnostic test and, based on the results, be directed to the relevant pronunciation units. Another common concern is over the pedagogic accuracy of content and technical reliability of tools. It is not unusual to find descriptions of phonological features or statements of rules which are either inaccurate or simplistic. For instance, a unit on intonation in a mobile app at one point states that ‘when asking questions in English, our voice goes up in pitch’ and, at another, ‘in English, we lower the pitch of our voice at the end of questions’, with no contextual information or clarification to support these statements. At another point, in a simple exchange – ‘How are you Dylan?’ ‘Great thanks and you?’ – the user is instructed to ‘select the part of the sentence that has rising tone’ in the response ‘Great thanks and you?’. However, the app does not allow the user to select ‘you’.

Such instances can both frustrate and misinform learners: technology needs to work, and content needs to be pedagogically accurate as well as relevant. Again, this relates to the gap between pronunciation experts and those who develop CAPT resources. Two examples can illustrate this. While pronunciation researchers and many teachers are now conversant with the ‘intelligibility’ versus ‘nativism’ debate (Levis, 2018) and many accept that intelligibility is generally a much more achievable goal for many learners, many CAPT resources still focus on ‘native-like pronunciation’ as a learner goal and some promote the opportunity to have pronunciation corrected by a native speaker as a key feature of the training materials. Similarly, while many teachers and researchers now question the need for most learners to acquire all of the phonological features of a language in order to be intelligible, most CAPT systems do not prioritise pronunciation features on this basis, or on other important criteria such as error frequency or severity, or learner goals.

Feedback

Accurate and timely feedback is essential to help learners notice discrepancies between their production or output and the target L2 model. This is particularly the case for pronunciation since studies have shown that learners are often unaware of L1 interference and may find it difficult to perceive differences between their interlanguage production and the L2 pronunciation target (Ehsani and Knodt, 1998; Flege, 1995).

Many apps and programmes give simple but immediate feedback on a learner's perception of pronunciation features, including individual sounds (e.g. *Sounds*, https://www.macmillaneducationapps.com/soundspron/) word stress (e.g. *Pronunciation Power*) and, to some extent, across an extended phrase or utterance (e.g. *Duolingo*). However, as Levis (2007) explains, '[t]echnologically, CAPT systems often suffer from difficulties in giving learners adequate, accurate feedback and an inability to provide accurate and automatic diagnosis of pronunciation errors' (2007: 185).

Some CAPT systems provide immediate feedback on speaker input through various types of visual displays – for instance, in the form of spectrograms or wave forms of the speech sample, often compared with a model display from a native speaker. Research suggests that such visual displays can be of value; for example, Ramírez-Verdugo's (2006) study of Spanish learners of English found that after instruction using visual displays of prosody the learners employed a wider variety of tones and their intonation was judged as more native-like in form and meaning. Similarly, various studies have supported the use of spectrograms to increase learners' awareness of segmental errors (Akahane-Yamada et al., 1997; Molholt and Hwu, 2008; Olson, 2014). However, the use of such visual displays raises some issues. One is they assume that to produce 'correct' pronunciation of an utterance, the wave form or spectrogram of both speakers needs to be identical, when in fact the same utterance can be pronounced well by both speakers but the wave form or spectrogram for each speaker might be quite different. Another issue is that such displays present raw data, which requires some degree of expertise or training to be interpreted and, while they may appear impressive, are not easily understood by a student without the guidance of a teacher. Finally, since such displays can tell the average learner little about their errors or their causes, the learner is often left to make random attempts to correct their pronunciation, which can ultimately lead to frustration and possibly further fossilisation of errors.

Other CAPT systems use automatic speech recognition (ASR) to give more implicit and realistic feedback. ASR works by comparing speech input from one speaker, usually with a native speaker model, generated from a database containing a large number of samples of native speaker speech. ASR technology offers great potential in terms of providing immediate, individualised feedback. It has improved considerably in recent years, with programmes becoming much better at native speaker voice recognition, although the level of accuracy for non-native speech is much lower. Levis (2007: 192), for instance, points out that *Dragon Naturally Speaking* (https://www.nuance.com/dragon.html), an ASR-based word-processing programme, is '95% or more accurate for native-speaking English users' but accuracy levels drop to near 70% for L2 speakers of English with advanced level proficiency, but accented speech.

One example of a CAPT programme which uses ASR and has been trained to recognise non-native speech, is the *Tell Me More* language learning software by Rosetta Stone (https://www.rosettastone.com/). The software provides a sequence of interactive dialogues where the learner has to choose the correct response from three utterances which are phonetically different. If the learner's recorded response is a good match to the stored pronunciation model, it is recognised by the computer and the dialogue continues; if not, the learner has to repeat the response. The software's voice recognition feature, SETS (Spoken Error Tracking System), provides a global pronunciation score and highlights words within a sentence which are incorrectly pronounced. However, reviews of *Tell Me More* suggest that the software has difficulties in voice recognition that limit its usability (see, e.g., Effective Language Learning reviews, 2013).

It is not uncommon for CAPT apps to give erroneous feedback, either not detecting learner errors or falsely labelling acceptable pronunciation as errors, both of which can be frustrating and demotivating for learners. This issue is compounded by the fact that the computer often does not give sufficient guidance to correct individual errors, so that while feedback might point out that a phrase or utterance is unintelligible, or even highlight a specific word, little help is given to the learner regarding how to correct the error. One possible way of dealing with this is by incorporating audiovisual feedback, for instance, using talking heads (see later section) to illustrate how a sound is articulated, so that a learner can use the visual model to correct their pronunciation error.

Technology has obvious potential to provide immediate, customised feedback on both a learner's perception and production of the target language pronunciation. Well-designed ASR feedback has been found to improve learners' phoneme production, for example, of the /x/ sound in Dutch (Cucchiarini et al., 2009). Similarly, in the *DISCO* system, ASR technology was used to detect errors in syntax, vocabulary, and pronunciation for L2 learners of Dutch and found to be highly accurate (Strik et al., 2012). However, at present, ASR seems better at giving precise feedback on segmental errors than suprasegmental features (Levis, 2007). Also, most off-the-shelf ASR software is unable to distinguish which accent features affect intelligibility and which do not; it is equally unable to adjust to learner speech with increased exposure, as a human does (although this may be possible with AI and increasing sophistication of databases underpinning programmes). At present, these shortcomings can lead to false-negative error detection and feedback, and consequent frustration and demotivation for learners.

Despite advances in CAPT and ASR software, it can be seen that issues remain with the use of technology for error detection and feedback. A more nuanced approach is needed for the analysis and correction of pronunciation errors in L2 learner speech and, as explained in Rogerson-Revell (2011) and Pennington and Rogerson-Revell (2019), more specific criteria are needed to prioritise remedial work.

Automated Assessment

The use of computer-based language assessment is increasingly common, either diagnostically within language learning programmes or as standalone assessment tools. The potential benefits are obvious in terms of eliminating human bias, fluctuations in attention, and error. For pronunciation assessment, technology has long been used to assess perception and discrimination of sounds and comprehensibility of speech. More recently, the incorporation of ASR has enabled the automated assessment of pronunciation production, providing a holistic evaluation of pronunciation quality, usually in the form of a numerical score and based on deviation from a standard speech model. However, the limitations in the accuracy of ASR noted earlier mean that the use of such computerised assessment of pronunciation carries considerable risk, particularly for high-stakes tests where an individual's educational or professional prospects depend on the outcome.

AI and automated scoring are being used increasingly in international and national tests, including tests used for immigration purposes. For example, Australia's immigration department uses the Pearson Test of English (PTE) Academic (https://pearsonpte.com/) as one of five tests. The PTE tests speaking ability using voice recognition technology and computer scoring of test-takers' audio recordings. However, L1 English speakers and highly proficient L2 English speakers have failed the oral fluency section of the English test, and in some cases it appears that L1 speakers achieve much higher scores if they speak unnaturally slowly and carefully. One English language teacher, quoted in *The Guardian* newspaper, explains that:

> .. I encourage students to exaggerate intonation in an over-the-top way and that means less making sense with grammar and vocabulary and instead focusing more on what computers are good at, which is measuring musical elements like pitch, volume and speed. (Davey, 2017)

Criticisms have been raised with other testing services using automated testing of speaking and pronunciation such as the British Council and Cambridge Assessment, suggesting that, although this is a promising area, there are still issues of reliability and validity with automated scoring and use of AI.

Some commercial tests, such as Pearson's *Versant* (https://www.pearson.com/english/versant.html), are used regularly for high-stakes purposes such as testing the language proficiency levels of pilots and air traffic controllers. The test includes assessments of fluency and pronunciation as part of overall oral proficiency. Pearson reports that the speaking test has been extensively trialled and validated in relation to human raters of oral proficiency (Balogh et al., 2011). The test is based on short samples of speech and highly predictable speaking tasks, and the question remains whether testing would be equally valid for evaluating extended stretches of natural speech.

One of the main issues in automated testing of pronunciation is being able to correlate automated, ASR scoring with human judgements of pronunciation proficiency. Various studies have shown only limited correlation between human and machine evaluations, particularly at the segmental level (Derwing et al., 2000; Kim, 2006). However, there does seem to be closer correlation between human and machine scoring of speech rate and syllable duration (Cucchiarini et al., 2000).

Another challenge for computer-based pronunciation assessment is defining the appropriate criteria with which to evaluate pronunciation proficiency. Many CAPT tools assess pronunciation using descriptive adverbs which have been proven to be ineffective and vague (Alderson, 1991; North, 2000), such as 'poor pronunciation' or 'non-native-like intonation'. Also, most systems evaluate proficiency on the basis of accuracy, in relation to a native speaker model, without any consideration of other criteria such as the frequency, persistence, and salience of errors underpinning intelligibility and fluency. A fundamental issue is the fact that machines do not hear like humans and cannot process the complexities and nuances of natural speech as well. As Scharenborg explains, humans are 'far better at dealing with accents, noisy environments, differences in speaking style, speaking rate, etc.' (2007: 344). Nevertheless, machine-based assessment is now being used in an increasing range of language tests and, as Levis (2007) suggests, it may be good enough at providing global pronunciation scores for some purposes.

Effectiveness

A key concern of any pronunciation learning and teaching approach, whether computer-based or otherwise, is whether it is effective. There is some evidence that CAPT can be effective for pronunciation learning, especially for perceptual training and providing visual feedback. Various studies have also demonstrated the value of technology-based training on different aspects of pronunciation, such as phonemes (Neri et al., 2006; Wang and Munro, 2004); rhythm and stress (Coniam, 2002); intonation (Cauldwell, 2012; Hardison, 2004; Kaltenboeck, 2002; Levis and Pickering, 2004); and speech rate and fluency (Hincks, 2005). Learning L2 pronunciation is a complex process and there are many factors that influence a learner's success, such as L1 transfer, age and other individual factors, and educational context. As well as these factors, other components which are key to effective pronunciation learning, as mentioned earlier, are sufficient and relevant input, output, and feedback.

Computer-based training can capitalise on the affordances of digital technologies to provide a wide variety of spoken input, using a range of multimedia. Exposure to variable input from multiple speakers appears to be particularly effective for acquiring L2 phonemic categories (Logan et al., 1991). An example of this is Thomson's (2012) *English Accent Coach* based on an approach referred to as high-variability perceptual/phonetic training, as first developed by Logan et al. (1991), whereby, instead of focusing on a single model, learners listen to multiple speaker models to develop awareness of, and tolerance to, phoneme variation.

Studies consistently confirm the benefit of multimodal input for pronunciation learning and, in particular, the use of visual displays, such as spectrograms and two-dimensional (2D) and three-dimensional (3D) computer animations of the lips and oral cavity (Elliot, 1995; Grant and Greenberg, 2001). The use of visual displays has been shown to be effective not only for input but also for students to monitor their pronunciation output and to enhance feedback. There is some evidence that such displays can help learners distinguish their own phoneme production – for instance, L2 learners of Spanish found that spectrograms helped them distinguish their own production of stop consonants from the target forms (Olson, 2014). At the suprasegmental level, various studies have shown that visual representations can help learners with perception and production of intonation (e.g. De Bot and Mailfert 1982; Hardison, 2004; Levis and Pickering, 2004). It has also been shown that simultaneously hearing and seeing speech articulations can improve both the perception and production of sounds (Massaro, 1987), suggesting that using technologies that enable the visualisation of mouth and facial movements can be helpful. It appears that the use of such displays may be most beneficial to demonstrate external articulations – that is, of the lips, rather than internal movements of the tongue and other internal articulators. For instance, Badin et al. (2010) showed the positive effects of training a range of vowels and consonants using visual displays but found a frontal view of the face was perceived better than a cutaway view of the head.

Personalized, immediate feedback appears to be particularly effective for language learning, and the use of immediate and personalised ASR feedback has been shown to have a positive impact on pronunciation learning. Cucchiarini et al. (2009) developed a CAPT system using ASR for L2 learners of Dutch and found a significant reduction in the number of mispronunciations in the experimental group using ASR feedback. Similarly, Mayfield-Tomokiyo et al. (2000) report a substantial reduction in error rate for the teaching of /θ/ and /ð/ to learners with different L1s, although in this case the ASR feedback was accompanied by visual feedback on articulations. As discussed earlier, providing effective feedback is still a challenging area for CAPT, especially feedback on learner production using ASR. To be optimally effective, systems not only need to provide accurate feedback specific to an individual learner, but also should be able to prioritise errors in relation to various criteria including frequency, salience, intelligibility, and fluency, rather than simply proximity to a native speaker model.

Future Directions

As described earlier, the key issue in CAPT, and indeed in CALL more broadly, is the tension between technology and pedagogy. When considering future directions, it would be easy to speculate on where technology is likely to lead the field, but a more important concern is what aspects of pedagogy will be at the forefront of developments.

Ubiquitous Learning

At the beginning of this paper, I described how many digital technologies have become ubiquitous: an essential but taken-for-granted part of our everyday lives. Mobile devices such as smartphones, tablets, and laptops offer learners the possibility to study anytime, anywhere, and at their own convenience: an experience referred to as ‘ubiquitous’ learning (Li et al., 2005; Yang, 2006). Smartphones are a prominent example of ubiquitous learning devices, having characteristics that lend themselves particularly well to language learning, including portability, social interactivity, context-sensitivity, connectivity, and individualisation (Sung et al., 2015). Mobile-assisted language learning is now a well-established field and one which has had an enormous impact on L2 teaching and learning (Pachler et al., 2010). The availability and functionality of mobile devices is leading to an increase in informal, self-directed, out-of-class mobile learning, where smartphones give learners the opportunity to create ‘impromptu sites of learning’ (Bachmair and Pachler, 2014; Kukulska-Hulme et al., 2017). A recent meta-analysis demonstrated that mobile learning was more effective in informal instructional settings than in formal settings (Sung et al., 2016). Such studies suggest that, given the relatively small amount of time spent on pronunciation teaching in the classroom, the increase in technologies that enable ubiquitous, out-of-class learning are to be welcomed.

A vast number of mobile apps for language learning are now available, including some aimed specifically at pronunciation training or incorporating a focus on pronunciation. In a survey of pronunciation apps, Foote and Smith (2013) found that most concentrated on individual phonemes, used pedagogically dubious methods, and did not necessarily have an effect on intelligibility. However, there are examples of good practice and positive outcomes. For instance, Foote and McDonough (2017) found that regular individual practice using shadowing with mobile technology improved participants' comprehensibility and fluency of speech (though not their accentedness). Fouz-González's (2020) study of 52 Spanish leaners of English found that the use of the *English File Pronunciation* app (https://elt.oup.com/student/englishfile/advanced3/pronunciation) helped learners improve both perception and production of some phonemes.

Intelligent Tutoring and Authentic Interaction

Technological advances are gradually enabling CAPT systems to move beyond giving simplistic, generic support and feedback towards providing more customised, intelligent tutoring and more opportunities for meaningful, authentic interaction. The potential of ASR to provide instant, personalised feedback is increasingly being promoted as a key feature of many apps, such as *Pronunciation Power* and *Say It* (https://elt.oup.com/catalogue/items/global/pronunciation), which now incorporate a voice recognition function to give individualised feedback on a user's speech recording. Further progress in ASR technology, including better performance in noisy environments, together with facial recognition technology to capture lips movement and facial expression, could further enhance the capabilities of mobile technology for pronunciation learning. Smartphone software developers are also increasingly integrating AI in apps to add the functionality of so-called 'intelligent' language tutors. Apps such as *Duolingo, ELSA Speak* (https://elsaspeak.com/en/), and *Busuu* (https://www.busuu.com/en/mobile) use voice recognition and AI to enable controlled interactions and simulate structured conversations with the user.

While we are still some way off from being able to have natural, spontaneous conversations with machines, rapid advances in speech technologies are making it more and more possible to have meaningful, authentic interactions with computers. Since the middle of the 20th century, computer programmes known as *chatbots* have been designed to simulate authentic interactions. Such programmes incorporate AI to appear to understand and take part in unscripted dialogues with a human by giving programmed responses to keywords and phrases. Recent advances in AI and computing power have enhanced the capability of such *bots*, including for language learning purposes, so that they respond to speech-based as well as text-based chats – for example, the 'conversational chatbot' in the *Mondly* app. Although somewhat limited at present, such apps could motivate the learner to practise pronunciation with an intelligent online tutor. Such tools represent an important step in CAPT towards the development of more sophisticated spoken dialogue systems, which combine the technologies of speech recognition, speech synthesis, and natural language processing to allow an individual to converse with a computer in spontaneous dialogues. Software developers and researchers have been exploring this field for some time to create *embodied conversational agents* (ECAs), or *talking heads*, which can function as language tutors and conversational partners.

Research clearly shows that visual information from the face aids intelligibility and communication (Benoît et al., 1994; Jesse et al., 2000). ECAs were developed originally in the speech sciences to help individuals with communication difficulties – for instance, *Baldi*, a 3D talking head, has been used effectively with deaf and autistic children (Bosseler and Massaro, 2003; Massaro and Light, 2004). Both 2D and 3D talking heads have also been used for language learning, including pronunciation training. For example, Baldi has been used as a language tutor (Ouni et al., 2005), and Alsabaan and Ramsay (2014) developed a talking head to help L2 learners of Arabic improve their pronunciation. Wik and Hjalmarsson (2009) developed two animated talking heads, *Ville* and *DEAL*, with different roles and functionality, for L2 learners of Swedish. *Ville* is described as '"a virtual teacher" whose role is to guide, encourage, and give corrections on a student's pronunciation and language use' (Wik and Hjalmarsson, 2009: 1025), while *Deal* acts as a 'role-play dialogue system for conversation training' (Wik and Hjalmarsson, 2009: 1025). While there is still some way to go in their development, many see the use of talking heads as holding great promise for CAPT (Engwall, 2008; Fouz-González, 2015; Liu et al., 2007; Pennington and Rogerson-Revell, 2019: 252–254) enabling both intelligent, personalised feedback and authentic interactions in the form of a virtual tutor.

Taking embodied agents one step further leads to the use of robots for learning, including language learning. In the near future, personal robots may be the next big change in our everyday lives, including how we communicate. Having been used for decades in the automotive industry, they are now being used in a much wider range of applications, such as lawn mowers, autonomous vehicles, and even as care home assistants and hotel receptionists. Although their use in language learning is still relatively new, the field of *robot-assisted language learning* (RALL) emerged in the mid-2000s and has grown particularly in countries like Japan, Korea, and Taiwan, where automated L1 English-speaking teaching assistants have been used to help young learners with their pronunciation learning, often in pre-school or after-school programmes. Han (2012) explains the use of robots as an alternative pedagogical approach in a South Korean English as a foreign language context, where the robot can take the role of the native speaker and interact with the learners. According to Han:

> Among the various instructional models in language learning, we should consider RALL, employing currently emerging robot technology. This anthropomorphized version of existing mobile devices is autonomous, with features such as image recognition through camera, voice recognition through microphone, and interaction based on various sensors. (2012: 5)

Another interesting technological development which is starting to have an impact on language learning is the use of *virtual reality* (VR) systems, either VR environments or, more recently, VR headsets. VR environments such as *Second Life* (https://secondlife.com/) and *Active Worlds* (https://www.activeworlds.com/) have been around for some time and their use for language learning is well documented (Lin and Yan, 2015; Wang and Vásquez, 2012), although less so for pronunciation training specifically. The affordances of VR include enabling the learner to immerse themselves in a wide variety of simulated, real-life social contexts and activities, which are often game-based, and to assume a persona or avatar, which affords personal anonymity and, therefore, potentially reduces anxiety (Peterson, 2012). In the last few years, some app developers have introduced VR to mobile apps, such as *Mondly* VR (https://www.mondly.com/vr-for-daydream), where the user can take their avatar through various scenarios, such as in a taxi cab or in a restaurant, simulating a tourist experience in a foreign country. *Mondly*'s latest development includes VR headsets, which enable users to participate in multiplayer events, using an avatar to practice speaking the target language, with the aid of speech recognition and AI.

As yet, there is limited research into the effectiveness of RALL or VR for pronunciation training, but studies suggest that robots can increase motivation and enhance language learning (e.g. Movellan et al., 2009; Park et al., 2011). As with other areas of CAPT, if robots or VR tools represent the next technological paradigm shift, great consideration will have to be given as to how best to use and integrate such devices from the perspective of pronunciation learning and pedagogy.

Goal-Oriented, Meaningful, Task-Based Learning

As well as providing opportunities for authentic interaction, other key elements of successful language and pronunciation learning are motivation and engagement, and technologies that adopt a games-based approach can provide users with fun and challenging scenarios in which to practice meaningful, task-based language use. The affordances of online games and simulations for both CAPT and CALL are increasingly being recognised (Godwin-Jones, 2014; Golonka et al., 2014). As Pennington and Rogerson-Revell (2019: 256–257) point out, the use of avatars and role plays can have positive effects in depersonalizing interactions and so reducing anxiety while the elements of competition, problem solving, and reward for successful task completion can motivate users. The immersive environment of games can also be absorbing and provide extensive exposure to the target language. If a speech function is available, the user typically has to converse rapidly, in real and meaningful ways, to complete tasks, and moreover needs to understand and be understood by a range of players from different language backgrounds. All of these features can be of benefit for teaching and learning pronunciation.

To date, not many game-based programmes or apps have been developed specifically for language learning, let alone pronunciation (see Pennington and Rogerson-Revell, 2019: 255–259, for an overview) because of the technological complexity and development costs. However, many commercial *massively multiplayer online* games such as *World of Warcraft* (https://worldofwarcraft.com/en-us/) or *Star Wars* (e.g. https://www.swtor.com/) are available in multiple languages. As such, they provide extensive opportunities for ubiquitous pronunciation learning through exposure to authentic spoken communication.

Conclusion

The current coronavirus pandemic has shown how important digital technologies are in many aspects of our lives, including communication, entertainment, and education. There are many unanswerable questions about the future role of technology in language learning generally and pronunciation learning specifically. Will technological developments make teachers redundant, as virtual tutors become the norm? How will technology change learning behaviours, needs, and goals? For instance, will the concept of intelligibility need to broaden not only beyond intelligibility to native speakers but also to include intelligibility to ASR-based personal assistants such as Apple's *Siri* and Google's *Alexa*? Will a combination of speech synthesis, speech recognition, and AI make language learning unnecessary altogether? Will we be able to put on a VR headset that automatically translates and interprets another language for us?

At present, the affordances of technology for language learning in general and pronunciation in particular are undeniable. The infinite source of a wide range of speaker input in terms of accent, L1 background, and speech style is far beyond what is available in a normal classroom. Technology offers the learner limitless choice of what, when, and how to learn. Continuous technological advances are increasingly enhancing the scope of learner output, so that rather than simply recording individual sounds or words, learners can interact in meaningful dialogues or participate in real-world games which test their ability to produce intelligible speech. What still remains problematic is automated feedback. While this has greatly improved in recent years, there are still issues in terms of the level of detail and accuracy of feedback, whether at the segmental or the suprasegmental level. Feedback, whether automated or not, must be correct and reliable; and despite advances in ASR, there are still limitations in providing real-time, robust, easy to interpret, automated feedback.

The other remaining issue with CAPT lies not with the technology itself but with the design and development of CAPT systems. Educational technology is only as good as the humans behind it. This leads back to the underlying tension between technology and pedagogy and the need for greater collaboration between educational technologists, pronunciation experts, and teachers to ensure that the designs and functionalities of apps and software reflect learner needs and that technologies are used to optimise pedagogical effectiveness.

Digital technologies offer great potential for pronunciation training, as for other areas of language learning. The rapid, continuous evolution of technologies makes it hard to keep up with this changing field, let alone predict future developments. While we may not have fulfilled Bax's (2003) vision of the normalisation of technology in language learning and teaching, the role of ubiquitous learning seems to offer considerable promise for pronunciation learning. Despite some of the issues outlined here, technological advances should bring further benefits to pronunciation teaching and learning, especially if harnessed to the needs and priorities of learners and teachers.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Pamela M Rogerson-Revell https://orcid.org/0000-0002-9403-7992

References

Akahane-Yamada R, Adachi T, and Kawahara H (1997) Second language production training using spectrographic representations as feedback. *Journal of the Acoustical Society of Japan* 18: 341–343.

Alderson JC (1991) Bands and scores. In: Alderson JC, North B (eds) *Language Testing in the 1990s*. London: Macmillan, 71–86.

Alsabaan M, Ramsay A (2014) Diagnostic CALL tool for Arabic learners. In: *CALL design: Principles and practice: proceedings of the 2014 EUROCALL conference* (eds Jager S, Bradley L, and Meima EJ, et al.), Groningen, the Netherlands, September 2014, pp.6–11. Research-publishing.net.

Bachmair B, Pachler N (2014) A cultural ecological frame for mobility and learning. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung* 24: 53–74.

Badin P, Tarabalka Y, Elisei F, et al. (2010) Can you 'read tongue movements'? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication* 52: 493–503.

Balogh J, Bernstein J, Suzuki M, et al. (2011) Automatically scored spoken language tests for air traffic controllers and pilots. *VERSANT White Paper, Pearson Education*. Available at: https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/versant-test/Paper-Automatically-Scored-Spoken-Language-Tests-for-ATCs-and-Pilots.pdf (accessed 1 September 2020).

Bax S (2003) CALL—Past, present and future. *System* 31: 13–28.

Benoît C, Mohammadi T, and Kandel S (1994) Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research* 37: 1195–1203.

Bosseler A, Massaro DW (2003) Development and evaluation of a computer-animated tutor for vocabulary and language learning for children with autism. *Journal of Autism and Developmental Disorders* 33(6): 653–672.

Cauldwell R (2012) Cool Speech app. Available at: http://www.speechinaction.org/cool-speech-2 (accessed 12 July 2020).

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.

Cucchiarini C, Neri A, and Strik H (2009) Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication* 51(10): 853–863.

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

Davey M (2017) Outsmarting the computer: The secret to passing Australia's English-proficiency test. *The Guardian*, 9 August. Available at: https://www.theguardian.com/australia-news/2017/aug/10/outsmarting-the-computer-the-secret-to-passing-australias-english-proficiency-test (accessed 12 July 2020).

De Bot K, Mailfert K (1982) The teaching of intonation: Fundamental research and classroom applications. *TESOL Quarterly* 16: 71–77.

Derwing TM, Munro MJ (2015) *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. Amsterdam: John Benjamins.

Derwing TM, Munro MJ, and Carbonaro MD (2000) Does popular speech recognition software work with ESL speech? *TESOL Quarterly* 34: 592–603.

Effective Language Learning (2013) Tell me more review. Available at: https://effectivelanguagelearning.com/language-course-reviews/tell-me-more-review/ (accessed 21 July 2020).

Ehsani F, Knodt E (1998) Speech technology in computer-aided learning: Strengths and limitations of a new CALL paradigm. *Language Learning and Technology* 2: 45–60. Available at: http://llt.msu.edu/vol2num1/article3/index.html (accessed 27 February 2002).

Elliot R (1995) Foreign language phonology: Field independence, attitude and the success of formal instruction in Spanish pronunciation. *Modern Language Journal* 79(4): 530–542.

Engwall O (2008) Can audio-visual instructions help learners improve their articulation? An ultrasound study of short term changes. In: *Proceedings of the 9th annual conference of the inter-*

national speech communication association, INTERSPEECH, Brisbane, Australia, 22–28 September 2008, pp.2631–2634. Brisbane: Interspeech.

Flege JE (1995) Second-language speech learning: Findings and problems. In: Strange W (ed) *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*. Timonium, MD: York Press, 233–273.

Foote JA, McDonough K (2017) Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation* 3(1): 34–56.

Foote JA, Smith G (2013) *Is there an app for that?* In: *Paper presented at the 5th pronunciation in second language learning and teaching conference*, Ames, Iowa, 19–21 September 2013.

Fouz-González J (2015) Trends and directions in computer assisted pronunciation training. In: Mompean J, Fouz-González J (eds) *Investigating English Pronunciation: Trends and Directions*. Basingstoke: Palgrave Macmillan, 314–342.

Fouz-González J (2020) Using apps for pronunciation training: An empirical evaluation of the English File Pronunciation app. *Language Learning & Technology* 24(1): 62–85.

Godwin-Jones R (2014) Games in language learning: Opportunities and challenges. *Language Learning & Technology* 18(2): 9–19.

Golonka EM, Bowles AR, Frank VM, et al. (2014) Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning* 27: 70–105.

Grant K, Greenberg S (2001) Speech intelligibility derived from asynchronous processing of auditory visual information. Paper presented at the Audio-visual Speech Processing Workshop 2001. Available at: http://www.icsi.berkeley.edu/ftp/pub/speech/papers/avsp01-av.pdf (accessed 12 July 2020).

Han J (2012) Emerging technologies: Robot-assisted language learning. *Language Learning & Technology* 16(3): 1–9.

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

Henderson A, Frost D, Tergujeff E, et al. (2012) The English pronunciation teaching in Europe survey: Selected results. *Research in Language* 10: 5–27.

Hincks R (2005) Measures and perceptions of liveliness in student oral presentation speech: A proposal for automatic feedback mechanism. *System* 33(4): 575–591.

Jesse A, Vrignaud N, Cohen M, et al. (2000) The processing of information from multiple sources in simultaneous interpreting. *Interpreting* 5: 95–115.

Kaltenboeck G (2002) Computer-based intonation teaching: Problems and potential. In: Talking computers, proceedings of the IATEFL pronunciation and computer special interest groups, pp.11–17. Farenham: IATEFL.

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

Kukulska-Hulme A, Gaved M, Jones A, et al. (2017) Mobile language learning experiences for migrants beyond the classroom. In: Beacco JC, Krumm HJ, , and Little D, et al. (eds) The Linguistic Integration of Adult Migrants: Some Lessons from Research. Berlin: De Gruyter, 219–224.

Levis J (2007) Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics* 27: 184–202.

Levis J (2018) *Intelligibility, Oral Communication, and the Teaching of Pronunciation*. Cambridge: Cambridge University Press.

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

Li L, Zheng Y, Ogata H and Yano Y (2005) Ubiquitous computing in learning: toward a conceptual framework of ubiquitous learning environment. *International Journal of Pervasive Computing and Communications* 1 (3): 207-216.

Lin TJ, Lan YJ (2015) Language learning in virtual reality environments: Past, present, and future, *Educational Technology & Society* 18(4): 486–497.

Liu Y, Massaro D, Chen T, et al. (2007) Using visual speech for training Chinese pronunciation: An in-vivo experiment. In: *Speech and Language Technology in Education* (*SLaTE)*, Farmington, PA, USA, 1-3 October 2007, pp.29–32. Farmington: SLaTE.

Logan JS, Lively SE, and Pisoni DB (1991) Training Japanese listeners to identify English /r/ and /l/ III: Long-term retention of new phonetic categories, *Journal of the Acoustical Society of America* 89: 874–886.

Massaro DW (1987) *Speech Perception by Ear and Eye: A Paradigm for Psychological Enquiry*. Hillsdale, NJ: Lawrence Erlbaum.

Massaro DW, Light J (2004) Using visible speech for training perception and production of speech for hard of hearing individuals. *Journal of Speech, Language, and Hearing Research* 47(2): 304–320.

Mayfield-Tomokiyo L, Wang L, and Eskenazi M (2000) An empirical study of the effectiveness of speech-recognition-based pronunciation tutoring. In: *Proceedings of the 6th international conference on speech and language processing*, Beijing, China, 16–20 October 2000, pp.677–680. Beijing: ICSLP.

Molholt G, Hwu F (2008) Visualization of speech patterns for language learning. In: Holland M, Fisher F (eds) *The Path of Speech Technologies in Computer Assisted Language Learning: From Research toward Practice*. London: Routledge, 91–122.

Movellan J, Eckhardt M, Virnes M, et al. (2009) Sociable robot improves toddler vocabulary skills. In: *Proceedings of the 4th ACM/IEEE international conference on human robot interaction*, La Jolla, CA, USA, 11–13 March 2009, pp.307–308. La Jolla: IEEE.

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

North B (2000) *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang.

Olson D (2014) Benefits of visual feedback on segmental production in the L2 classroom. *Language Learning and Technology* 18: 173–192.

Ouni S, Cohen MM, and Massaro DW (2005) Training Baldi to be multilingual: A case study for an Arabic Badr. *Speech Communication* 45(2): 115–137.

Pachler N, Bachmair B, and Cook J (2010) *Mobile Learning: Structures, Agency, Practices*. New York: Springer.

Park S, Han J, Kang B, et al. (2011) Teaching assistant robot, ROBOSEM, in English class and practical issues for its diffusion. In: *Proceedings of IEEE workshop on advanced robotics and its social impacts*, Menlo Park, California, USA, 2-4 October 2011, pp.8–12. California: IEEE. Available at: http://www.davidbutterworth.net/bibtex/pdf/service_robots/park2011a.pdf (accessed 21 December 2020).

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

Pennington MC, Rogerson-Revell P (2019) *English Pronunciation Teaching and Research: Contemporary Perspectives*. London: Palgrave Macmillan.

Peterson M (2012) EFL learner collaborative interaction in Second Life. *ReCALL* 24(1): 20–39.

Ramírez-Verdugo M (2006) A study of intonation awareness and learning in non-native speakers of English. *Language Awareness* 15: 141–159.

Rogerson-Revell P (2011) *A Study English Phonology and Pronunciation Teaching*. London: Bloomsbury.

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

Strik H, Colpaert J, Doremalen J, et al. (2012) The DISCO ASR-based CALL system: Practicing L2 oral skills and beyond. In: *Proceedings of the conference on international language resources and evaluation (LREC)*, Istanbul, Turkey, 21–27 May 2012, pp.2702–2708. Istanbul: LREC.

Sung YT, Chang K, and Liu TC (2016) The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education* 94: 252–275.

Sung YT, Chang KE, and Yang JM (2015) How effective are mobile devices for language learning? A meta-analysis. *Educational Research Review* 16: 68–84.

Thomson RI (2012) English Accent Coach: Not quite a fairy godmother for pronunciation instruction, but a step in the right direction *Contact* 38(1): 18–24.

Wang S, Vásquez C (2012) Web 2.0 and second language learning: What does the research tell us? *CALICO Journal* 29(3): 412–430.

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

Wik P, Hjalmarsson A (2009) Embodied conversational agents in computer assisted language learning. *Speech Communication* 51(10): 1024–1037. Available at: http://dx.doi.org/10.1016/j.specom.2009.05.006 (accessed 17 July 2020).

Yang SJ (2006) Context aware ubiquitous learning environments for peer to peer collaborative learning. *Educational Technology and Society* 9: 188–201.