Development of an Articulatory Visual-Speech Synthesizer to Support Language Learning

Ka-Ho WONG, Wai-Kim LEUNG, Wai-Kit LO and Helen MENG
Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Hong Kong
{khwong, wkleung, wklo, hmmeng}@se.cuhk.edu.hk

***Abstract* — This paper presents a two-dimensional (2D) visual-speech synthesizer to support language learning. A visual-speech synthesizer animates the human articulators in synchronization with speech signals, e.g., output from a text-to-speech synthesizer. A visual-speech animation can offer a concrete illustration to the language learners on how to move and where to place the articulators when pronouncing a phoneme. We adopt a 2D vector-based viseme models and compiled a collection of visemes to cover the articulation of all English phonemes (42 visemes for the 44 English phonemes). Morphing between properly selected vector-based articulation images achieves articulatory animations. In this way, we have developed an articulatory visual speech synthesizer that can accept free-text input and synthesize articulatory dynamics in real-time. Evaluation involving 32 subjects based on "lip-reading" shows that they can identify the appropriate word(s) based on articulation animation alone nearly ~80% of the time**

Keywords - computer-assisted language learning, articulatory phonetics, visual-speech synthesizer, text-to-audiovisual systhesis

I. INTRODUCTION

There has been an increasing demand for resources in language learning. The number of language learners, especially second language (L2) English learners, is increasing at a very fast pace. It has been estimated that there are over 5 million English second language (ESL) learners in China and India alone [1]. This large demand will aggravate the existing shortage of language teachers. Computer-assisted pronunciation training (CAPT) technology offers a viable solution with automatic, computer-based training. CAPT also has the advantages that it provides round-the-clock service, with personalized lessons to reduce anxiety and support self-based training.

Pronunciation learning involves perceptual training and productive training. Perceptual training aims at enabling the learners to discriminate different sounds in a language. On the other hand, productive training helps facilitate learning of articulatory movements for correct pronunciation.

Conventional perceptual training and productive training rely on audio only. In particular, productive training usually takes a "listen-and-repeat" approach. However, there can be many possible reasons that a learner cannot pronounce accurately. A learner may in fact be unable to perceptually distinguish between the target sound and its mispronunciation version. Even if the learner notices the difference, he/she may not know how to produce the sound. For example, Chinese learners may have difficulty in perception and production of inter-dental fricative (e.g. /ð/) because it does not exist in their primary language. When they are informed of the existence of this phoneme, it is difficult for them to reproduce an inter-dental by hearing due to lack of perceptual and productive experience.

The "listen-and-repeat" approach can be enhanced to a "perceive-and-repeat" one. We can provide multimodal (both audio and visual) examples to the learner for them to imitate. A visual illustration of the articulator movements, together with the audio can be useful. This is especially important for learners with hearing impairment where the visual channel can enhance perception. The University of Iowa has made available the animation and video illustrating the articulation of the sounds of American English, German and Spanish. These visual materials include a recorded video of the lips of a speaker and a pre-built flash animation of the articulatory movement in the mid-sagittal plane [2]. Wik and Hjalmarsson [3] developed the Ville system which teaches Swedish. Ville has a virtual three-dimensional language teacher who can guide and provide feedback to the learners.

We have developed a system that can synthesize a synchronized animation of articulatory movement with the synthetic speech signals for any free-text input. We collected a set of two-dimensional (2D) viseme models that include 44 English phonemes, covering phonemes in both TIMITDict [4] and CMUDict[5]. We used the FreeTTS [6], an open-source text-to-speech synthesizer based on Flite [7, 8], to synthesize the speech signals, phoneme sequence and timing information for text input. The animation is created by morphing the visemes from the 2D viseme models. The system flow is illustrated in Figure 1.

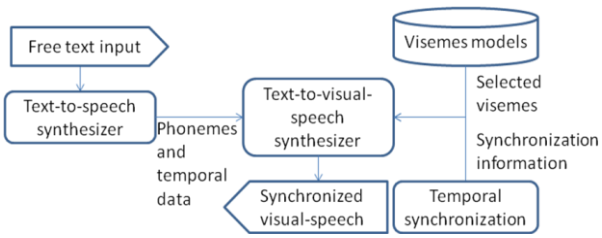

Figure 1: The flow chart illustrated the work of articulatory visual-speech synthesizer.

II. TWO-DIMENSIONAL (2D) VISEME MODEL

We adopted a set of 2D viseme models to represent the articulation of the English phonemes. We have 44 phonemes and 42 visemes. We obtain 42 scanned viseme images from Nilsen [9], which provides diagrams depicting the production of each sound based on American English. The scanned images cover 38 of the phonemes in our inventory. For the remaining 6, we share among existing models as summarized in Table 1. Based on the scanned images, we have devised 42 visemes in total (see Table 1).

Phonemes do not exist in Nilsen	We substitute by the models from
ɚ (dollar /d ɑ l ɚ/)	/ə/ and /ɝ/
ʌ (cut /k ʌ t/)	/ə/
l̩ (cattle /k æ t l̩/)	/ə/ and /l/
n̩ (certain /s ɝ t n̩/)	/ə/ and /n/
ɨ (spotted /s p ɑ t ɨ d/)	/ɪ/
ɝ (pure /p j ʊ ɝ/)	Designed manually

Table 1: The supplement of missing phonemes in Nilsen [6].

The scanned bitmap images are converted to vector-based images. They are manually enhanced to succinctly and clearly highlight the shapes and places of the articulators for pedagogical purposes [10]. Figure 2 shows an example viseme for the phoneme /g/ (**g**ap).

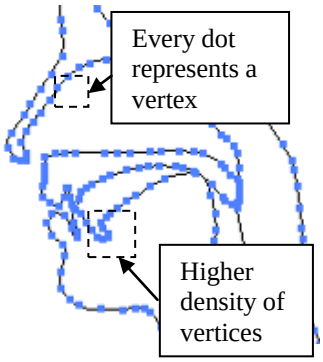

Figure 2: This is the viseme for /g/ (e.g., in /**g**ap/). The viseme is a vectorized image and the vertices are shown as dots here. We have intentionally assigned a higher density of vertices for the tongue for its flexibility in shape and dynamic movements during articulation.

III. VISEMES MODELS

Every viseme in the model is an ordered list of vertices. Every vertex (a data structure that describes a point in 2D) in a viseme has three attributes: coordinates of the vertex and two coordinates for the direction handles which are provided by Adobe Illustrator® . A direction handle controls the radian and shape of the line between two vertices as illustrated in Figure 3.

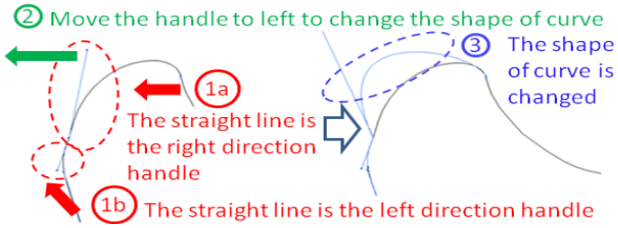

Figure 3: The line extending from the right-hand side of a vertex is right direction handle (1a). The line extending from the left-hand side of a vertex is left direction handle (1b). When we move the handle (2), the shape of curve between two vertices is changed (3).

We used 183 vertices for all visemes in the model. The model is represented as

$$V_j = \left\{\left(\mathbf{x}_{jk}, \mathbf{h}^l_{jk}, \mathbf{h}^r_{jk}\right)\right\} \quad (1)$$

where V_j is viseme *j;*

k is the index of the vertex ($1 \le k \le 183$),

x_{jk} is the coordinate for the *k*-th vertex,

h^l_{jk} is left direction handle vector for the *k*-th vertex and

h^r_{jk} is right direction handle vector for the k-th vertex

Since we currently use a 2D viseme model, each of these vectors contains the x- and y- coordinates, making up a total of 6 values for every vertex. The detailed mapping between the visemes and phonemes is described in the following section.

IV. PHONEME-TO-VISEME MAPPING

We employed a many-to-many mapping between the set of English phonemes and visemes in order to cater for the differences in articulation complexity of the phonemes. In general, a viseme in the model can be shared by several phonemes (e.g., voiced and voiceless counterparts). Phonemes with more complicated articulation (e.g., diphthongs) are allocated with more visemes.

1) Diphthongs

Figure 4 shows the 2 visemes for the diphthong /aʊ/.

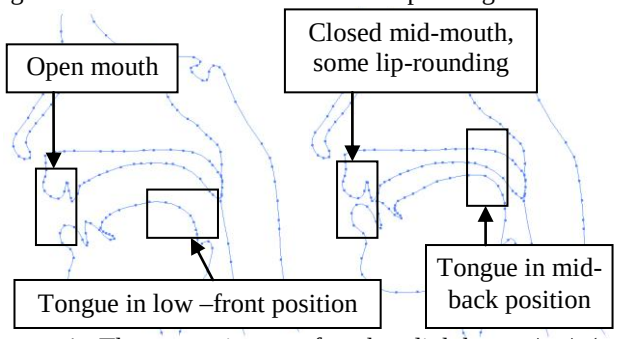

Figure 4: The two visemes for the diphthong /aʊ/ (e.g., b**ou**gh). The left viseme shows an open mouth articulation for the /ɑ/ (c**a**r) in /aʊ/. The right viseme shows the tongue in a mid, back position with some lip-rounding (/ʊ/ (b**oo**k) in /aʊ/).

2) Bilabials

Some phonemes, such as the bilabial plosives /b/ (**b**ig) and /p/ (**p**ut), have quick lip motions involving a closure and a burst. We use two visemes to represent them. Taking the articulation animation of /b/ as an example, there are two stages in the articulation. The first stage is mouth closing while keeping the lip shape normal until the lips touch (mouth closed). The second stage is the pressing of the lips against each other. The two stages should be done sequentially. Since there are two distinct actions, two visemes are used. The first viseme is when the lips just touch each other. The second viseme will be the lips pressing against each other firmly and hence deformed a little. Figure 5 illustrates the animation of /b/.

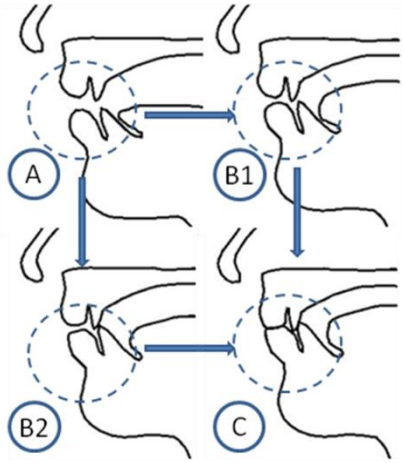

Figure 5: The sequence of visemes for the phoneme /b/ (as in **b**ig). In the animation of /b/, from viseme ***A*** transits to ***C*** via ***B1***. In this path, we will see the mouth close before the lips are deformed, which is more realistic. However, if the animation changes directly from viseme ***A*** to ***C*** (i.e. with only the single viseme ***C*** being used), the actions at mouth closure and lip deformation will happen simultaneously and will result in an abnormal transition such as ***B2***, which shows an intermediate transition where the lips are deformed before touching.

3) Others phonemes

For the remaining phonemes such as fricatives (e.g. /f/ (**f**ork)), we can use a single viseme, V_f. But to maintain consistency with other phonemes for morphing, we simple duplicate the viseme to make a pair. Hence all phonemes are represented in 2 visemes. Table 2 shows the summary of all mappings.

V. Articulation Motion Synthesis

The articulation motion in our visual-speech synthesizer is obtained by morphing the visemes in the 2D viseme model with reference to the input sequence of phonemes with timing information.

A. Morphing Between Two Visemes

We apply a simple blending process in the visual-speech synthesizer as shown by Equation (2). This process is based on the weighted morphing technique [11].

Phonemes		Number of distinct visemes
Diphthong	/l̩/, /n̩/, /i/, /eɪ/, /u/, /aɪ/, /aʊ/, /oɪ/, /ɚ/, /oʊ/, /ʒ/, /dʒ/, /j/, /h/	2
Bilabial	/p/, /b/, /m/	2
Others	/ɑ/, /æ/, /ʧ/, /ɔ/, /d/, /ð/, /ɛ/, /ʌ/, /ə/, /f/, /g/, /ɪ/, /i/, /k/, /l/, /n/, /ŋ/, /r/, /ɝ/, /s/, /ʃ/, /t/, /θ/, /ʊ/, /v/, /w/, /z/	1

Table 2: Phoneme represented one or two visemes. The two phonemes (/i/, /u/) are slightly diphthongized.

$$V(t) = V_a \times w(t) + V_b \times \big(1 - w(t)\big)$$
$$w(t) = \frac{(d-t)}{d} \qquad (2)$$

where $V(t)$ is the morphing result at time t,

$w(t)$ is the blending weight at time t and

d is the duration from viseme V_a to viseme V_b

Viseme morphing is realized by changing the blending weight from zero to one. Such blending is applied to all the three attributes for all vertices in the visemes. Finally, the transitional images in the animations are rasterized using the blended vertices.

B. Synchronization between audio and video

We control the blending process in order to achieve proper synchronization between the audio and video. We make use of the FreeTTS (the Java version of Flite) 1.2 [6] based on The CMUDict [5] to generate speech signals together with the timing information for each of the phonemes. For example, when the input is "map", the output from the TTS is: /m/ (duration: 0.07865931s), /æ/ (duration: 0.21247324s) and /p/ (duration: 0.15425742s). With this information, we can control the precise instances of appearance of visemes in the animation and also compute the duration of the phonemes for morphing.

1) Blending in general phonemes

To achieve more realistic and smooth animation, we reserve a short period of transition time empirically (e.g., 20% of the phoneme duration) for articulator transitions both at the beginning and at the end of a phoneme. This is necessary because in reality, no matter how fast the articulation, each articulator movement takes finite duration. For example the phoneme /ɔ/ with duration of d, we have assigned two visemes, V_1 and V_2. Let w_1 and w_2 be the respective blending weights. The blending weights w_1 and w_2 reach 1 at the instances $0.2d$ and $0.8d$, respectively.

2) Blending in plosives

In plosives, closure is needed before the actual sounding of the phoneme, we change the instances discussed above to $0.05d$ and $0.2d$ respectively. For example the phoneme /p/ with duration of d, we have assigned two visemes, V_1 and V_2. Let w_1 and w_2 be the respective blending weights. The blending weights w_1 and w_2 reach 1 at the instances $0.05d$ and $0.2d$, respectively.

3) Blending in consonants

For consonants other than plosives, we have also made similar allowances for articulator movements. Considering the faster movement of consonants, we set the values to be 0.05*d* and 0.8*d*. For the example /*m*/ (**m**at), we close the lips very quickly at the beginning and then keep them closed until near the end of the phoneme. Therefore, w_1 reaches 1 at 0.05*d* and w_2 reaches one 0.8*d*. Table 3 summarizes the assignment of transition times to different classes of phonemes.

	Blending weight of V_1 equals 1	Blending weight of V_2 equals 1
General phonemes	0.2**d*	0.8**d*
Plosives (/b/, /p/, /d/, /t/, /g/, /k/)	0.05**d*	0.2**d*
Consonants (/m/, /n/, /ŋ/, /f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, /ʧ/, /l/, /w/, /r/, /j/, /h/)	0.05**d*	0.8**d*

Table 3: Blending weights, w_1 and w_2, corresponding to two visemes V_1 and V_2 for a phoneme with duration *d*. We show the relative duration where the weights reach 1.

VI. EVALUATION

We evaluate the visual-speech synthesizer outputs by a subjective user test. The test consists of two parts: Part I tests with still images and Part II tests with animation of articulation.

1) Part I: Testing for the English phonemes knowledge

Part I aims to test (i) a subject's background knowledge of articulation based on the still image, whether they can tell what is being pronounced, (ii) and, whether they can discriminate between possible confusions with the articulation of other phonemes. We only select consonants to cover various articulatory motions which can be distinguished by the cross-sectional (mid-sagittal) view. A total of 16 articulation images are selected. For each image, subjects need to choose a phoneme from a list of 4 to 5 options (in IPA with sample words). The options are specifically designed to cover different combinations of place and manner of articulations. For example, given the image of the phoneme /b/, the options test the knowledge of two features, labial and non-nasal, and other combinations (Figure 6). The subjects are also instructed to select "*I don't know*" if they are uncertain.

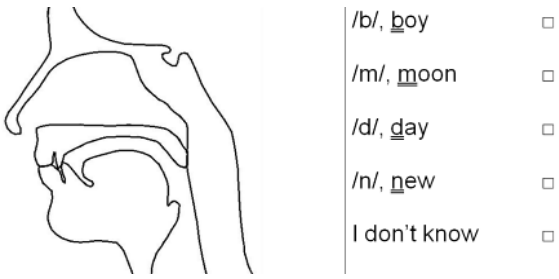

Figure 6: The articulation image of the phoneme /b/ (**b**oy) is shown to subjects with five options. The possibly confusing articulations are labial vs. alveolar, nasal vs. non-nasal.

The list of tested phonemes include: /b/, /d/, /g/, /z/, /r/, /ð/, /f/, /w/, /n/, /l/, /ŋ/, /m/, /j/, /ʃ/, /h/ and /ʧ/. All of them have distinct cross-sectional views. Since a normal language learner may not be aware of the velar flap, we have a briefing about the difference between /b/ and m/ during the introduction. We invited 32 subjects to the test. The average correctness in the part I is 60.4% (excluding "*I don't know*" which occupied 15.6% of answers). The correctness in distinguishing different articulatory motions is shown in Figure 7. Key observations include:

- *Perfect discrimination is achieved for four pairs of phonemes : /θ/ vs /d/, /g/ vs /m/, /b/ vs /n/ and /f/ vs /d/*
- *Confusion between retroflex and lateral (namely /r/ and /l/) is 59.6%*
- *Unfamiliar phonemes:*34% of subjects answered *"I don't know"* in the questions asking /j/ (**y**ellow) (the best is 3% /θ/).

In general, articulation pairs toward the right have higher correctness which implies that it can be helpful to visual-speech to offer productive training to language learners. Towards the left hand side, lower correctness for these pairs indicates that the learners may not be aware of the differences in articulation and more instruction is necessary.

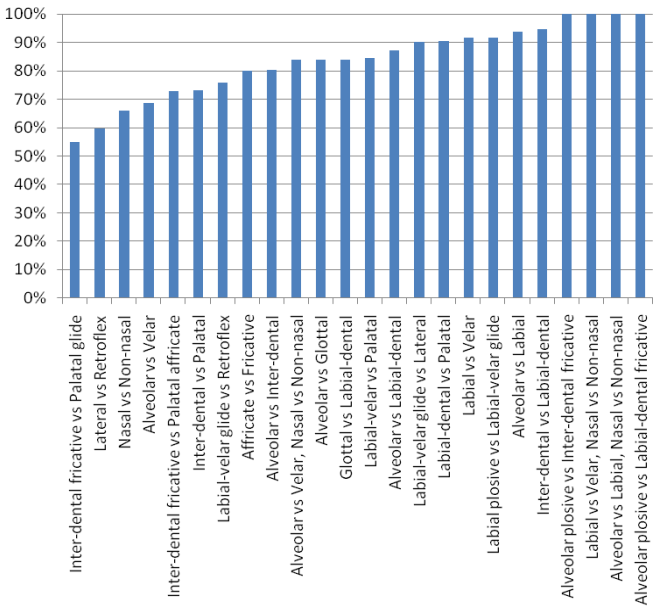

Figure 7: The correctness of each articulatory motion excluding the answer "*I don't know*".

2) Part II: Perception of articulatory animations

The purpose of the test in part II is to find out whether a learner can distinguish between a minimal pair of words based on articulatory animations. We presented 36 animations and asked the subjects to discriminate between minimal word pair, or choose "*I don't know*". These word pairs are selected according to phonological rules [12] [13] used in our system which captures as well as capturing common mispronunciations of Chinese learners of English, with reference to [9]. For example, to test the difference in articulation with or without an alveolar plosive deletion, we selected two pairs of words:

1. *bombed /bamd/* and *bomb /bam/* (play video without /d/ deletion)
2. *bids /bɪdz/* and *biz /bɪz/* (play video with /d/ deletion)

As shown in these examples, for each articulatory gesture (e.g. with or without /d/), two prompts, each in a word pairs are used to test it. The 2 prompts are separated for apart during the test. 36 prompts cover 18 articulatory gesture.

Every video is shown to the subjects three times in slow motion (around 50wpm, where the normal rate is 150wpm). Our test results show that our subjects achieve 79.8% correctness on average (excluding the answer of "*I don't know"*). A summary is given in Table 4.

In general, prompts which have low correctness require more exaggerated expressions in animation. More instructions from the teacher may be needed for prompts with a high percentage of subjects choosing of "*I don't know".* As for prompts with a high percentage of correctness and a low percentage of subjects choosing "*I don't know*", visual-speech may be a useful illustration to instruct the learners of proper articulations.

	Correct-ness	*I don't know*
Post-alveolar fricative vs Post-alveolar affricate substitution	55.7%	**25.0%**
Retroflex vs Labial-velar approximant substitution	70.3%	10.9%
Velar nasal vs Alveolar nasal substitution	71.0%	**20.3%**
Retroflex vs Alveolar lateral substitution	71.3%	12.5%
Velar plosive deletion	76.1%	**18.8%**
Alveolar nasal vs Alveolar lateral substitution	76.4%	7.8%
Labial-velar approximant vs Palatal approximant substitution	79.3%	9.4%
Labial plosive deletion	79.5%	8.6%
Retroflex deletion	80.0%	10.9%
Alveolar fricative deletion	83.3%	10.9%
Glottal fricative vs Alveolar fricative substitution	84.0%	12.5%
Alveolar plosive deletion	84.1%	10.9%
Voiced Inter-dental fricative vs Voiced unaspirated alveolar plosive substitution	84.3%	7.8%
Retroflex vowel vs Non-retroflex vowel substitution	85.6%	14.1%
Voiced labial-dental fricative vs Labial-velar approximant substitution	90.0%	6.3%
Voiceless inter-dental fricative vs Voiceless labial-dental fricative substitution	91.9%	4.7%
Voiceless inter-dental fricative vs Voiceless alveolar fricative substitution	93.6%	1.6%

Table 4: Results of the test on perception of visual-speech. The second column shows the percentage of correct choices (excluding those chosen *"I don't know*") among different substitution and deletion of phonemes. The third column shows the percentage of subjects chosing "*I don't know"* with each type of tested articulation contrast and the top three values are boldfaced.

VII. Conclusions and Future Work

We have developed a visual-speech synthesizer which explicitly shows the motion of lips, tongue and the opening of nasal passage to support language learning. We adopted vector-based mid-sagittal representation for the visemes which provides greater flexibility for modification to exaggerate the articulation when needed. We applied a simple blending technique for morphing to generate animation of articulators for free text input. We also designed detailed temporal mapping between visemes and phonemes in order to achieve audio-visual synchronization. Subjective evaluation shows that learners can distinguish either the substitution or deletion of articulation in 79.8% of the time (excluding those chosen *"I don't know*"). This initial attempt encourages us to further enhance the presentation by including the frontal lips view to the visual-speech synthesizer for those visemes which is similar in the mid-sagittal view. Another direction is to enhance the animation for co-articulation effects and allophonic variations of phonemes (e.g. the /l/ in **l**eaf and fee**l**, /r/ in **r**eed and dee**r** [10]).

Acknowledgment

The work has been partially supported by the National Natural Science Foundation of China (60928005).

References

[1] B. B. Kachru, Asian Englishes: Beyond the Canon. Hong Kong: Hong Kong University Press, 2005.

[2] Phonetics: The Sounds of Spoken Language, http://www.uiowa.edu/~acadtech/phonetics/.

[3] P. Wik and A. Hjalmarsson, "Embodied conversational agents in computer assisted language learning," Speech communication, Volume 51, Issue 10, pp. 1024-1037, October 2009.

[4] TIMITDict, http://www.ldc.upenn.edu/Catalog/docs/LDC93S1/TIMITDIC.TXT.

[5] CMUDict, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[6] FreeTTS 1.2, http://freetts.sourceforge.net/docs/index.html.

[7] Flite, http://www.cmuflite.org/.

[8] Festival, http://www.cstr.ed.ac.uk/projects/festival/.

[9] D. L. F. Nilsen and A. P. Nilsen, "Pronunciation Contrasts in English," 1973.

[10] P. Ladefoged, "A Course in Phonetics", 2006.

[11] J. Q. Wang, K. H. Wong, P. A. Heng, H. Meng and T. T. Wong, "A Real-Time Cantonese Text-to-Audiovisual Speech Synthesizer," in the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Fairmont Queen Elizabeth Hotel, Montreal, Quebec, Canada, 17-21 May 2004.

[12] H. Meng, Y. Y. Lo, L. Wang and W. Y. Lau, "Deriving Salient Learners' Mispronunciations from Cross-Language Phonological Comparisons, " in the Proceedings of Automatic Speech Recognition and Understanding (ASRU), Kyoto, Japan, 9-13 December 2007.

[13] A. M. Harrison, W. Y. Lau, H. Meng and L. Wang, "Improving Mispronunciation Detection and Diagnosis of Learners' Speech with Context-sensitive Phonological Rules based on Language Transfer," in the Proceedings of Interspeech, Brisbane, Australia, 22-26 September 2008.