Perceptual Feedback In Computer Assisted Pronunciation Training: A Survey

Renlong Ai
Language Technology Laboratory, DFKI GmbH
Alt-Moabit 91c, Berlin, Germany
renlong.ai@dfki.de

Abstract

This survey examines the feedback in current Computer Assisted Pronunciation Training (CAPT) systems and focus on perceptual feedback. The advantages of perceptual feedback are presented, while on the other hand, the reasons why it has not been integrated into commercial CAPT systems are also discussed. This is followed by a suggestion of possible directions of future work.

1 Introduction

In the last decades, CAPT has proved its potential in digital software market. Modern CAPT software aims no longer at simply assisting human teachers by providing various attractive teaching materials, but rather at replacing them by providing the learners with a private learning environment, self-paced practises, and especially instant feedback. Different types of feedback have always been highlighted in CAPT systems. However, it remains to be seen whether these types of feedback are really helpful to the learners, or are rather a demonstration of what modern technology can achieve. Considering whether a feedback is effective and necessary in CAPT systems, Hansen (2006) described four criteria in his work, namely:

- *Comprehensive*: if the feedback is easy to understand.
- *Qualitative*: if the feedback can decide whether a correct phoneme was used.
- *Quantitative*: if the feedback can decide whether a phoneme of correct length was used.
- *Corrective*: if the feedback provides information for improvement.

Ways of providing feedback grow as far as technology enables, but the four points above should be considered seriously while designing a practical and user-friendly feedback.

In Section 2 the existing feedback in available CAPT systems is examined. In Section 3 recent works on perceptual feedback are reviewed, which is still not quite common in commercial CAPT systems. In Section 4, some suggestions on integrating perceptual feedback into current CAPT systems in a more reliable way are sketched. Finally, conclusions are presented in Section 5.

2 Feedback In CAPT Systems

Feedback nowadays has been playing a much more significant role than simply telling the learner ”You have done right!” or ”This doesn’t sound good enough”. Thanks to the newer technologies in signal processing, it can pinpoint specific errors and even provide corrective information (Crompton and Rodrigues, 2001). One of the earliest type of feedback, which is still used in modern CAPT systems like TELLMEMORE (2013), is to show the waveform of both the L1 (the teacher’s or native’s) speech and the L2 learner’s one. Although the difference of the two curves can be perceived via comparison, the learner is still left with the question why they are different and what he should do to make his own curve similar to the native one. He might then try many times randomly to produce the right pronunciation, which may lead to reinforcing bad habits and result in fossilisation (Eskenazi, 1999). To solve this, forced alignment was introduced. It allowed to pinpoint the wrong phoneme, and give suggestion to increase or decrease the pitch or energy, like in EyeSpeak (2013), or mark the wrong pronounced phoneme to notify the learner, like in FonixTalk SDK (2013).

Another common type of feedback among CAPT systems is to provide a score. A score of the

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 1–6,
Hissar, Bulgaria, 9-11 September 2013.

overall comprehensibility of learner's utterance is usually acquired via Automatic Speech Recognition (ASR), like in SpeechRater Engine (Zechner et al., 2007), which is part of TOFEL (Test of English as a Foreign Language) since 2006. Many CAPT systems also provide word-level or even phoneme-level scoring, like in speexx (2013). Although scoring is appreciated among language students due to the immediate information on the quality it provides (Atwell et al., 1999) , it is regarded merely as an overall feedback, because if no detail follows, the number itself will not show any information for the learner to improve his speech.

To provide more pedagogical and intuitive feedback, the situation of classroom teaching is considered. Imaging a student makes a wrong pronunciation, the teacher would then show him how exactly the phoneme is pronounced, maybe by slowing down the action of mouth while pronouncing or pointing out how the tongue should be placed (Morley, 1991). After investigating such behaviours, Engwall et. al. (2006) presented different levels of feedback implemented in the ARTUR (the ARticulaton TUtoR) pronunciation training system. With the help of a camera and knowledge of the relation between facial and vocal tract movements, the system can provide feedback on which part of the human vocal system did not move in the right way to produce the correct sound, the tongue, the teeth or the palate, and show in 3D animations how to pronounce the right way.

These types of feedback are known as visual feedback and automatic diagnoses (Bonneau and Colotte, 2011) that show information with graphic user interface. Besides these, perceptual feedback, which is provided via speech and/or speech manipulations, is also used more and more common in modern CAPT systems.

3 Types Of Perceptual Feedback

Simple playback of the native and learner's speech and leaving the work of comparing them to the learners will not help them to perceive difference between the sound they produced and the correct targets sound because of their L1 influence (Flege, 1995), hence, the importance of producing perceivable feedback has been increasingly realised by CAPT system vendors and many ways of enhancing learns' perception have been tried.

3.1 Speech Synthesis For Corrective Feedback

Meng et. al. (2010) implemented a perturbation model that resynthesise the speech to convey focus. They modified the energy, max and min f0 and the duration of the focused speech, and then use STRAIGHT (Kawahara, 2006), a speech signal process tool, for the resynthesising. This perturbation model was extended later to provide emphasis (Meng et al., 2012). A two-pass decision tree was constructed to cluster acoustic variations between emphatic and neutral speech. The questions for decision tree construction were designed according to word, syllable and phone layers. Finally, Support vector machines (SVMs) were used to predict acoustic variations for all the leaves of main tree (at word and syllable layers) and subtrees (at phone layer). In such way, learner's attention can be drawn onto the emphasised segments so that they can perceive the feedback in the right way.

In the study of De La Rosa et. al. (2010), it was shown that students of English Language benefit from spoken language input, which they are encourage to listen; in particular this study shows that English text-to-speech may be good enough for that purpose. A similar study for French Language was presented in (Handley, 2009), where four French TTS systems are evaluated to be used within CALL applications. In these last two cases speech synthesis is used more as a complement to reinforce the learning process, that is, in most of the cases as a way of listen and repeat, without further emphasis.

3.2 Emphasis And Exaggeration

Yoram and Hirose (1996) presented a feedback in their system which produces exaggerated speech to emphasis the problematic part in the learner's utterance, as a trial to imitate human teachers, e.g. if the learner placed a stress on the wrong syllable in a word, the teacher would use a more extreme pitch value, higher energy and slower speech rate at the right and wrong stressing points to demonstrate the difference. As feedback, the system plays a modified version of the learner's speech with exaggerated stress to notify him where his problem is. A Klatt formant synthesiser was used to modify the f0, rate and intensity of the speech.

Lu et. al. (2012) looked into the idea of exaggeration further by investigating methods that modified different parameters. They evaluated duration-based, pitch-based and intensity-based stress exaggeration, and in the end combined these three to perform the final automatic stress exaggeration, which, according to their experiment, raised the perception accuracy from 0.6229 to 0.7832.

3.3 Prosody Transplantation Or Voice Conversion

In the previous sections we have seen that speech synthesis techniques can be used to provide feedback to the learner by modifying some prosody parameters of the learner's speech in order to focus on particular problems or to exaggerate them. Other forms of feedback intend to modify the learner's voice by replacing or "transplanting" properties of the teacher's voice. The objective is then that the learner can hear the correct prosody in his/her own voice. This idea has been motivated by studies that indicate that learners benefit more from audio feedback when they can listen to a voice very similar to their own (Eskenazi, 2009) or when they can hear their own voice modified with correct prosody (Bissiri et al., 2006) (Felps et al., 2009).

Prosody transplantation tries to adjust the prosody of the learner to the native's, so that the learner can perceive the right prosody in his own voice. According to the research of Nagano and Ozawa (1990), learners' speech sounds more like native after they tried to mimic their own voice with modified prosody than to mimic the original native voice. The effect is more remarkable if the L1 language is non-tonal, e.g. English and the target language is tonal, e.g. Mandarin (Peabody and Seneff, 2006). Pitch synchronous overlap and add (PSOLA) (Moulines and Charpentier, 1990) has been widely used in handling pitch modifications. Many different approaches, namely time-domain (TD) PSOLA, linear prediction (LP) PSOLA and Fourier-domain (FD) PSOLA, have been applied to generate effective and robust prosody transplantation.

Felps et. al. (2009) provided prosodically corrected versions of the learners' utterances as feedback by performing time and pitch scale before applying FD PSOLA to the user and target speech. Latsch and Netto (2011) presented in their PS-DTW-OLA algorithm a computationally efficient method that maximises the spectral similarity between the target and reference speech. They performed dynamic time warping (DTW) algorithm to the target and reference speech signals so that their time-warping become compatible to what the TD PSOLA algorithm requires. By combining the two algorithms, pitch-mark interpolations was avoided and the target was transplanted with high frame similarity. Cabral and Oliveira (2005) modified the standard LP-PSOLA algorithm, in which they used smaller period instead of twice of the original period for the weighting window length to prevent the overlapping factor to increase above 50%. They also developed a pitch synchronous time-scaling (PSTS) algorithm, which gives a better representation of the residual after prosodic modification and overcomes the problem of energy fluctuation when the pitch modification factor is large.

Vocoding, which was originally used in radio communication, can be also utilised in performing prosody transplantation and/or voice conversion. By passing the f0, bandpass voicing and Fourier magnitude of the target speech and the Mel-frequency cepstral coefficients (MFCCs) of the learner's speech, the vocoder is able to generate utterance with L2 learner's voice and the pitch contours of the native voice. Recently, vocoder techniques have been also used in flattening the spectrum for further processing, as shown in the work of Felps et. al. (2009).

An overview of the different types of perceptual feedback, the acoustic parameters they changed and the techniques they used, is summarised in Table 1.

4 Perceptual Feedback: Pros, Cons And Challenges

Compared to other feedback, the most obvious advantage of perceptual feedback is that the corrective information is provided in a most comprehensive way: via the language itself. To overcome the problem that it is hard for L2 learners to perceive the information in a utterance read by a native speaker, methods can be applied to their own voice so that it is easier for them to tell the difference. However, the most directly way to tell the learners where the error is located is still to show them via graphic or text. Hence, the ideal feedback that a CAPT system should provide is a combination of visual and perceptual feedback in the way that automatic diagnoses identify the errors and show them, while perceptual feedback

Perceptual Feedback	**Ref**	**Modify/replaced parameters**	**Method or technique**
Speech synthesis	(Meng et al., 2010)	F0, duration	STRAIGHT
	(Meng et al., 2012)	F0, duration	decision tree, support vector machines
Emphasis and exaggeration	(Yoram and Hirose, 1996)	F0, rate and intensity	Klatt formant synthesiser
	(Lu et al., 2012)	F0, duration and intensity	PSOLA
Voice conversion or prosody transplantation	(Felps et al., 2009)	duration, pitch contour, spectrum	FD-PSOLA, spectral envelope vocoder
	(Latsch and Netto, 2011)	duration, pitch contour	TD-PSOLA, DTW
	(Cabral and Oliveira, 2005)	pitch and duration	LP-PSOLA, time-scaling

Table 1: Perceptual feedback, acoustic parameters modified or replaced and the techniques used.

helps to correct them.

One argument about perceptual feedback is: in most works, only prosodic errors like pitch and durations are taken care of, and in most experiments that prove the feasibility of perceptual feedback, the native and L2 speech that are used as input differ only prosodically. Although the results of these experiments show the advantage of perceptual feedback, e.g. the learners did improve their prosody better after hearing modified version of their own speech than simply hearing the native ones, it is not the real case in L2 language teaching, at least not for the beginners, who might usually change the margins between syllables or delete the syllables depending on their familiarity to the syllables and their sonority (Carlisle, 2001). These add difficulties to the forced alignment or dynamic time warping procedure, which is necessary before the pitch modification, and hence the outcome will also not be as expected (Brognaux et al., 2012).

Perceptual feedback has been widely discussed and researched but not yet fully deployed in commercial CAPT systems. In order to provide more reliable feedback, the following considerations should be taken into account:

- For the moment, perceptual feedback should be applied to advanced learners who focus on improving their prosody, or to the case that only prosodic errors are detected in the learner's speech, i.e. if other speech errors are found, e.g. phoneme deletion, the learner gets notified via other means and corrects it; if only a stress is misplaced by the learner, he will hear a modified version of his own speech where the stress is placed right so that he can perceive his stress error.

- More robust forced alignment tool for non-native speech has been under development for years. In the near future, it should be able to handle pronunciation errors and provide right time-alignment even if the text and audio do not 100% match. Until then, an L1 independent forced alignment tool, which is one of the bottlenecks in speech technology nowadays, will be open to researchers, so in the near future, more accurate perceptual feedback can be generated.

5 Conclusions

In this paper first, various visual and diagnostic feedback in current CAPT systems are examined. Then existing research on providing perceptual feedback via multiple means is summarised. After the literature review presented in this paper, it has been found that the perceptual feedback in CAPT systems can be classified in 3 types: via speech synthesis, providing emphasis and exaggeration, and performing prosody transplantation. The three methods modify or replace prosody parameters like F0 and durations and the most used speech signal processing technology is PSOLA. Subsequently, the pros and cons of perceptual feedback are analysed taking into consideration the difficulties of its implementation in commercial CAPT systems. Finally, a suggestion on integrating perceptual feedback in future work is made.

6 Acknowledgement

This work is supported by the project SPRINTER, funded by the German Ministry of Education and Research (BMBF) within the program KMU-innovativ under number 01IS12006A. The author would like to express his very great appreciation to Marcela Charfuelan for her valuable suggestions during the survey.

References

Eric Atwell, Dan Herron, Peter Howarth, Rachel Morton, and Hartmut Wick. 1999. Pronunciation training: Requirements and solutions. *ISLE Deliverable*, 1.

Maria Paola Bissiri, Hartmut R Pfitzinger, and Hans G Tillmann. 2006. Lexical stress training of german compounds for italian speakers by means of resynthesis and emphasis. In *Proc. of the 11th Australasian Int. Conf. on Speech Science and Technology (SST 2006). Auckland*, pages 24–29.

Anne Bonneau and Vincent Colotte. 2011. Automatic feedback for l2 prosody learning. *Speech and Language Technologies*, pages 55–70.

Sandrine Brognaux, Sophie Roekhaut, Thomas Drugman, and Richard Beaufort. 2012. Train&align: A new online tool for automatic phonetic alignment. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 416–421. IEEE.

Joao P Cabral and Luıs C Oliveira. 2005. Pitch-synchronous time-scaling for prosodic and voice quality transformations. In *Proc. Interspeech*, pages 1137–1140.

Robert S Carlisle. 2001. Syllable structure universals and second language acquisition. *IJES, International Journal of English Studies*, 1(1):1–19.

P. Crompton and S. Rodrigues. 2001. The role and nature of feedback on students learning grammar: A small scale study on the use of feedback in call in language learning. In *Proceedings of the workshop on Computer Assisted Language Learning, Artificial Intelligence in Education Conference*, pages 70–82.

Kevin De-La-Rosa, Gabriel Parent, and Maxine Eskenazi. 2010. Multimodal learning of words: A study on the use of speech synthesis to reinforce written text in l2 language learning. In *Proceedings of the Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan.

Olov Engwall, Olle Bälter, Anne-Marie Öster, and Hedvig Kjellström. 2006. Feedback management in the pronunciation training system artur. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 231–234. ACM.

Maxine Eskenazi. 1999. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language learning & technology*, 2(2):62–76.

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832 – 844.

EyeSpeak. 2013. Language learning software. Online: http://www.eyespeakenglish.com.

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

James E Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, pages 233–273.

FonixTalk SDK. 2013. Speech FX Text to Speech. Online: http://www.speechfxinc.com.

Zöe Handley. 2009. Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication*, 51(10):906 – 919.

Thomas K Hansen. 2006. Computer assisted pronunciation training: the four'k's of feedback. In *4th Internat. Conf. on Multimedia and Information and Communication Technologies in Education, Seville, Spain*, pages 342–346. Citeseer.

Hideki Kawahara. 2006. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6):349–353.

V. L. Latsch and S. L. Netto. 2011. Pitch-synchronous time alignment of speech signals for prosody transplantation. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, Rio de Janeiro, Brazil.

Jingli Lu, Ruili Wang, and LiyanageC. Silva. 2012. Automatic stress exaggeration by prosody modification to assist language learners perceive sentence stress. *International Journal of Speech Technology*, 15(2):87–98.

Fanbo Meng, Helen Meng, Zhiyong Wu, and Lianhong Cai. 2010. Synthesizing expressive speech to convey focus using a perturbation model for computer-aided pronunciation training. In *Proceedings of the Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan.

Fanbo Meng, Zhiyong Wu, Helen Meng, Jia Jia, and Lianhong Cai. 2012. Generating emphasis from neutral speech using hierarchical perturbation model by decision tree and support vector machine. In *Proceedings of International Colloquium on Automata, Languages and Programming (ICALP 2012)*, Warwik, UK.

Joan Morley. 1991. The pronunciation component in teaching english to speakers of other languages. *Tesol Quarterly*, 25(3):481–520.

Eric Moulines and Francis Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467.

Keiko Nagano and Kazunori Ozawa. 1990. English speech training using voice conversion. In *First International Conference on Spoken Language Processing*.

Mitchell Peabody and Stephanie Seneff. 2006. Towards automatic tone correction in non-native mandarin. In *Chinese Spoken Language Processing*, pages 602–613. Springer.

speexx. 2013. Language learning software. Online: http://speexx.com/en/.

TELLMEMORE. 2013. Language learning software. Online: http://www.tellmemore.com.

M. Yoram and K. Hirose. 1996. Language training system utilizing speech modification. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1449–1452 vol.3.

Klaus Zechner, Derrick Higgins, and Xiaoming Xi. 2007. Speechrater: A construct-driven approach to scoring spontaneous non-native speech. *Proc. SLaTE*.