ENUNCIATE: AN INTERNET-ACCESSIBLE COMPUTER-AIDED PRONUNCIATION TRAINING SYSTEM AND RELATED USER EVALUATIONS

Ka-Wa Yuen, Wai-Kim Leung, Peng-fei Liu, Ka-Ho Wong, Xiao-jun Qian,Wai-Kit Lo and Helen Meng

Human-Computer Communications Laboratory
Shun-Hing Institute of Advanced Engineering
Department of Systems Engineering and Engineering Managements
The Chinese University of Hong Kong
Hong Kong SAR
{kwyuen, wkleung, pfliu, khwong, xjqian, wklo, hmmeng}@se.cuhk.edu.hk

ABSTRACT

Abstract—This paper presents our group's latest progress in developing Enunciate — an online computer-aided pronunciation training (CAPT) system for Chinese learners of English. Presently, the system targets segmental pronunciation errors. It consists of an audio-enabled web interface, a speech recognizer for mispronunciation detection and diagnosis, a speech synthesizer and a viseme animator. We present a summary of the system's architecture and major interactive features. We also present statistics from evaluations by English teachers and university students who participate in pilot trials. We are also extending the system to cover suprasegmental training and mobile access.

Index Terms— Internet-accessible, User Evaluation, CAPT, Language Learning, System Architecture

1. INTRODUCTION

Acquisition of L2 (secondary language) spoken language tends to be influenced by L1 (primary language) features. Chinese has stark linguistic contrasts in comparison with English, so negative transfer from L1 often leads to pronunciation inaccuracies of L2 production for Chinese learners of English. The situation is even worse when the learner fails to distinguish certain contrastive pairs between the two languages.

As pronunciation training is considered an iterative self-feedback process, pronunciation improvement requires persistent practice in both productive and perceptual training. An online computer-aided pronunciation system can serve as an anxiety-free, widely accessible and personalized tutor that fills the gap between the shortage of qualified language teachers and the growing need from the students. To support productive training (i.e., eliciting speech from the learner for analysis), we have been applying automatic speech recognition techniques to our system that enables detection and diagnosis of targeted pronunciation inaccuracies predicted by some prior knowledge. To support perceptual training (i.e., developing the learners' skills to accurately discriminate among sounds of the target language), we have been developing automatic response generation that provides multimodal visualization of the production process (e.g. through text-to-audiovisual speech synthesis) in addition to speech synthesis. The generated responses are intended as helpful instructions that guide error correction and improvement.

The rest of this paper is organized as follows: Section 2 briefly describes our previous work in CAPT. Section 3 gives a high level description of the Enunciate architecture. Section 4 describes some important components of Enunciate system in detail. Section 5 first describes a pronunciation learning cycle in Enunciate and then summarizes the interactional features in Enunciate. Section 6 presents the feedback from subjects after using Enunciate. Finally, Section 7 presents conclusions and future directions.

2. PREVIOUS WORK

The field of computer-aided pronunciation training (CAPT) is a wide topic covering speech analysis, speech recognition, audio and visual-speech synthesis, animated agent, feedback generation and even architectural development. Many research groups are actively involved in different aspects [1-3] and some systems have also entered the market. Products include those that can provide phone-level scoring, pitch, fluency and emphasis analysis, e.g. [4], as well as others that focus on stress teaching [5] by giving feedback on an exaggerated stress through modifying pitch, duration and intensity. Aside from detection, researchers also developed several feedback methods, including those using automatically generated scores that correlated well with human scores [6] and those with animated agents [7]. Our work in CAPT [8] includes automatic mispronunciation detection and diagnosis for English, as well as expressive

speech and visual speech with articulatory animation that synthesized during feedback generation.

3. SYSTEM ARCHITECTURE

The Enunciate system is developed on the WAMI Toolkit from MIT [9] and adopts the Spring [10] and Hibernate [11] frameworks. As an internet-accessible web application, Enunciate utilizes automatic speech recognition (ASR) technologies to detect and diagnose mispronunciations in real time. It prompts for leaner's requests (e.g. speech input) and prepares corresponding responses via invoking and coordinating backend components such as the speech synthesizer and the speech recognizer.

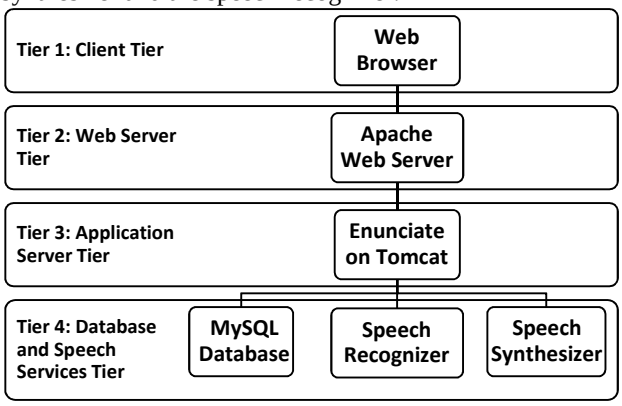

Figure 1: High level Enunciate architecture.

As illustrated in Figure 1, the Enunciate architecture is composed of four tiers: the "Client Tier", the "Web Server Tier", the "Application Server Tier" and the "Databases and Speech Services Tier". The description below focuses on how these tiers work together from top to bottom.

When a learner visits Enunciate via a web browser, a Java applet for audio recording and playback will be loaded together with the web interface.

As soon as the learner starts to record speech, the audio stream is relayed to the Apache web server in Tier 2, which is a proxy to dispatch related requests to the Tomcat server in Tier 3 for further processing. For example, the learner's recording from the Java applet is sent to the Tomcat server in the form of a bytes stream, the text input from the web interface in the form of an HTTP request, etc.

The Tomcat server is the runtime container of the Enunciate web application, which accumulates the received audio bytes stream and submits it chunk by chunk to the backend speech recognizer in Tier 4.

The speech recognizer identifies a list of phones from the received audio bytes stream and returns the recognition results to the component in Tier 3. Mispronunciation detection and diagnosis is done in Tier 3 by phonetic alignment between the recognized phones and the canonical phones generated by the speech synthesizer in Tier 4.

Both the speech recognizer and the speech synthesizer are implemented as XML-RPC [12] servers in C++.

The audio recordings and the recognition results are saved in the MySQL [13] database, which is also used to store user accounts, pre-defined lessons and Enunciate sessions.

Enunciate also includes a user account manager, a lesson manager and a session manager. All the learners' recordings are saved in each individual's user account. Some pre-defined targeted lessons are designed to help Chinese learners practice English. The session manager is introduced to support concurrency.

The Enunciate architecture can be extended to support mobile devices as well. For example, we have developed a beta version of mobile client on the Android platform.

4. COMPONENT TECHNOLOGIES

4.1 Audio Controller

Audio recording is not supported in current web browsers. Hence a Java applet, which is part of the WAMI Toolkit, is used for audio recording and playback on the client tier. An audio controller is included in the applet to stream speech data between Tier 1 and Tier 2. The learner's speech input is recorded using a microphone and is passed down the tiers, while synthesized speech is returned from the synthesizer and passed up the tiers.

4.2 Speech Synthesis

The speech synthesis component, developed based on the open-source Text-to-Speech Flite [16] system, is used to generate canonical speech for the learners. The phone-level canonical transcription is also fed to the recognition component (see section 4.4), as well as to the animation component (see section 4.5) together with the timing information of the phones.

4.3 Mispronunciation Prediction

Given a canonical phone sequence returned from the speech synthesis component, as well as a set of pre-defined rules either authored from knowledge or derived from collected data [17], this component forms an Extended Recognition Network [8] that encodes a graph of possible mispronunciations. The network is later traversed to construct an *extended pronunciation dictionary* which contains not only canonical pronunciations but also predicted mispronunciations.

4.4 Speech Recognition

This component is implemented with ATK [14] (API for HTK [15]) and incorporate the extended pronunciation dictionary.

For a given sentence prompt and the corresponding learner's speech, both the canonical phone-level

transcription by the speech synthesizer and the recognized phone-level transcription by ASR are aligned at the phone level. In the alignment, mismatched phone pairs indentified as insertions, deletions or substitutions are regarded as mispronunciations. The phone accuracy of the speech recognizer is about 73.02%, based on experiments with our CHLOE corpus [18].

4.5 Articulatory Animation

The articulatory animation component is a separate library developed in Java to help learners visualize motions of lips, tongue, and mouth as well as the opening of nasal passage. This component accepts speech signals, phoneme sequences and time boundary information from the speech synthesizer and morphs a set of two-dimensional viseme models to generate the animation. Details can be found in [19].

5. TYPICAL INTERACTIONAL FEATURES

5.1 The Pronunciation Learning Cycle

Figure 2 shows the pronunciation learning cycle available in Enunciate to help learners improve their pronunciation. Learners can first listen to the canonical pronunciation and then record their own speech. Enunciate helps pinpoint detected mispronunciations. Learners can further examine the appropriate articulatory movements. They can then repeat the learning cycle with the knowledge of their mispronunciations in contrast to the canonical pronunciations.

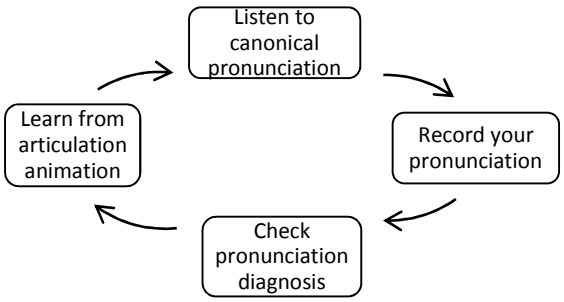

Figure 2: Pronunciation learning cycle in Enunciate.

5.2 Pre-defined Lessons and Self-Practice

In Enunciate, learners can either work through pre-defined lessons to practice various pronunciations, or input free text in the self-practice section. The pre-defined lessons are designed based on contrastive analysis of L1 and L2, and include typical English words that native Cantonese speakers tend to mispronounce. In addition, the self-practice section allows learners to enter any words or sentences for practicing.

5.3 System Feedback

System feedback is provided through the modalities of text, audio and visual features. Learners can read the pronunciation of any English words or sentences in International Phonetic Alphabet. They can also listen to the synthesized speech of the input words. Enunciate translates the recognition results into comprehensible feedback by highlighting the mispronounced words and providing phonetically aligned transcriptions of both the canonical pronunciation and the learners' own pronunciation, as shown in Figure 3.

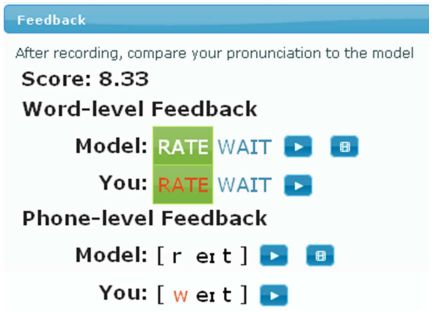

Figure 3: An example of diagnostic feedback.

Learners can replay their own recordings and compare it with the canonical pronunciation. Enunciate also supports audio playback of individual words in the sentence.

The articulatory animation, as shown in Figure 4, assists learners in getting the whole picture of the articulation in order to acquire the correct pronunciation. It supports two rates of playback: normal speed (150 words per minute) and slow speed (18.75 words per minutes). It also provides both the midsagittal view and the front view. Similarly, every single word of the animation can be replayed based on the learners' selection.

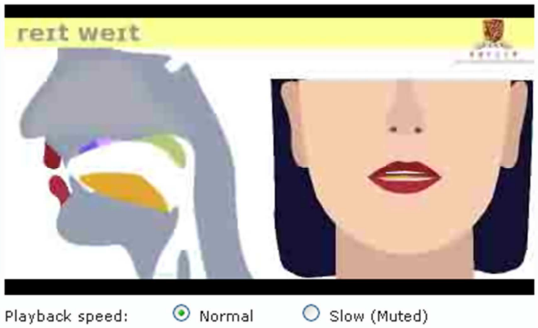

Figure 4: An example of articulatory animation.

6. USER EVALUATION

We conduct pilot trial of the Enunciation system in a master level phonetic course and an undergraduate English enhancement course at CUHK for the purpose of evaluation. Subjects do the evaluation test in groups of 20 people. They all use the same model of headsets (Sennheiser PC156), and

are required to go through 71 exercises and complete the evaluation form. Two groups of people participated in this evaluation. Group A (the masters level Phonetics course) consists of 19 English teachers from primary or secondary schools. Group B (the undergraduate English enhancement course) has 156 students from 9 English enhancement classes in CUHK.

6.1 Demographics of Subjects

We receive 19 evaluation forms from Group A and 106 from Group B. Most of subjects are native Cantonese or Mandarin speaker whose are English L2 learner. The demographics of group A and group B are shown in Table 1.

Native language		
Cantonese	17	72
Mandarin	0	27
English	1	5
Other	1	2

Table 1: Demographics of Group A and B.

6.2 Evaluation Design

The evaluation questions are asked to Group A and Group B are shown in Table 2. Subjects are asked to rate their opinion in a five-point Likert scale and write down their comments and feedback.

Questions for Group A
(1) The level of difficulty of this assignment.
(2) The length of the assignment.
(3) How useful is the assignment in helping you identify problems in your pronunciation?
(4) Does the assignment help you become more aware of your problems in pronunciation?
(5) Is the speed of them system acceptable?
(6) Is the system easy to use?
(7) What is your opinion of the design of this assignment?
(8) Would you like to do assignments using a computer system in the future?
(9) Would you like do this assignment again with a new set of materials?

Questions for Group B
(1) The system was easy to use.
(2) The system speed was acceptable.
(3) The model pronunciation was helpful.
(4) The feedback was helpful.
(5) You would like to use system again.

Table 2: Evaluation questions for two groups. Users provide answer based on a five-point Likert scale.

6.3 Evaluation Analysis for Group A

We receive a lot of positive comments from subjects. They appreciate the system, especially the designed prompts and corrective feedback. They feel that the assignment is worthwhile, and they spot some of their frequent mistakes. The statistical summary is shown in Table 3.

Comments are received from the subjects include:

(1) "*The software is either ultra sensitive in a narrow range or not sensitive enough in picking up certain syllabic sounds, for instance some of my 'z' sounds were perceived as 's' sounds and some of my end 'p' sounds as the 'p' in 'camp' wasn't picked.*"

(2) "*For the assignment we need longer texts for learners to learn the intonation (rhythmic patterns) of English.*"

(3) "*If the system is giving feedback to our performance in terms of the tone choice expected, it would be even better. Now, it seems to test consistency more than accuracy.*"

(4) "*For the minimal pair check, it may be even better if the words are put in a sentence / in a context since performance in connected speech may vary than reading isolated items."*

Questions	**Legend**		**Likert scale**				
			1	**2**	**3**	**4**	**5**
(1)	1- difficult	5- easy	0	1	16	1	1
(2)	1- short	5- long	0	0	10	9	0
(3)	1- not useful	5- useful	0	3	2	13	1
(4)	1- not useful	5- useful	0	3	2	13	1
(5)	1- slow	5- fast	0	0	18	1	0
(6)	1- difficult	5- easy	0	0	1	11	7
(7)	1- poor	5- good	0	0	8	11	0
(8)	1- no 3- neutral	5- yes	0		2		17
(9)	1- no 3- neutral	5- yes	0		5		14

Table 3: Evaluation Summary for Questions in Table 2 (Group A).

6.4 Evaluation Analysis for Group B

We list some representative examples of subject's comments in *italic* with our explanations as follow. The statistical summary is shown in Table 4.

(1) *"It will lag or even hang if too many people are using the system simultaneously"*
This can be explained by network instability and bugs in the system at that time. However, after we fixed the bugs and the network configuration, subjects are satisfied with the speed and the stability of the system.

(2) *"I don't know how to pronounce the diagnostic feedback. It's hard to imitate the correct pronunciation"*
Subjects fail to reproduce the canonical pronunciation since they do not know which part of the articulator should be used. As the visual-speech animator is not integrated to our system at that time, we hope the new visual-speech feedback showing the articulator

movement can help them indentify pronunciation problems from another angle.

(3) *"It does not work well, because some of the sounds pronounced are not recorded due to the background noise"*
There were around 20 subjects in a classroom doing the system evaluations simultaneously. Although we already provide high quality close-talking head-mount headsets (Sennheiser PC156) for recording, background noise is still inevitably captured and degrades the accuracy of recognition, if the voice of the subject is weak.

Questions	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
(1)	3	23	15	56	9
(2)	4	30	10	59	3
(3)	6	24	19	54	3
(4)	6	20	32	45	3
(5)	8	18	25	52	3

Table 4: Evaluation Summary for Questions in Table 2 (Group B).

6.5 Evaluation Analysis of Mispronunciation Detection

We have received 24,339 valid recordings containing 278,989 phonemes in total, of which 50,378 (18%) mispronounced phonemes are detected. The top seven mispronounced phonemes, which contribute to 65% of all the mispronunciations, are summarized in Table 5, as well as the corresponding most frequent mispronunciation for each phoneme.

Mispronounced Phoneme	Occurrence Percentage	Mispronounced as	
		Phoneme	Percentage
/ih/	8,050 (16%)	/iy/	86%
/z/	5,666 (11%)	/s/	98%
/n/	4,995 (10%)	/ng/	93%
/r/	4,124 (8%)	_	89%
/dh/	3,374 (7%)	/d/	78%
/ae/	3,288 (7%)	/eh/	64%
/er/	3,070 (6%)	/eh/	94%

Table 5: Top seven mispronounced phonemes, e.g., the mispronunciation occurrence percentage of /ih/ is around 16%, and 86% of the errors in /ih/ are mispronounced as /iy/.

From Table 5, we can see how negative language transfer effect affects English L2 learners. L2 learners may use a similar sound in their L1 to pronounce L2. For example, native Cantonese English L2 learners may pronounce voiced consonant (e.g. /z/) as voiceless consonants (e.g. /s/).

To summarize, subjects agree that our system is useful in helping learner identify errors and become more aware of their pronunciation problems. Subjects are satisfied with the user interface and the performance. The high percentage of willingness to use our system in the future strongly encourages our further research.

7. CONCLUSIONS AND FUTURE DIRECTIONS

This paper introduced an Internet-accessible CAPT system — Enunciate, from the high level architecture and the detailed components to the major interactive features as well as the statistics of users' evaluations during the pilot trials. Enunciate has improved the capability of mispronunciation detection in ASR and provided pre-defined lessons and self-practice section, which are suitable for teachers and students to use during in-class training as well as self-practicing by students.

We are trying to implement the feature that can simultaneously offer the reference and the learner's articulatory animation at the same time so that learners can benefit from the differentiated articulatory movement. Meanwhile, we are working on incorporating a suprasegmental component, so that learners can not only improve their segmental quality but practice some of the prosodic features including stress and intonation as well. A release on the Android mobile platform will be available soon and we will investigate the performance of our recognizer for mobile use.

8. ACKNOWLEDGEMENTS

The work has been supported by the CUHK Teaching Development Grant and the NSFC/RGC Joint Research Scheme (project no. N_CUHK 414/09).

9. REFERENCES

[1] T. Kawahara, H. Wang, Y. Tsubota, and M. Dantsuji, "English and Japanese CALL Systems Developed at Kyoto University," *Proc. of APSIPA*, 2010.

[2] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing Suprasegmental English Through Parroting," *Proc. of Speech Prosody*, Chicago, May 2010.

[3] W. Chen, J. Mostow, and G. Aist, "Exploiting Predictable Response Training to Improve Automatic Recognition of Children's Spoken Questions," *Proc. of ITS*, Pittsburgh, June 2010.

[4] Y. J. Lin, "Introduction to L Labs Inc.," NSC Research Project – In the *Summer Workshop Lecture* of A Study on the Next-Generation Automatic Speech Recognition, Taiwan, 31 August 2009.

[5] J. L. Lu, R. L. Wang, L. C. De Silva, Y. Gao, and J. Liu, "CASTLE: a Computer-Assisted Stress Teaching and Learning Environment for Learners of English as a Second Language," *Proc. of Interspeech*, 2010.

[6] EduSpeak®,
http://www.speechatsri.com/products/eduspeak.shtml

[7] D. W. Massaro, Y. Liu, T. H. Chen, and C. Perfetti, "A Multilingual Embodied Conversatrional Agent for Tutoring Speech and Language Learning," *Proc. of Interspeech*, 2006.

[8] H. Meng, W. K. Lo, A. M. Harrison, P. Lee, K. H. Wong, W. K. Leung, and F. Meng, "Development of Automatic Speech Recongition and Synthesis Technologies to Support Chinese Learners of English: The CUHK Experience," *Proc. of APSIPA*, 2010.

[9] WAMI Tookit, http://wami.csail.mit.edu

[10] Spring Framework, http://www.springsource.org

[11] Hibernate, http://www.hibernate.org

[12] XML-RPC, http://www.xmlrpc.com

[13] MySQL, http://www.mysql.com

[14] ATK, http://mi.eng.cam.ac.uk/research/dialogue/atk_home

[15] HTK, http://htk.eng.cam.ac.uk

[16] Flite, http://www.speech.cs.cmu.edu/flite

[17] W. K. Lo, S. Zhang, and H. Meng, "Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer Assisted Pronunciation Training System," *Proc. of Interspeech2010*, 2010.

[18] H. Meng, Y. Y. Lo, L. Wang and W. Y. Lau, "Deriving Salient Learners' Mispronunciations from Cross-Language Phonological Comparisons," *Proc. of ASRU*, Kyoto, Japan, 9-13 December 2007.

[19] K. H. Wong, W. K. Lo, and H. Meng, "Allophonic Variations in Visual Speech Synthesis for Corrective Feedback in CAPT," *Proc. of ICASSP*, 2011.