COPiL CAMBRIDGE OCCASIONAL PAPERS *in* LINGUISTICS
Volume **11**, Article 5: 120–130, 2018 | ISSN 2050-5949

Conceptualizing an Ideal Computer-Assisted Pronunciation Training System: Educational Technology for L2 Phonetic Features*

Nick Ballou
University of Cambridge

Abstract Computer-assisted pronunciation training (CAPT) has been growing in popularity, computational robustness, and pedagogical relevance for the better part of two decades, but the features included for identifying pronunciation errors and giving relevant feedback to users vary widely and often are not grounded in research. Studies in second language speech acquisition and related fields have identified a number of effective strategies for pronunciation training, including perceptual training (high variability phonetic training), user voice matching, visual prosodic feedback, and gamification. However, despite this work and the relative technological simplicity of these features, many CAPT systems do not include them and do not provide justification for the features that are included. This paper is a first attempt to create a complete framework of relevant and effective phonetic feedback mechanisms in CAPT, with the goal of delineating the outline of an 'ideal' or standardised CAPT program. The framework is then used in a brief evaluation of two sample CAPT programs (*Rosetta Stone* and *NativeAccent English*) to demonstrate how it can be applied to these and other systems.

1 Introduction

In this paper, a framework for the evaluation of phonetic feedback in computer-assisted pronunciation training (CAPT) will be presented. CAPT, which refers to the use of computer technology to teach, correct, and improve pronunciation, offers a number of benefits to language learners. It can be used in a private, stress-free environment, allowing learners to choose their own pace of progression and schedule. By providing learners with unlimited patience from an individualised tutoring system, CAPT can build confidence and help erode the negative emotional and motivational factors like anxiety, self-consciousness, and boredom, which Krashen collectively calls the affective filter (Krashen 1985), that impede language learning. The affective filter is particularly relevant for speech; oral tasks have been shown to generate significantly higher levels of anxiety among learners than listening, reading, and writing (Young 1990). It can be useful not just for L2 learners,

* This paper is an abridged version of my long essay for the MPhil in Theoretical and Applied Linguistics at the University of Cambridge. I want to thank my supervisor, Dr Calbert Graham, for his astute comments and support and an anonymous examiner for their insightful feedback.

but also for people with hearing impairments or speech pathologies (Popovici & Buică-Belciu 2012).

The majority of modern CAPT programs use phonetic feedback, or the automatic recognition of subsegmental, segmental, and suprasegmental pronunciation errors with suggestions for correction, and it is this aspect that will be discussed here. Research, however, is lacking in this area; the inclusion or exclusion of a given feature is often presented in the context of researchers' or developers' intuitions, and not on empirical evidence of its effectiveness. To educe the key components of a maximally effective or 'ideal' system, it is therefore necessary to synthesise research from a number of fields including second language speech acquisition, computer-assisted language learning (CALL), and CAPT itself. In section 5, the framework will be applied to two prominent examples of CAPT software (*Rosetta Stone* and *NativeAccent*) in order to both illuminate ways in which those programs could improve their efficacy and demonstrate how these criteria can be used to evaluate other similar programs.

2 CAPT structure

In evaluating CAPT phonetic feedback systems, it is first important to understand the various ways in which those systems may vary pedagogically. Phonetic feedback can be categorised according to the following major factors (adapted from Chen & Li 2016, Hansen 2006, Engwall 2012): (i) level of detail, (ii) medium, (iii) input, and (iv) adaptation to the user's native language (L1).

Level of detail refers to the specificity of pronunciation error detection. Early phonetic feedback was limited to binary ('acceptable' or 'unacceptable') feedback at the utterance level, but newer CAPT software has made more ambitious attempts to give precise feedback on the learner's error. This is theoretically possible at the sub-segmental (place/manner of articulation, e.g.), segmental (phone), word, and suprasegmental (pitch, intonation, duration, etc.) levels. Subsegmental feedback, however, requires the highest level of precision from the automatic speech recognition (ASR) system and was not present in any of the CAPT programs reviewed for this paper.

Next, CAPT programs vary in the medium of feedback they give, which can be visual and/or auditory in form. Different aspects of pronunciation may be better suited to particular mediums; Chen & Li (2016), for example, hypothesise that visual feedback is especially well-suited for suprasegmentals. Other researchers have endeavoured to adapt feedback mediums to particular learning styles, though results on this topic have been mixed (Hsu 2016).

Input in CAPT is typically one or a combination of text-to-speech (TTS), native speaker recordings, and modified user speech, each with its own strengths and weaknesses. TTS modules are inexpensive and easily extendable to new utterances, but lack the direct imitability of native speakers and are prone to segmental and prosodic errors. Native speaker recordings require additional resources from the developer, but provide the best examples of well-formed utterances like those a learner will encounter in the outside world. Finally, some systems have mixed one

of those types of input with modified versions of the learner's own voice by using technology to change certain acoustic features of an utterance and make it more nativelike (Felps, Bortfeld & Gutierrez-Osuna 2009). User speech modification will be discussed further in sections 3 and 4.

Finally, CAPT systems vary in the extent to which they incorporate knowledge of the user's L1. By training error detection models on non-native speaker data from a particular L1 background, the system can better generalise about the types of mistakes a group of speakers may make (devoiced word-final consonants by German learners of English, for example). This added precision, however, comes at the cost of decreased portability-a system intended for a particular L1 may require new training data to be effective for users from a different linguistic background. Some researchers have proposed 'weak' versions of L1 adaptation that allow CAPT to make coarse-grained adjustments to a given L1 without the need for training data. One of these methods, *L1-L2map*, will be examined in section 4.

3 Second language speech acquisition

It has been noted many times over that only a fraction of late (post-pubescent) language learners will ever achieve nativelike pronunciation (Avery & Ehrlich 1992). A number of explanations have been proposed for this, ranging from neurological maturation and reduced sensorimotor plasticity (Scovel 1988, Penfield & Roberts 1959), to insufficient motivation, and to establishment of incorrect habits in the early stages of learning (Richards, Platt & Platt 1992). Though notions of a sensitive period in L2 speech remain controversial, a well-supported line of research has shown that certain features of L2 pronunciation are not acquired naturalistically, but can be improved with instruction. Most teachers using modern methods do this implicitly with recasts and the incorporation of multimedia teaching material. Others also teach pronunciation explicitly with corrections and methods like CAPT.

In response to research on ultimate L2 attainment and the changing social dynamics of immigration and globalisation, the primary goals in language education for pronunciation have shifted to intelligibility and comprehensibility, rather than nativelike pronunciation (Munro & Derwing 2011, Nair-Venugopal 2003). As early as 1949, researchers began to espouse the view that 'most language learners need no more than a comfortably intelligible pronunciation' (Abercrombie 1949: p. 120). This has resulted in focus on the pronunciation errors that are most detrimental to successful communication at lower L2 proficiencies (A1-B1 CEFR); advanced learners may attempt to correct more subtle deviations from native speech later in the language learning process.

One of the first prominent models to predict areas of difficulty for L2 learners came from the Contrastive Analysis Hypothesis, which claimed that 'those elements that are similar to [the learner's] native language will be simple for him, and those elements that are different will be difficult' (Lado 1957: p. 2). Though it is now quite clear that L1 transfer is not the only factor affecting L2 pronunciation, L1 features can nonetheless provide a starting point that may help inform the development of exercises and teaching material (Avery & Ehrlich 1992).

Further work has shed light on the more precise interactions between L1 and L2 phonetics and phonology. The Speech Learning Model (SLM) (Flege 1995) contends that 'the processes and mechanisms that children use when establishing the sound system of the L1 remain intact across the life span, and remain accessible for use in L2 learning', but that those processes are influenced by the common phonological space shared by the L1 and L2, and evolve via category assimilation and category dissimilation (Flege 2007: p. 385). SLM predicts that learners will have the greatest difficulty with L2 sounds that are similar but not identical to sounds in their L1, followed by sounds that are unlike any L1 categories. Flege (1987), for example, found evidence for categorical assimilation among American women living in France (average length of residency: 10 years). The study found that subjects produced French voiceless stops with VOTs longer than are typical of French monolinguals, but shorter than typical of English monolinguals.

A second important tenet of SLM is that many L2 production errors have a perceptual basis, and that perception leads production for certain (if not all) unfamiliar L2 sounds. Evidence for this has been given by studies like Bradlow, Pisoni, Akahane-Yamada & Tohkura (1997), who showed that Japanese speakers' production of the contrast between /l/ and /ɹ/ improved after a period of perceptual training, despite the participants receiving no explicit pronunciation instruction.

The SLM alongside related, but competing, models like the Perceptual Assimilation Model (PAM) (Best 1995) make similar predictions for the purposes of CAPT. Both SLM and PAM claim that phonemic perception is a) influenced by the L1 and b) a requisite for accurate production. These models and the empirical support for them motivate the use of perceptual training in CAPT, and we will return to this idea in the next section.

4 Components of a maximally effective CAPT system

In this section, the SLA research discussed above will be synthesised with research on CAPT itself in order to identify a set of features for a maximally effective or 'ideal' CAPT system.

One of the most important findings from modern pronunciation research has been the significance of suprasegmental features in listeners' perception of non-native speech, alongside evidence that suprasegmental errors can be reduced with training (Trouvain & Gut 2007). Derwing & Munro (1997) and Munro & Derwing (1995) showed that native evaluations of non-native speakers' prosodic naturalness strongly correlated with overall accent scores and intelligibility. Hahn (2004) found that native English speakers recalled 26% more content and evaluated non-native speakers significantly more favourably when primary stress was placed correctly. In Hardison (2004), French learners (English L1) significantly improved prosodic naturalness after explicit audiovisual prosodic training with pitch contour graphs compared to students who only received auditory feedback. Indeed, CAPT is uniquely positioned to provide feedback on prosody; features like duration and pitch are subtle and gradient, but easily measurable. Suprasegmental feature training is a rich topic and can only be superficially addressed in this paper, but the existing

evidence about their importance in the intelligibility and comprehensibility of L2 speech strongly supports their inclusion in CAPT.

Evidence suggests that it is advantageous in L2 learning for the input voice to be matched to the speaker. When using the Fluency CAPT system (Probst, Ke & Eskenazi 2002), users who practiced with an input voice matched to their own for F_0 and rate of speech reduced their segmental error rate by 43.3%, compared to a reduction of only 21.2% for learners who used a dissimilar input voice. Another variation of input voice matching is possible as well; early work by Nagano & Ozawa (1990) exploited the ability to modify acoustic features of the users' utterances to effectively create a perfectly matched input voice. Their study showed that users who trained with prosodically modified versions of their own voice improved prosodic naturalness more than learners who trained only with native speaker recordings (improvement of .88 vs. .33 on a 7-point Likert scale).

Input voice matching, however, conflicts somewhat with work on High Variability Phonetic Training (HVPT) (Logan, Lively & Pisoni 1991). HVPT research has shown that allophonic variation through the use of multiple speakers and varied segmental contexts facilitates categorical abstraction and leads to improved perception and production. Wong (2012), for example, showed that Cantonese learners of English improved their rate of target-like perception and production of /e/ vs. /æ/ more than twice as much after HVPT compared to a control group that received low variability phonetic training. Thomson (2011) showed this to be effective in CAPT as well; in his study, 21 of 24 Chinese learners of English improved vowel pronunciation after HVPT training with a novel CAPT system (two speakers and six different CV contexts per vowel).

Thus, given the lack of clarity about the most effective type or combination of input, a comprehensive option would include one or more 'default' speaker models matched to the user as closely as possible for sex, F_0, and/or speaking rate—potentially including voice modification as supplemental input—as well as certain exercises or options to hear other speakers produce the same sounds and utterances in the style of HVPT.

A further feature relevant to perceptual training is the ability for learners to listen back to their own productions. This topic has yet to be directly studied, but has been argued for by a number of researchers including Neri, Cucchiarini & Strik (2007) and Probst et al. (2002), and incorporated into their respective CAPT systems, *Dutch-CAPT* and *Fluency*. Given the technological simplicity of saving certain utterances and its non-intrusive nature (the user can choose whether or not to replay the utterance), utterance playback constitutes a form of perceptual training whose presence is well-motivated in an 'ideal' CAPT program.

User utterance playback can also be combined with scoring and gamification, defined as the application of game-like elements like point scoring or competition to non-game domains. Because pronunciation scoring is an inherent part of any CAPT system that uses ASR, this can be used overtly to allow the system to keep a record of the user's 'best' (highest-scoring) pronunciation, for example. Allowing learners to rehear their best pronunciations can be a source of encouragement, in that the learner knows they have produced a high(er) quality utterance in the past and are

capable of repeating or improving upon it. Gamification is still in its infancy as a research topic, but has been shown to increase user motivation across a wide variety of domains (see Hamari, Koivisto & Sarsa 2014 for a review) including language learning (Perry 2015).

Though it has not been investigated directly, there is a consensus in the research community that phonetic feedback needs to be clear and comprehensible to non-specialists. This view arose in response to CAPT programs in the early 2000s like Auralog's *TellMeMore*, which included waveform diagrams for what appear to be marketing purposes. Researchers like Neri, Cucchiarini, Strik & Boves (2002) vehemently argue that highly technical feedback like waveforms and spectrograms should be omitted. The nuances of particular feedback mediums and their relationship to different levels of feedback constitutes an area that warrants further research, but unless empirical evidence surfaces that supports the use of complex acoustic diagrams, they likely should be excluded.

As clearly seen with SLM and other L2 speech models, the L1 is implicated in the types of phonetic errors made by learners (Jones 1997). This suggests that an ideal system would include knowledge of the L1-L2 pairing, incorporating both universal mistakes made by almost all learners of the target language as well as more fine-grained L1-specific errors. Given the near impossibility of training ASR models with data from every possible L1, however, some attempts have been made to find a middle ground. Husby, Øvregaard, Wik, Bech, Albertsen, Nefzaoui, Skarpnes & Koreman (2011) proposed a method using only IPA charts to create an 'L1-L2map' that can inform the system about potential areas of difficulty. While clearly not as precise as a system trained on spoken learner corpora, their method was nonetheless able to make accurate predictions about L1-influenced error patterns. An alternative strategy is to have new CAPT users produce a special set of utterances, allowing the system to use multiple instances of each phoneme to determine which ones need remediation.

Regardless of which scoring algorithm is used by the speech recognition system, research supports the notion that it is better for CAPT systems to err on the side of acceptance in borderline utterances—an unidentified mispronunciation can be detected in the future, but labelling an acceptable pronunciation as badly-formed can be discouraging for learners. Evidence of this appears in qualitative feedback from users of HAFSS (Abdou, Hamid, Rashwan, Samir, Abdel-Hamid, Shahin & Nazih 2006) and ARTUR (Engwall, Bälter, Öster & Kjellström 2006) and is argued for by a range of other authors including Neri et al. (2007) and Popovici & Buică-Belciu (2012). It is important to note that even trained human instructors do not agree on all pronunciation errors. One can therefore establish the upper bound for error detection in CAPT as the interrater correlation across a wide test set. Studies in this area have yielded numbers from $r = 0.77$ (Franco, Neumeyer, Digalakis & Ronen 2000) to 0.89 (Hardison 2005), suggesting that at least 10% of pronunciation errors are ambiguous even to practitioners (though this likely varies cross-linguistically). This supports the notion that CAPT should not strive for 100% accuracy but instead should focus on high precision and recall, ignoring mistakes it identifies as marginal for the sake of user trust.

To briefly summarise, the components of an ideal CAPT system that have been identified here include, but are not necessarily limited to:

- Suprasegmental feedback
- Default input voice matching
- Multiple speaker models
- User recordings and playback control
- Scoring/gamification
- (Basic) L1 adaptation
- Exclusion of highly technical feedback
- Disregarding of errors below a certain confidence threshold

5 Evaluation of *Rosetta Stone* and *NativeAccent*

The features identified in the previous section will now be used to compare two CAPT systems. For space reasons, the evaluations here are by no means intended to be exhaustive, but instead simply help demonstrate how the framework can be used to illuminate strengths and weaknesses of a given CAPT system.

The first commercial system is *Rosetta Stone*, included because of its popularity and the nature of its evolution, having acquired a number of innovative CAPT and CALL systems like *TellMeMore* and *Livemocha* over its lifetime. *Rosetta Stone* is a complete language learning system based on immersion and picture identification, with special exercises for pronunciation training. It is entirely L1-independent. Information was gathered from the English and German versions.

During speaking exercises, *Rosetta Stone* first plays a recording of a native speaker with text and prompts the user to either repeat or reply to the utterance. If the ensuing pronunciation is deemed acceptable, the user moves on to the next task. If it is unacceptable, the user is allowed to either try again or hear the native speaker once more. *Rosetta Stone* varies native speakers throughout and between exercises, but does not qualify as HVPT because each utterance only has one speaker model. It does not include any segmental or suprasegmental feedback, instead providing feedback at the word and utterance levels, and does not correct every mistake (in one example during testing by the author, the system accepted 'weiß' [vaɪs] instead of the target word 'heiß' [haɪs]), though it is unclear whether errors are ignored based on ASR confidence thresholds, relative importance for intelligibility at a given proficiency, or other reasons. Simple gamification is included with a bar that fills up based on the quality of the user's attempt.

Under the framework established in the previous section, it is clear that the CAPT portion of *Rosetta Stone* lacks a number of features that have been shown to be effective. The hypothesis is therefore that *Rosetta Stone* would achieve larger learning gains if users had the ability to listen back to their own attempt, received

suprasegmental feedback, and could hear the same utterance produced by more than one speaker, including one with a voice as similar as possible to their own.

The other commercial system to be discussed here is *NativeAccent* (Carnegie Speech). *NativeAccent*, a standalone CAPT system for English, was chosen because of its unique emphasis on segmental and suprasegmental feedback. *NativeAccent* adapts to 63 different L1s using contrastive language analysis and uses both auditory and visual feedback. Visual feedback includes diagrams of the vocal tract, arrows representing pitch movement, and frontal views of native speakers. Before beginning the training, users complete a 30-minute pronunciation assessment that allows the system to determine which sounds need the most improvement. The program includes different exercises for 38 phonemes (both consonants and vowels) and various suprasegmental features including pitch, duration, and pausing.

From the structure of the program, it is clear that *NativeAccent* represents a commercial system whose strategy is much more closely in line with speech acquisition research. Not all criteria are met, however, resulting in the hypothesis that features like voice modification, gamification, and HVPT are could further increase learning gains.

6 Conclusions and future directions

In this paper, a framework was presented for evaluating phonetic feedback in CAPT. A list of features with empirical support was assembled, including explicit prosodic feedback, perceptual training and user playback control, input matching, HVPT, L1 adaptation, and gamification. Two example CAPT systems were evaluated in accordance with that list, leading to the hypothesis that *Rosetta Stone* is not implementing phonetic feedback as effectively as NativeAccent, by virtue of the fact that *NativeAccent* includes a wider range of individually effective features. One of the ensuing predictions is that *NativeAccent* will lead to larger or faster learning gains than *Rosetta Stone* after a fixed period of pronunciation training. This type of study, comparing either two different CAPT systems or two variations of the same CAPT system with different phonetic feedback features, constitutes a key area of inquiry going forward. Other more equivocal pedagogical features including tongue awareness training and synthesised speech also warrant further investigation.

CAPT is already a powerful tool for a difficult aspect of language learning that classroom instruction often fails to address, but there is significant room for improvement. As further research emerges, ASR technology continues to grow in power and accessibility, and developers begin to consider more carefully which feedback strategies are worth including, CAPT finds itself well-poised to play a major role in the future of language learning and speech pathology.

References

Abdou, Sherif Mahdy, Salah Eldeen Hamid, Mohsen Rashwan, Abdurrahman Samir, Ossama Abdel-Hamid, Mostafa Shahin & Waleed Nazih. 2006. Computer aided

pronunciation learning system using speech recognition techniques. In *Ninth International Conference on Spoken Language Processing*.

Abercrombie, David. 1949. Teaching pronunciation. *ELT Journal* 3(5). 113–122.

Avery, Peter & Susan Ehrlich. 1992. *Teaching American English pronunciation* Oxford handbooks for language teachers. Oxford [England] ; New York: Oxford University Press.

Best, Catherine T. 1995. A direct realist view of cross-language speech perception. In *Speech perception and linguistic experience: Issues in cross-language research*, 171–204. Baltimore: York Press.

Bradlow, Ann R., David B. Pisoni, Reiko Akahane-Yamada & Yoh'ichi Tohkura. 1997. Training Japanese listeners to identify English/r/and/l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America* 101(4). 2299–2310.

Chen, N. F. & H. Li. 2016. Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1–7.

Derwing, Tracey M. & Murray J. Munro. 1997. Accent, intelligibility, and comprehensibility. *Studies in second language acquisition* 19(1). 1–16.

Engwall, Olov. 2012. Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning* 25(1). 37–64.

Engwall, Olov, Olle Bälter, Anne-Marie Öster & Hedvig Kjellström. 2006. Designing the user interface of the computer-based speech training system ARTUR based on early user tests. *Behaviour & Information Technology* 25(4). 353–365.

Felps, Daniel, Heather Bortfeld & Ricardo Gutierrez-Osuna. 2009. Foreign accent conversion in computer assisted pronunciation training. *Speech Communication* 51(10). 920–932.

Flege, James E. 2007. Language contact in bilingualism: Phonetic system interactions. *Laboratory phonology* 9. 353–382.

Flege, James Emil. 1987. A critical period for learning to pronounce foreign languages? *Applied linguistics* 8(2). 162–177.

Flege, James Emil. 1995. Second language speech learning: Theory, findings, and problems. In *Speech perception and linguistic experience: Issues in cross-language research*, 233–277. Timonium, MD: York Press.

Franco, Horacio, Leonardo Neumeyer, Vassilios Digalakis & Orith Ronen. 2000. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication* 30(2). 121–130.

Hahn, Laura D. 2004. Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL quarterly* 38(2). 201–223.

Hamari, J., J. Koivisto & H. Sarsa. 2014. Does gamification work? – A literature review of empirical studies on gamification. In *2014 47th Hawaii International Conference on System Sciences*, 3025–3034.

Hansen, Thomas K. 2006. Computer assisted pronunciation training: The four 'K's of feedback. In *Current Developments in Technology-Assisted Education*, 324–326.

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

Hardison, Debra M. 2005. Contextualized computer-based L2 prosody training: Evaluating the effects of discourse context and video input. *CALICO Journal* 22(2). 175–190.

Hsu, Liwei. 2016. An empirical examination of EFL learners' perceptual learning styles and acceptance of ASR-based computer-assisted pronunciation training. *Computer Assisted Language Learning* 29(5). 881–900.

Husby, Olaf, Åsta Øvregaard, Preben Wik, Øyvind Bech, Egil Albertsen, Sissel Nefzaoui, Eli Skarpnes & Jacques Koreman. 2011. Dealing with L1 background and L2 dialects in Norwegian CAPT. In *Speech and Language Technology in Education*.

Jones, Rodney H. 1997. Beyond 'listen and repeat': pronunciation teaching materials and theories of second language acquisition. *System* 25(1). 103–112.

Krashen, Stephen D. 1985. *The input hypothesis: Issues and implications*. London ; New York: Longman.

Lado, Robert. 1957. *Linguistics across cultures: Applied linguistics for language teachers*. Ann Arbor: Univ. of Michigan Press.

Logan, John S., Scott E. Lively & David B. Pisoni. 1991. Training Japanese listeners to identify English /r/ and /l/: A first peport. *The Journal of the Acoustical Society of America* 89(2). 874–886.

Munro, Murray J. & Tracey M. Derwing. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45(1). 73–97.

Munro, Murray J. & Tracey M. Derwing. 2011. The foundations of accent and intelligibility in pronunciation research. *Language Teaching* 44(03). 316–327.

Nagano, Keiko & Kazunori Ozawa. 1990. English speech training using voice conversion. In *Proceedings of ICSLP 90*, 1169–1173. Kobe.

Nair-Venugopal, Shanta. 2003. Intelligibility in English: Of What Relevance Today to Intercultural Communication? *Language and Intercultural Communication* 3(1). 36–47.

Neri, Ambra, Catia Cucchiarini & Helmer Strik. 2007. Pronunciation training in Dutch as a second language on the basis of automatic speech recognition. *Stem-, Spraak- en Taalpathologie* 15(2).

Neri, Ambra, Catia Cucchiarini, Helmer Strik & Lou Boves. 2002. The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning* 15(5). 441–467.

Penfield, Wilder & Lamar Roberts. 1959. *Speech and brain mechanisms*. Princeton University Press.

Perry, Bernadette. 2015. Gamifying French language learning: A case study examining a quest-based, augmented reality mobile learning-tool. *Procedia - Social and Behavioral Sciences* 174. 2308–2315.

Popovici, Doru-Vlad & Cristian Buică-Belciu. 2012. Professional challenges in computer-assisted speech therapy. *Procedia - Social and Behavioral Sciences* 33.

518–522.
Probst, Katharina, Yan Ke & Maxine Eskenazi. 2002. Enhancing foreign language tutors — in search of the golden speaker. *Speech Communication* 37. 161–173.
Richards, Jack C., John Platt & Heidi Platt. 1992. *Longman dictionary of language teaching and applied linguistics*. Harlow: Longman 2nd edn.
Scovel, Thomas. 1988. *A time to speak: A psycholinguistic inquiry into the critical period for human speech* Issues in second language research. Cambridge [England] ; New York: Newbury House.
Thomson, Ron I. 2011. Computer assisted pronunciation training: targeting second language vowel perception improves pronunciation. *CALICO Journal* 28(3). 744–765.
Trouvain, Jürgen & Ulrike Gut (eds.). 2007. *Non-native prosody: Phonetic description and teaching practice* (Trends in linguistics 186). Berlin ; New York: Mouton de Gruyter.
Wong, Janice Wing Sze. 2012. Training the perception and production of English /e/ and /æ/ of Cantonese ESL learners: A comparison of low vs. high variability phonetic training. In *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Sydney.
Young, Dolly J. 1990. An investigation of students' perspectives on anxiety and speaking. *Foreign Language Annals* 23(6). 539–554.

Nick Ballou
University of Cambridge
nb567@cam.ac.uk