Development of Automatic Speech Recognition and Synthesis Technologies to Support Chinese Learners of English: The CUHK Experience

Helen Meng, Wai-Kit Lo, Alissa M. Harrison, Pauline Lee,
Ka-Ho Wong, Wai-Kim Leung and Fanbo Meng
The Chinese University of Hong Kong, Hong Kong SAR, China
E-mail: {hmmeng, wklo, alissa, khwong, wkleung, fbmeng}@se.cuhk.edu.hk Tel: +852-26098327

Abstract — **This paper presents our group's ongoing research in the area of computer-aided pronunciation training (CAPT) for Chinese learners of English. Our goal is to develop technologies in automatic speech recognition (ASR) to support productive training for learners. We focus on modeling possible errors due to negative transfer from L1 (i.e. Chinese) of Chinese learners to L2 (English). ASR techniques are used for fine phonetic analyses to achieve mispronunciation detection and diagnoses. Learners' inputs that are grossly erroneous are handled by an utterance rejection technique. We also present initial work in the development of audio and visual speech synthesis to provide corrective feedback for learners. This paper presents an overview of the technologies, related experimental results and ongoing work as well as future plans.**

I. INTRODUCTION

In recent years, our group at CUHK has been engaged in the development of speech technologies to support computer-aided pronunciation training (CAPT), especially for Chinese learners of English. English is the *lingua franca* of our world. It has been estimated [1] that by 2010 there will be 2 billion English learners worldwide and the proportion in Asia alone will exceed the number of native speakers. With such a huge demand, there is an acute shortage of qualified teachers. Computer-assisted language learning (CALL), including CAPT applications, can supplement existing learning resources and provide unique benefits to the learner in terms of accessibility, reduced anxiety and individualized instructions. Our work focuses on the language pair of Chinese L1 (primary language) and English L2 (secondary language), due to local relevance, as bilingual competence has long been a competitive edge that underpins Hong Kong as an international Chinese city.

Pronunciation training involves correct perception and production of sounds in the target language. The learning process tends to be influenced by the transfer of L1 features in L2 productions. Negative transfer leads to pronunciation inaccuracies and errors in the second language. These inaccuracies tend to fossilize with age and present specific challenges to adult L2 learners. Chinese has stark linguistic contrasts in comparison with English. We often observe negative transfer effects in L2 English productions of L1 Chinese learners. Pronunciation improvement requires persistent practice in productive and perceptual training. In order to support productive training (i.e. eliciting speech from the learner for analysis), we have been developing automatic speech recognition techniques that enable detection and diagnoses of targeted pronunciation inaccuracies (i.e. mispronunciations) due to negative language transfer effects [2-7]. In order to support perceptual training (i.e. developing the learners' skills to accurately discriminate among sounds of the target language), we have begun to develop automatic response generation that provides multimodal visualization (i.e. through text-to-audiovisual speech synthesis) of the speech production process. The generated responses are intended as helpful instructions that guide error correction and improvement.

The rest of this paper is organized as follows: Section II presents a brief description of previous work. Section III describes our effort in L2 English speech corpora collection. Sections IV to VII describes our linguistically-motivated approach for developing automatic speech recognition (ASR) technologies that target mispronunciations due to negative language transfer, as well as a fusion with the conventional pronunciation scoring approach. Section VIII presents a pre-filtering technique for grossly erroneous inputs. Sections IX and X discusses the use of text-to-audiovisual synthesis for corrective feedback generation. Finally, Section XI discusses our ongoing and future research plans.

II. PREVIOUS WORK

The field of CALL is flourishing with uses of speech analysis, speech recognition, speech synthesis, language understanding and generation, machine translation and dialog modeling technologies to assist learning and assessment in the areas of pronunciation, vocabulary, grammar, comprehension as well as overall fluency. A sample of recent works may be found in [8-29]. Previous and ongoing work covers a variety of L1 and L2. English remains the most popular L2 language due to international usage. Research efforts have studied English speech as the "interlanguage" of secondary language learners who have not acquired native-like proficiency. The

interlanguages came from native speakers of Chinese (Cantonese and Mandarin) [14][30][31], German [32], Hindi and Bengali [33], Italian [22][32], Japanese [34] and many other languages. Examples of previous initiatives include the ISLE (Interactive Spoken Language Education) project funded by the European Commission to develop pronunciation training technologies for Italian and German learners of English; a project in the Cambridge-MIT Institute on Mandarin Chinese learning by native speakers of US or UK English; as well as DARPA's DARWARS program that includes teaching Arabic for military training [35]. The International Speech Communication Association Special Interest Group on speech and language technology in education (ISCA SIGSLaTE)[1] has been established since Fall of 2007.

III. Design and Collection of the Chinese Learners of English Corpus

We have designed and collected the Chinese University CHinese Learners Of English Corpus (CU-CHLOE) to support our research and development in L2 language learning.

The corpus is collected as prompted speech collected in a quiet room. A head-mount close-talking microphone (Sennheisser PC155/PC156) is used for recording. We developed a Windows-based computer program for the recording process. The text prompts are presented individually to the speaker (i.e. subject). The speaker is allowed to control his/her recording pace by the press-to-start and press-to-stop buttons. This approach also saves us from post-processing efforts in sentence segmentation. An illustration is provided in Figure 1. This recording tool has been made available free-of-charge with source code to members of the AESOP (Asian English Speech cOrpus Project) consortium.

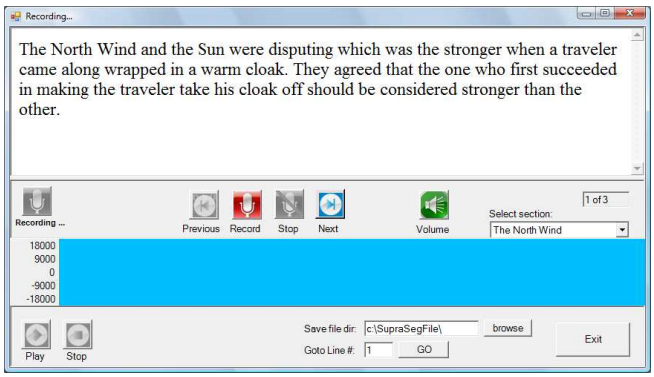

Figure 1. User interface of the recording tool used in the collection of the CU-CHLOE corpus

The prompts of CU-CHLOE focus on capturing L2 learners' pronunciation variants in the context of a word list, sentences as well as paragraphs. The aim is to gather speech data for deriving salient mispronunciations made by Chinese learners of English. Table 1 summarizes the design of the reading materials in this corpus.

The corpus includes data from 100 (50 male + 50 female) Cantonese subjects and 111 (61 male + 50 female) Mandarin subjects. The recordings are phonetically transcribed by experienced linguists.

Table 1. Summary of reading materials in different parts of the CU-CHLOE corpus.

The North Wind and The Sun (AESOP's Fable)	
Quantity	1 paragraph with 6 sentences
Rationale	Provides rich phonetic coverage of the English phonemes.
Example	"The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak." …
Phonemic sounds	
Quantity	20 sentences
Rationale	Include sentences specially designed by experienced English teachers to cover common English mispronunciations by Chinese students.
Example	"These ships take cars across the river."
Confusable words	
Quantity	10 lists
Rationale	Include lists of frequently mispronounced words by Chinese students. Summarized by experienced English teachers.
Example	"debt doubt dubious"
Minimal pairs	
Quantity	50 lists
Rationale	Include lists of words that are similar in pronunciations.
Example	"look full pull foot book"

IV. Capturing Language Transfer Effects through Contrastive Phonological Analyses

We believe in pronunciation training, the exact pairing of L1 and L2 is important. Hence we devise an approach for mispronunciation prediction, which is based on *contrastive phonological analysis* between L1-L2 pair. Contrastive analysis is grounded in the theory of language transfer. The Contrastive Analysis Hypothesis [36] states that sounds similar to the learner's first language will be easy for the learner to acquire while different sounds will present difficulty [15]. We conduct a contrastive analysis of Cantonese and English by examining the phonetic inventory and phonotactic constraints of the languages to determine phones and phone sequences present in English but lacking in Cantonese. Those phones which are not present in Cantonese are hypothesized to be substituted by Cantonese learners with phonetically-similar phones that do exist in Cantonese. Furthermore, [37] explicitly recognized the importance of examining actual errors within a corpus of data. Figure 2 presents and overview of our approach. We perform contrastive phonological analyses in order to *identify* phones that are susceptible to mispronunciations by L2 learners. We then perform error analyses based on field data (L2 speech recordings) to *capture* common errors on these phones. We then *summarize* these errors by deriving phonological rules for the observed errors.

[1] http://www.cs.cmu.edu/~max/mainpage_files/index.html

These rules are then used for predicting pronunciations variants of L2 learners.

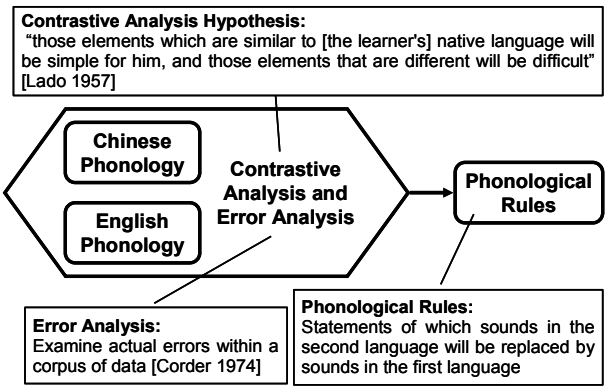

Figure 2. Overview of the proposed approach for mispronunciation prediction, which aims to focus on language transfer effects.

	Labial	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Plosives	p pʰ		t tʰ			k kʰ kʷ kʷʰ	
Affricates			ts tsʰ				
Nasals	m		n			ŋ	
Fricatives	f		s				h
Approximants					j	w	
Laterals			l				

Table 2. Consonants in Cantonese, organized according to the manner and place of articulation [38].

	Labial	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Plosives	p b		t d			k g	
Affricates				ʧ ʤ			
Nasals	m		n			ŋ	
Fricatives	f v	θ ð	s z	ʃ ʒ			h
Approximants				ɹ	j	w	
Laterals			l				

Table 3. Consonants in American English [36]. Consonants that do not exist in Cantonese are highlighted.

As a brief illustration, we present the consonant tables of Cantonese Chinese and American English in Table 2 and 3 respectively. The phonemes colored in Table 3 represent those that are present in English but absent from Chinese. We observe that there are two English dental fricatives: /θ/ is voiceless and is often mispronounced as the voiceless Cantonese labiodental /f/; /ð/ is voiced and is often mispronounced as the voiceless unaspirated alveolar plosive in Cantonese /t/. Examples include:

- "three" /θ r i/ vs. "free" /f r i/
- "there" /ð ɛ r/ vs. "dare" /t ɛ r/

(where the correct pronunciation for "dare" should be /d ɛ r/)

To model such mispronunciations, we make use of context-sensitive phonological rules, of the format:

$$\varphi \rightarrow \psi \; / \; \lambda \_ \rho$$

which denotes that phone φ may be substituted by the phone ψ when it is preceded by the phone λ and followed by the phone ρ. Similarly, by including the null symbol ϵ, we can encode the phone insertion rule by $\epsilon \rightarrow \psi$ and phone deletion by $\varphi \rightarrow \epsilon$. Rules that commonly occur regardless of context may be applied in a context-free manner, as $\varphi \rightarrow \psi$. Hence, to model the mispronunciations above, we may use the phonological rules:

- */th/* → [*f*] (context-free)
- */dh/* → [*d*] / # _ */ae/* (where # denotes a word boundary)

Contrastive comparisons between Cantonese and English vowels, or between the phonological spaces between different Chinese dialects and English accents, can be conducted in a similar manner. Further details about contrastive phonological analyses can be found in [36].

V. Mispronunciation Prediction with Manually and Automatically Derived Phonological Rules

As mentioned in the previous section, we have made use of knowledge from language transfer to predict the possible mispronunciations made by L2 English learners. Such knowledge is encoded in phonological rules.

A. Manually written phonological rules

We first developed a list of 43 *context-insensitive rules* (i.e. $\varphi \rightarrow \psi$ where the phone φ in the canonical pronunciation may be pronounced as ψ by the learner). When each of these rules is applied as a rewrite rule on the canonical pronunciation, we can generate hypothesized pronunciation variants that may appear in the learners' speech. We find two significant problems with using context-insensitive rules to extend the canonical pronunciation dictionary with possible mispronunciations (hereafter referred as extended pronunciation dictionary, EPD). The dictionary grows exponentially and many pronunciations generated are rare or implausible in the learner's speech. For example, Cantonese does not have voiced stops (e.g. /b/, /d/, /g/) or consonant clusters (e.g. /s t r/) while English does. Cantonese learners may substitute voiceless counterparts (e.g. /p/, /t/, /k/) or delete consonants to cope with these difficult sounds. So our list must include rules like (1) */d/* → [*t*], (2) */d/* → ϵ and (3) */k/* → ϵ. Admittedly, we can see that rules (2) and (3) do not fully represent the knowledge gained from our contrastive analysis (i.e. deletion only occurs in consonant clusters). When these rules are applied to a word like 'could' /k uh d/, we generate pronunciation variants such as: /k uh t/, /uh d/, /uh/, etc. Note that while /k uh t/ is a plausible mispronunciation of 'could', the variants /uh d/ and /uh/ generated from (2) and (3) are so phonetically-distant from the canonical pronunciation of 'could' that they are considered implausible mispronunciations.

To reduce the number of implausible pronunciations in the extended pronunciation dictionary, context-sensitive rules were developed from the contrastive analysis. The list of context-

sensitive rules was compiled using the same list of context-insensitive rules but additionally specifying the phonetic environments that constrain its application. A total of 51 context-sensitive rules were developed using the immediate neighboring segments and symbols for various linguistic classes: C for consonants, V for vowels, F for fricatives, and # for word-boundaries. Basically, context-sensitive rules solve the problem of over-generating implausible variants by reconsidering the variants /uh d/ and /uh/ generated by context-insensitive rules. These variants were generated because context-insensitive rules had no representational means to specify that deletion of consonants should only occur in consonant clusters. Context-sensitive rules solve this problem by allowing us to specify a phonetic environment that must be satisfied for the rule to apply. Thus, the consonant deletion rule from the previous section can be rewritten as "*/d/ → ϵ / C _*" where the left-hand side specifies that /d/ must be preceded by a consonant in order for the rule to apply. Note that manually written context-sensitive rules allow the left or right context to be a wildcard (i.e. with no specific constraints). When these context-sensitive versions of the previous rules are used to generate variants for a word like 'could', we see that the conditions of the deletion rules are not satisfied and thus implausible variants like /uh d/ and /uh/ are not generated. Details of the manual process of phonological rule authoring may be found in [3].

B. Automatically derived phonological rules

Manually authoring phonological rules requires expertise in both the mother language and also the L2 being learned. This means that the feasible language pairs will be limited by the availability of such kind of experts. As an initial effort to search for an alternative solution to the manual process, we have proposed and investigated an automatic, data-driven phonological rule derivation approach [4]. This method makes use of phonetically transcribed L2 speech data, together with canonical pronunciations. Our approach is based on a few assumptions: (i) differences in the phonetic transcriptions and the canonical pronunciations are due to negative language transfer effects, (ii) other interferences such as misread prompts, unknown words, transcription errors, ambiguity due to multiple accented pronunciations, etc., do not dominate. The proposed approach is summarized as shown in Figure 3.

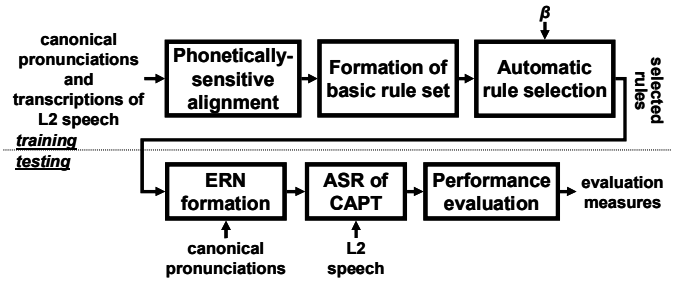

Figure 3. Schematic diagram of an automatic, data-driven derivation method for phonological rules in a CAPT system.

We make use of the Cantonese CU-CHLOE corpus in our investigation of automatic rule derivation. First, we aligned the canonical pronunciations with the manual transcriptions. From the aligned results, we can obtain a set of all phonetic substitutions, insertions, and deletions. This makes up the basic rule set. We then perform the rule selection process by keeping the top-N rules in the basic rule set and evaluate the coverage of the top-N rules by computing the F1-score. Figure 4 (a) shows the number of occurrences of particular rule and (b) depicts the computed F1-scores at different value of rules selected. The optimal (in the sense of F1-score) number of rules is found to be 216 for the Cantonese CU-CHLOE corpus.

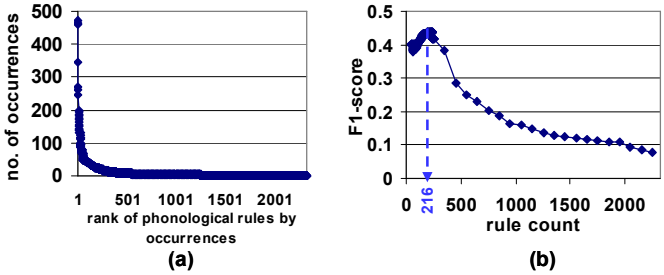

Figure 4. (a) Ranking of context-sensitive phonological rules (in the basic rule set) by occurrence counts. (b) F1-score based on the selected top-N rules (optimal N=216).

These phonological rules, whether manually authored or automatically derived, may be applied to the canonical pronunciations to obtain possible mispronunciations in L2 English speech.

VI. Mispronunciation Detection and Diagnoses

The phonological rules can be used for predicting pronunciation errors based on language transfer. This section presents our approach for mispronunciation detection and diagnoses.

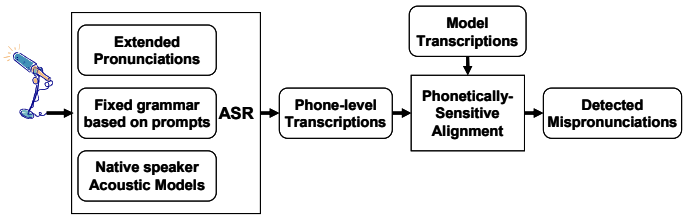

Figure 5. Overview of the ASR-based system to learners' detect and diagnose second language mispronunciations.

Figure 5 shows the basic idea of using automatic speech recognition (ASR) to detect mispronunciations with the use of the extended pronunciation dictionary, which includes the canonical pronunciation in a dictionary and the predicted mispronunciation(s) for the given word(s). For example the word "north" has the canonical pronunciation /n ao r th/. Application of the rule "*/r/ → ϵ / ao _*" generates the pronunciation /n ao th/ as an extension to the canonical pronunciation. The process is repeated for all rules to generate the extended pronunciation dictionary, see Figure 5 for an

illustration. The recognized phone sequences are then aligned with the canonical phone sequences. Phones that cannot be aligned properly can then be easily identified as deletions, insertions and substitutions. Furthermore, this approach can provide diagnostic feedback telling the learners what kinds of mistakes have been made and suggest the remedial actions.

A. Representation of Extended Pronunciations

Initially, we use the extended pronunciation dictionary with ASR, as shown in Figure 6 [3]. This approach is intuitively simple but algorithmic generation of the dictionary via exhaustive search-and-replace is inefficient and leads to more redundancy in the recognition network.. We then devise the Extended Recognition Network (ERN) (see Figure 7) [5] as a compact representation of the same information. Meanwhile, we also make use of the finite state transducer [39][40] as a vehicle to represent the rules (see Figure 8 for an example of r-deletion). This method is computationally more efficient as the application of phonological rules is a simple composition of finite state transducers and redundant paths can be easily removed before using for ASR.

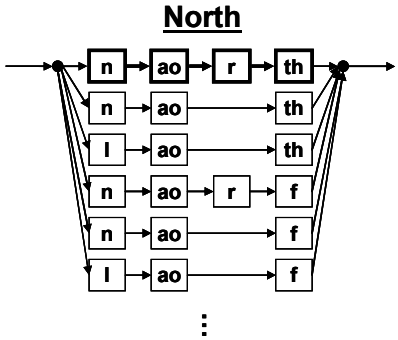

Figure 6. An example of the EPD for the word "north" with canonical pronunciation /n ao r th/. The prediction pronunciation variants are listed as optional paths.

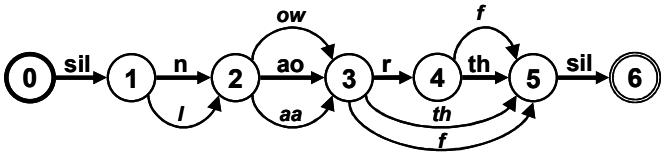

Figure 7. With the ERN, the ASR can detect whether the learner pronounced accurately (the middle path) or made any predicted pronunciation mistakes.

With the extended canonical pronunciations represented as an ERN, we then perform forced-alignment of the learner's speech by using the ERN as the recognition grammar. As a result, the canonical pronunciation will be obtained as the ASR output if the learner pronounces accurately. Otherwise, the ASR output will indicate exactly what kind of pronunciation mistakes have been made, by forced alignment between the recognizer's output and the canonical pronunciations. For example, if a pronunciation of /n ao th/ is detected, we can produce the diagnostic message for the learner, informing that "the retroflexion /r/ is deleted, the tongue needs to be curled back". The ability to generate diagnostic feedback is an advantage over the conventional approach of pronunciation scoring where only the evaluation score is provided for the learner.

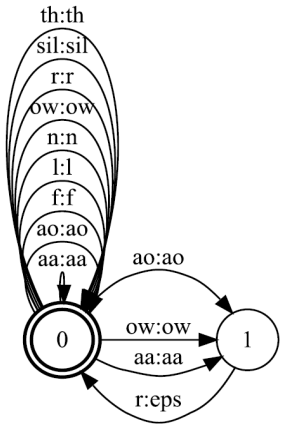

Figure 8. An example finite state transducer expressing /r/-deletion rule.

VII. Enhancing Mispronunciation Detection by Fusion with Pronunciation Scoring

The CAPT approach described in the previous section focuses on detection of salient mispronunciations related to language transfer phenomena. Hereafter, we refer to it as the linguistically-motivated approach. Not all the possible mispronunciations are predicted by the approach. This may be due to: (i) the expansion rule may be absent due to pruning or lack of relevant language transfer knowledge; (ii) the quality of the acoustic models is poor which hinders recognition accuracy; and (iii) the mispronunciations may be caused by factors other than language transfer (e.g. misread words, or incorrect letter-to-sound conversion). As a result, there are cases where the mispronunciations cannot be detected by the linguistically-motivated approach. In order to address this issue, we investigate a fusion technique to with conventional pronunciation scoring to enhance detection performance, as will be described in this section.

Conventional pronunciation scoring is based on the posterior probability of a speech unit being produced by the speaker, given the acoustic observations and the speech recognizer (including model pronunciation, acoustic models, etc.) as shown in Equation (1).

$$P\left(p|\,\vec{O},\,\Lambda\right)\approx\frac{ac\left(\vec{O}\,|p,\;\Lambda\right)}{\sum_{\forall p'\in P}ac\left(\vec{O}\,|p',\;\Lambda\right)}\qquad(1)$$

where ac is the acoustic likelihood score, p is the speech unit in focus, P is the set of all units (e.g. phoneme set), $\vec{O}$ is the acoustic observation, and Λ is the speech recognizer.

This posterior probability can be used directly as the pronunciation score [41][42], or further categorized into grades [43] (e.g., normative scale from 1 to 5). Scoring pronunciation in this way can leverage existing ASR technologies to offer a quantitative (or categorical) assessment for the users. In pronunciation training, the system may regard a speech unit to be mispronounced if its score falls below a pre-set threshold.

This method can, in principle, detect all possible mispronunciations.

The philosophy behind our fusion technique is simple and intuitive: take the decision that is more reliable in mispronunciation detection, with the objective of minimizing the total detection error (i.e. total number of false rejections and false acceptances) [4]. Hence the question to investigate is – how we may decide which of the two approaches, i.e. the linguistically-motivated approach or pronunciation scoring approach, is more reliable. To address this issue, we adopt the strategy of phone-dependent thresholds for decision making, and experimented with the Cantonese CU-CHLOE corpus, using half of the data for training and the remaining for testing. We first optimize individual thresholds for every English phone for pronunciation scoring. The detection performance of the linguistically-motivated approach and the phone-dependent pronunciation scoring is then evaluated over the training data to obtain a list of phones that is better handled by the pronunciation scoring approach. This forms our "backoff phone list" which is used in our fusion strategy, as illustrated in Figure 9.

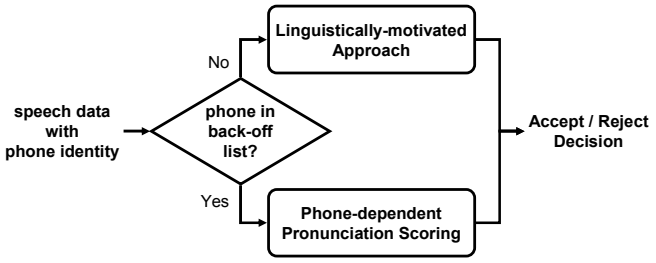

Figure 9. Fusion strategy that combines the linguistically-motivated approach with pronunciation scoring.

Experimental results based on the CU-CHLOE Cantonese dataset (the first time we publish such results) is shown in Figure 10. We observe that fusion brings significant improvement in detection performance over the individual approaches. To have a closer look at the detection performance, the total number of detection errors (TotErr) is depicted in Figure 11.

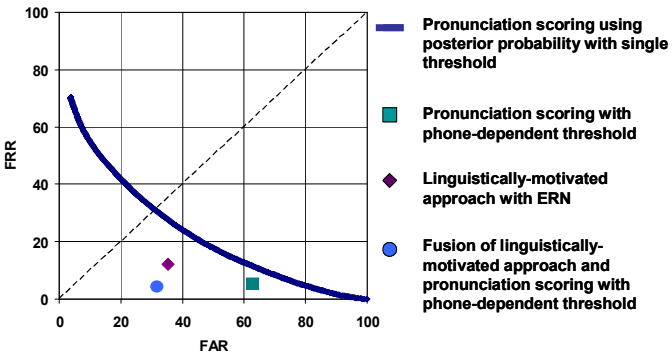

Figure 10. ROC for pronunciation detection performance using single threshold for all phones. The diamond is the operating point of the linguistically-motivated detection approach, the box is the operating point for phone-dependent pronunciation scoring and the circle is the operating point after fusion with phone-dependent pronunciation scoring.

We observe that fusion improves the mispronunciation detection performance by about 40% relative to the individual approaches. This fused approach may be used in a CAPT system in the following way: For phones that can be handled reliably by the linguistically-motivated approach, results in mispronunciation detection and diagnosis can both be provided for the learners. Otherwise, the CAPT can still return an appropriate mispronounced decision to the learners.

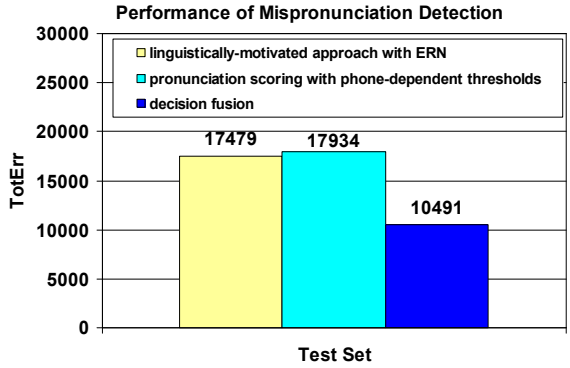

Figure 11. Comparison in mispronunciation detection performance among different approaches – the linguistically-motivated approach, pronunciation scoring with phone-dependent thresholds and the fusion of the two. (Remark: total no. of phones in the test set is 101,633).

VIII. Utterance Rejection for Pre-Filtering

Thus far we have presented an approach that uses ASR technologies to achieve fine phonetic analysis for mispronunciation detection and diagnoses. These technologies are intended for handling input utterances from learners based on text prompts in a CAPT system. However, learners' inputs that are grossly erroneous should be more appropriately handled by a pre-filtering mechanism. Anecdotal observations based on new users show that there are several common factors that may cause corruptions to the input utterances. For instance, there may be disfluencies (such as false starts, repairs, repetitions). Users may stop reading before completing the prompt text, due to distractions, side conversations, etc. The recording may also be truncated possibly due to the user pressing the <stop> button too early. These corrupted utterances should be handled differently by the system, as compared with an *intact* utterance whose spoken content corresponds well with the text prompt. More specifically, our system generates corrective feedback for an intact input utterance to inform the user of discovered phonetic errors. However, appropriate feedback for *non-intact* input should prompt the user to record again. Hence, there is strong motivation to develop a pre-filtering mechanism that separates the two types of utterances.

Confidence measures have been used in earlier work to verify that an input utterance has appropriate content for the speech application [44]. For example, a phone-dependent confidence measure is used for utterance rejection in [45]. In [46], the generalized word posterior probability is computed for each word and utterance rejection is performed based on a

combination of word scores. Phone duration has been used as feature for computing confidence measures in ASR applications for embedded and noisy environments [47][48] as well as verifying selected utterances in a language learning application [49]. In our work, the forced alignment nature of our CAPT approach (explained in the previous section) can exaggerate the phone duration variations in corrupted utterances. Hence we investigate the use of a statistical phone duration model to pre-filter for intact utterances that can subsequently undergo detailed phonetic analysis for mispronunciation detection and diagnosis.

It is known that phone durations vary across speakers and utterances. Phone durations have often been modeled statistically by the Gamma distribution [47][50][51]. We verify this based on the corpus statistics of the TIMIT corpus for native American English speakers and the CU-CHLOE for non-native speakers. The duration distributions of certain phones (especially consonants) tend to fit well with the exponential distribution – a special case of Gamma.

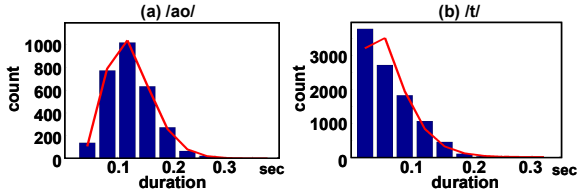

Figure 12. Histogram of the phone duration of native American English speaker for (a) the diphthong /ao/ and (b) the consonant /t/. The statistics are estimated based on the TIMIT corpus and the fitted Gamma distribution is also shown.

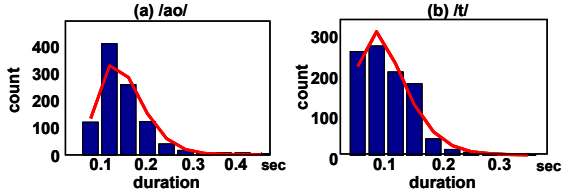

Figure 13. Histogram of the phone duration of native American English speaker for (a) the diphthong /ao/ and (b) the consonant /t/. The statistics are estimated based on the Cantonese CU-CHLOE corpus and the fitted Gamma distribution is also shown.

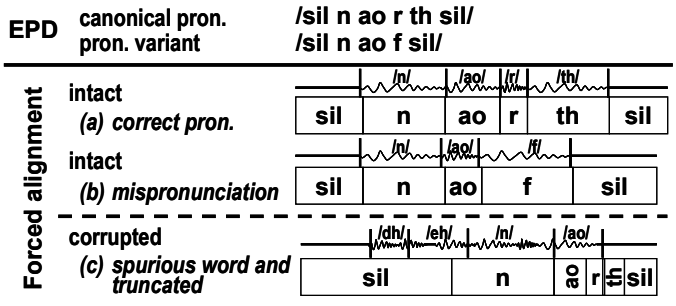

Figure 14. This figure illustrates forced alignment between an input utterance and the best-matching phone sequence from the extended pronunciation dictionary (EPD). Forced alignment produces reasonable phonetic durations for an intact utterance. On the other hand, phonetic durations of corrupted utterances tend to be overly long or short.

To design a filtering mechanism for intact utterances that are appropriate for subsequent phonetic analysis for mispronunciation detection and diagnosis, we assume that the phones in an intact utterance should largely carry their respective inherent durations. Our filtering approach references the durations obtained by the recognizer through forced alignment. As an illustration, Figure 14 shows the word "north" in a sentence which belongs to one of the system's text prompts. If the learner utters the text prompt with correct pronunciation as in (a), the phone durations should resemble their inherent values. In (b), the learner mispronounces the word and the best alignment selects the pronunciation variant that is among those predicted in the EPD. The phone durations in the forced alignment should also resemble their inherent values.

In (c), the input utterance (with a phone sequence of /dh eh n ao/) does *not* correspond in any way to the prompted text (that includes the word "north" with reference phoneme sequence /n ao r th/). Forced alignment makes the best effort possible to align the input utterance with one of the pronunciations in the EPD. This results in the frames of spurious phones (e.g. /dh/ and /eh/ that do not appear in the pronunciation of "north") being absorbed by the SILENCE segment or a non-silence phone segment(s) (e.g. /eh/ being absorbed into the /n/ segment). The latter causes lengthening of the absorbing phone segment. As for missing phones (e.g. /r/ and /th/ that occurs in the word "north" but are absent in the input utterance) in the EPD pronunciation that do not correspond to any acoustic frames, they tend to be assigned very short durations by the alignment algorithm. Hence, if forced alignment produces phone durations that are overly long or short, as compared with their inherent values, it may suggest that the input utterance is *not intact* and should not be subjected to further detailed phonetic analysis. As such, we can design a filtering approach based on phonetic durations to identify intact utterances that are analyzed phonetically for further mispronunciation detection and diagnoses.

In phone duration scoring, we incorporate an anti-model to increase the discriminative power of the phone duration model. A likelihood ration test is applied as shown in Equation (2):

$$\frac{1}{\|S\|}\log\left(\prod_{p\in S}\frac{P(dur(p)|\,p)}{P_{anti-model}(dur(p))}\right) \qquad (2)$$

where S is the set of phones in the utterance, $\|S\|$ is the number of phones in the utterance, $P(dur(p)|p)$ is the statistical duration model of phone p and we chose the Gamma distribution.

The statistical duration model of a phone p is obtained by estimating a Gamma model using the phone duration of all phone p in the CU-CHLOE corpus. For the anti-model, we have experimented with different techniques, including the "catch-all" anti-model. We first shuffle the utterances in the corpus such that the recordings will not be matching to the prompting texts. A forced-alignment is then performed using this intentionally shuffled prompts. A Gamma distribution is then trained using all aligned phone durations in the shuffled

corpus. The rationale is to obtain an anti-model for use in the likelihood ratio test where each phone duration model is trained with non-corresponding phonetic segments in the utterance. In our experiments with the different pre-filtering strategies, we consider the equal error rate (EER), i.e. the operating point where FAR=FRR (FAR: false acceptance rate, i.e. accepting a non-intact utterance for an intact one; false rejection rate, i.e. rejecting an intact utterance as a non-intact one). Our experiments show that the "catch-all" model gave the best detection performance at an EER of 17.16% [6].

IX. Synthesizing Expressive Speech to Convey Emphasis in Feedback Generation

This section, as well as the next one, presents our initial work in the development of speech synthesis technologies for corrective feedback generation in CAPT. We choose to begin with synthesizing *focus*. Words carrying focus in the synthesized response aims to draw the attention of the learner, highlighting the speech segments where correction is needed. For example, if the learner has trouble discriminating between the phones /th/ and /f/, as in the sentence, "Fighting *thirst* is the *first* thing to be done in this country," we aim to place focus on the words "thirst" and "first".

Focus is supported primarily by prosodic features, such as intensity, pitch changes and phone durations. We analyze the difference in these prosodic features between an utterance that carry neutral intonation and one that carry expressive intonation to convey focus. Our analysis involves the classification of phone in focus and neutral words, as follows:

For a focus word with a syllable carrying primary stress:

Class 1: Phones in the stressed syllable
Class 2: Phones before the stressed syllable
Class 3: Phones after the stressed syllable

For words without focus:

Class 4: Phones in the word before the focus word
Class 5: Phones in the word after the focus word
Class 6: All other (remaining) phones

Figure 15 illustrates this method of phone classification. "Peterson" and "occasion" are the focus words in the sentence.

I	have	met	PETERSON	on	one	OCCASION.
6	4	1	3	5	4	2 1 3

Figure 15. An example of phone classification based on the location of stressed syllables in focus words.

We also extract the following acoustic features from the phones to capture focus:

- maximum f0 (Max, in Hz),
- f0 range (R, in Hz),
- minimum f0 (Min, in Hz),
- mean f0 (Mean, in Hz),
- absolute value of f0 slope (S, in Hz/ms),
- mean of RMS energy (E, in dB), and
- duration per phone (D, in ms).

Thereafter, we develop a perturbation model for each phone class, based on the ratio of a feature in focus speech and its counterpart in neutral speech, as shown in Equation (3):

$$R = \frac{1}{n}\sum \frac{F_{i,focus}}{F_{ii,neu}} \quad (3)$$

where $F_{i,\ \mathrm{neu}}$ is the value of one feature for the ith phone in neutral speech, $F_{i\ ,\mathrm{focus}}$ is its counterpart in focus speech and n is the number of the phones in the particular phone class.

This perturbation model is used to modify neutral speech recordings to synthesize focus. Results from a listening test show that the 13 subjects can identify the focus words with an accuracy of over 98%. The perceived degree of focus in the identified words achieves a mean score of 4.5 in a five-point Likert scale. Details of this work can be found in [52].

X. Synthesizing Visual Speech with Articulator Animation

We believe that visualization of articulatory motions will also be effective for corrective feedback generation. There has been other visual animation work used in CAPT. For instance, Ville teaches Swedish with the help of an animated avatar [53]. The University of Iowa also offers a flash animation of articulators in a midsagittal view for English phones [54]. We take the idea one step further, with the aim to synthesize visually animated speech based on free-text input, which can be synchronized with synthesized audio to provide multimodal corrective feedback generation.

We collected the full set of visemes corresponding to English phones with reference to [55]. In our approach, multiple visemes can be mapped to multiple phonemes. Altogether, we have 42 visemes for 44 English phonemes. For a given text input, a text-to-speech synthesizer (FreeTTS [56], a Java interface to Flite [57]) is used to generate the audio and the time boundaries of the phones. Based on the phone identity and the corresponding time boundaries, we apply a blending approach [58][59] to synthesize the animated articulatory motions. Our preliminary visual perception test shows that when the subjects are asked to perform "articulator-reading" for a minimal word pair, they can correctly discriminate between the words about 75% of the time. Details are presented in [60].

XI. Ongoing Work and Future Directions

There are many areas of research that we are currently exploring with the aim to develop speech technologies for CAPT. We strive to improve the basic capability of mispronunciation detection in ASR. As mentioned earlier, conventional ASR technologies are developed for LVCSR task such as dictation. The objective is to return the correct words to the users even if the pronunciations are inaccurate. However, for a CAPT task, the requirement is more stringent that the ASR is now required to discern (i) how much the pronunciations in the utterance deviate from the reference? (ii)

Does the deviation constitute a mispronunciation? (iii) If so, how can we pinpoint the error? For example, the confusion between /th/ and /f/ in the word "thin" by a Chinese learner should have led to "thin" being spoken as "fin", but this may be rectified automatically by language modeling constraints in conventional ASR. Our approach described above uses the ERN to provide explicitly modeled mispronunciations to capture the error. However, this also places a very high requirement on the discriminative ability of the acoustic models. Hence, we are exploring the use of discriminatively trained acoustic models, with reference to predicted mispronunciations [60]. These models should bring about improvements in mispronunciation detection and diagnoses, as well as pronunciation scoring.

We have also expanded the scope of our investigation from segmental phonology to suprasegmental phonology. English and Chinese have stark contrasts in suprasegmental phonology. We focus on prosodic features that are critical for communication, namely, stress and intonation. Our work includes design and collection of an L2 suprasegmental corpus [61] (also under the AESOP umbrella), running perceptual tests to assess whether Chinese learners can perceive the relevant prosodic cues in English [62], as well as the development of technologies to detect and verify appropriate intonation in L2 speech. Additionally, we note that Chinese has syllable-timed rhythm while English has stress-timed rhythm. We are developing a technique in a prototype known as MusicSpeak, that can make use of synthesized musical rhythm to help Chinese learners acquire the appropriate rhythmic productions of English [63].

It is our hope to be able to incorporate all our technologies in our prototype system, in order for it to help teachers and students during in-class training, as well as students in self-directed learning and practicing.

ACKNOWLEDGMENTS

This paper presents an overview of our work in the past few years. The work has been supported by the CUHK Teaching Development Grant, funding from the Shun Hing Institute of Advanced Engineering, the NSFC/RGC Joint Research Scheme (project no. N CUHK 414/09) and Innovation and Technology Fund, Hong Kong SAR government (ITS/286/09). The synthesis work was conducted while Fanbo Meng from Tsinghua University was a visiting research intern at CUHK. We thank Professor Lianhong Cai and Dr. Zhiyong Wu of Tsinghua University for their helpful suggestions.

REFERENCES

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

[2] H. Meng, Y. Y. Lo, L. Wang and W. Y. Lau, "Deriving Salient Learners' Mispronunciations from Cross-Language Phonological Comparisons," *Proc. of ASRU2007*, 2007.

[3] 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," *Proc. of Interspeech2008*, 2008.

[4] W. K. Lo, A. M. Harrison, H. Meng, and L. Wang, "Decision Fusion for Improving Mispronunciation Detection using Language Transfer Knowledge and Phoneme-dependent Pronunciation Scoring," *Proc. of the 6th International Symposium on Chinese Spoken Language Processing*, 2008.

[5] A. M. Harrison, W. K. Lo, X-J Qiang, and H. Meng, "Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training," *Proc. of the 2nd ISCA Workshop on Speech and Language Technology in Education*, 2009.

[6] W. K. Lo, A. M. Harrison, and H. Meng, "Statistical Phone Duration Modeling to Filter for Intact Utterances in a Computer-Assisted Pronunciation Training System," *Proc. ICASSP2010*, 2010.

[7] 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.

[8] G. Kawai and K. Hirose, "A Call System Using Speech Recognition to Teach the Pronunciation of Japanese Tokushuhaku," *Proc. of STiLL1998*, pp. 73-76, 1998.

[9] T. Kawahara, *et al*., "Practical Use of English Pronunciation System for Japanese Students in the CALL Classroom," *Proc. of Interspeech2004*, pp. 1689-1692, 2004.

[10] J. Mostow, A. G. Hauptmann, L. L. Chase, and S. Roth, "Towards a Reading Coach that Listens: Automated Detection of Oral Reading Errors," *Proc. of the Eleventh National Conference on Artificial Intelligence (AAAI93)*, pp. 392-397, 1993.

[11] Ordinate Corporation, "The PhonePass Test," Menlo Park, CA, 1998.

[12] W. Byrne, E. Knodt, S. Khudanpur, and J. Bernstein, "Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English," *Proc. of STiLL1998*, pp. 37-40, 1998.

[13] S. M. Witt and S. J. Young, "Performance Measures For Phone-Level Pronunciation Teaching in CALL," *Proc. of STiLL1998*, pp. 99-102, 1998.

[14] B. Mak, *et al.*, "PLASER: Pronunciation Learning via Automatic Speech Recognition," *Proc. of HLT-NAACL*, pp. 23-29, 2003.

[15] J. Mostow, "Is ASR accurate enough for automated reading tutors, and how can we tell?" *Proc. of Interspeech2006*, pp. 837-840, 2006.

[16] Y. Tsubota, T. Kawahara, and M. Dantsuji, "Recognition and verification of English by Japanese students for computer assisted language learning system," *Proc. of ICSLP2002*, pp. 1205-1208, 2002.

[17] B. Townshend, J. Bernstein, O. Todic, and E. Warren, "Estimation of Spoken Language Proficiency," *Proc. of STiLL1998*, 1998.

[18] M. Eskenazi, "Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype," *Language Learning and Technology*, 1999.

[19] L. M. Tomokiyo, L. Wang, and M. Eskenazi, "An empirical study of the effectiveness of speech-recognition-based pronunciation training," *Proc. of ICSLP2000*, 2000.

[20] A. Neri, C. Cucchiarini, and H. Strik, "Effective Feedback on L2 Pronunciation in ASR-based CALL," *Proc. of the Workshop on Computer Assisted Language Learning, Artificial Intelligence in Education Conference*, 2001.

[21] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, "Modeling and Automatic Detection of English Sentence Stress

for Computer-Assisted English Prosody Learning System," *Proc. of ICSLP2002, 2002*.
[22] R. Delmonte, "Prosodic Modeling for Automatic Language Tutors," *Proc. of the STiLL1998*, 1998.
[23] B. Granström "Towards a Virtual Language Tutor," *Proc. of the InSTIL/ICALL Symposium on Computer Assisted Learning, NLP and Speech Technologies in Advanced Language Learning Systems*, 2004.
[24] L. Oppelstrup, M. Blomberg, and D. Elenius, "Scoring Children's Foreign Language Pronunciation," *Proc. of FONETIK*, 2005.
[25] J. Mostow, "Is ASR Accurate Enough for Automated Reading Tutors, and How Can We Tell?" *Proc. of Interspeech2006*, 2006.
[26] T. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation Verification of Children's Speech for Automatic Literacy Assessment," *Proc. of Interspeech2006*, 2006.
[27] S. Peterson, and M. Ostendorf, "Text Simplification for Language Learners: A Corpus Analysis," *Proc. of SLaTE2007*, 2007.
[28] A. Black, "Speech Synthesis for Educational Technology," *Proc. of SLaTE2007*, 2007.
[29] S. Seneff, "Web-based Dialog and Translation Games for Spoken Language Learning," *Keynote Speech and paper in Proc. of SLaTE2007*, 2007.
[30] W. Liang, J. Liu, and R. Liu, "Automatic Spoken English Test for Chinese Learners," *Proc. of Communications, Circuits and Systems*, 2005.
[31] S. Chevalier, "Speech Interaction with Saybot Player, a CALL Software to Help Chinese Learners of English," *Proc. of SLaTE2007*, 2007.
[32] T. Oba, and E. Atwell, "Using the HTK speech recogniser to anlayse prosody in a corpus of German spoken learner's English," *Proc. of International Conference on Corpus Linguistics*, 2003.
[33] A. Verma, K. Lal, Y. Y. Lo, and J. Basak, "Word Independent Model for Syllable Stress Evaluation," *Proc. of the ICASSP2006*, 2006.
[34] M. Muto, Y. Sagisaka, T. Naito, D. Maeki, A. Kondo, and K. Shirai, "Corpus-based Modeling of Naturalness Estimation in Timing Control of Non-native Speech," *Proc. of Eurospeech2003*, 2003.
[35] W. L. Johnson, S. Marsella, N. Mote, M. Si, H. Vilhjalmsson, and S. Wu, "Balanced Perception and Action in the Tactical Language Training System," *Proc.. of the International Conference on Autonomous Multiagent Systems*, 2004.
[36] R. Lado, *Linguistics Across Cultures: Applied Linguistics For Language Teachers*. University of Michigan Press. 1957.
[37] S. Corder, "Idiosyncratic Dialects and Error Analysis," *Error analysis: Perspectives on Second Language Acquisition*, pp. 158-171, Richards, J. (Ed.), Longman.
[38] E. Zee, "Chinese (Hong Kong Cantonese)," *Journal of International Phonetic Association*, 21(1), 1991.
[39] C. Allauzen, M. Riley, B. Harb, J. Schalkwyk, M. Mohri, R. Sproat, and W. Skut, "OpenFST v1.1," http://www.openfst.org, 2009.
[40] M. Mohri, F. C. N. Pereira, and M. Riley. "Weighted Finite-State Transducers in Speech Recognition," *Computer Speech and Language*, 16(1):69-88, 2002.
[41] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 30(2-3):95-108, 2000.
[42] L. Neumeyer, F. Horacio, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, 30:83-93, 2000.
[43] H. Franco, *et al.*, "Combination of Machine Scores for Automatic Grading of Pronunciation Quality," *Speech Communication*, 30:21-130, 2000.
[44] H. Jiang, "Confidence measures for speech recognition: a survey," *Speech Communication*, 45:455-470, 2005.
[45] M. Cohen *et al.*, "A phone-dependent confidence measure for utterance rejection," *Proc. of ICASSP1996*, pp. 515-518, 1996.
[46] W. K. Lo and F. Soong, "Generalized posterior probability for minimum error verification of recognized sentences," *Proc. of ICASSP2005*, pp. 85-55, 2005.
[47] B. L. Pellorn and J. H. L. Hansen, "A duration-based Confidence Measure for Automatic Segmentation of Noise Corrupted Speech," *Proc. of ICSLP1998*, 1998.
[48] S. Goronzy *et al.*, "Phone-duration-based Confidence Measures for Embedded Applications," *Proc. of ICSLP2000*, pp. 500-503, 2000.
[49] J. Doremalen *et al.*, "Utterance Verification in Language Learning Applications," *Proc. of the SLaTE2009*, 2009.
[50] S. E. Levinson, "Continuously Variable Duration Hidden Markov Models for Speech Analysis," *Proc. of ICASSP1986*, pp. 1241-1244, 1986.
[51] F. Ramus, "Acoustic Correlates of Linguistic Rhythm: Perspectives," *Proc. of SpeechProsody2002*, pp. 115-120, 2002.
[52] F. B. Meng, H. Meng, Z. Y. Wu, and L. H. Cai, "Synthesizing Expressive Speech to Convey Focus using a Perturbation Model for Computer-Aided Pronunciation Training," *Proc. of Second Language Studies: Acquisition, Learning, Education and Technology*, 2010.
[53] P. Wik, and A. Hjalmarsson, "Embodied conversational agents in computer assisted language learning," *Speech Communication*, 51(10):1024-1037, October 2009.
[54] University of Iowa, "The Sound of Spoken Language," http://www.uiowa.edu/~acadtech/phonetics.
[55] D. L. F. Nilsen and A. P. Nilsen, "Pronunciation Contrasts in English," Simon and Schuster, 1973.
[56] FreeTTS 1.2, http://freetts.sourceforge.net.
[57] Flite, http://www.speech.cs.cmu.edu/flite.
[58] J. Q. Wang, K. H. Wong, P. A. Heng, H. Meng, and T. T. Wong, "A Real-Time Cantonese Text-to-Audiovisual Speech Synthesizer," *Proc. of ICASSP2004*, 2004.
[59] K. H. Wong, W. K. Leung, W. K. Lo, and H. Meng, "Language Learning with Articulatory Visual-Speech Synthesizer," *submitted to ISCSLP2010*, 2010.
[60] X-J Qian, F. Soong, and H. Meng, "Discriminative Acoustic Model for Improving Mispronunciation Detection and Diagnosis in Computer-Aided Pronunciation Training (CAPT)," *Proc. of Interspeech2010*, 2010.
[61] H. Meng, C-Y Tseng, M. Kondo, A. M. Harrison and T. Viscelgia, "Studying L2 Suprasegmental Features in Asian Englishes: A Position Paper," *Proc. of Interspeech2009*, 2009.
[62] S. Zhang, K. Li, W. K. Lo, and H. Meng, "Perception of English Suprasegmental Features by non-native Chinese Learners," *Proc. of SpeechProsody2010*, 2010
[63] H. Wang, P. Mok, and H. Meng, "MusicSpeak: Capitalizing on Musical Rhythm for Prosodic Training in Computer-Aided Language Learning," *Proc. of Second Language Studies: Acquisition, Learning, Education and Technology,* 2010.