Capturing L2 Segmental Mispronunciations with Joint-sequence Models in Computer-Aided Pronunciation Training (CAPT)

Xiaojun Qian[1,2,3], Helen Meng[1,2] and Frank Soong[1,2,3]
[1]CUHK MoE-Microsoft Key Laboratory of Human-Centric Computing and Interface Technologies
[2]Department of SEEM, The Chinese University of Hong Kong, Hong Kong SAR
[3]Speech Group, Microsoft Research Asia, Beijing
{xjqian, hmmeng}@se.cuhk.edu.hk, frankkps@microsoft.com

***Abstract*—In this study, we present an extension to our previous efforts on automatically detecting text-dependent segmental mispronunciations by Cantonese (L1) learners of American English (L2), through modeling the L2 productions. The problem of segmental mispronunciation modeling is addressed by joint-sequence models. Specifically, a grapheme-to-phoneme model is built to convert the prompted words to their corresponding possible mispronunciations, instead of the previous characterization of language transfer through phonological rules. Experiments show that the proposed approach can better capture the mispronunciations compared with the use of phonological rules.**

I. Introduction

"Mispronunciations" refer to incorrect or inaccurate pronunciations, or simply "errors". Generally speaking, for non-native speakers, there can be *supra-segmental* errors [1] - occurring in lexical stress, utterance-level stress, intonation and phrasing, etc.. There can also be *segmental* errors, e.g. a common mispronunciation made by Cantonese learners of English is to produce $/b\ ow\ f/$ for $/b\ ow\ th/$ ("both"). Here, a phonetic unit in the target language (especially if non-existent in the learner's mother tongue) may be substituted with on that exists in the mother tongue.

The goal of Computer-Aided Pronunciation Training (CAPT) for language learning is to detect mispronunciations produced by non-native learners and provide appropriate feedback to help them improve. In a typical scenario, the system prompts the learner with a sentence or paragraph to read aloud, and preferably detailed feedback is presented to the learner after the recorded speech is analyzed. For example, a learner may mispronounce the word "rice" as $/l\ ay\ s/$ ("lice"), and the CAPT system should be able to respond: "You have mispronounced the phone $/r/$ as $/l/$."

There has been a great deal of research on mispronunciation detection to promote Computer-Aided Pronunciation Training (CAPT) during the past two decades [2]. Most of them can be classified into two categories: (1) use of confidence measures based on ASR, e.g. GOP [3] and Scaling Posterior Probability [4]; and (2) classification using other acoustic-phonetic features, e.g. LDA on formants and durations [5], etc.

In our previous work, we adapted the ASR-based framework for mispronunciation detection by the incorporation of linguistic knowledge and the introduction of an extended pronunciation dictionary or network [6][7][8][9]. We also show that optimizing the recognizer's performance metrics in terms of "*false acceptances*", "*false rejections*" and "*diagnostic errors*" is equivalent to minimum word error discriminative training, and the error minimization can lead to significant performance boost for the acoustic models [10]. The bottleneck of the recognizer is the inability to capture as many possible mispronunciation patterns as possible.

The paper aims to break the bottleneck by showing how the use of grapheme-to-phoneme generation [11] can be applied to the mispronunciation modeling task.

Our operating assumption is that we can use the ARPABET to phonetically transcribe L2 English.

The paper is organized as follows: The second section deals with the corpus preparation. In the third section, the generation of text-dependent mispronunciation is modeled. The joint-sequence model originally proposed for grapheme-to-phoneme conversion is briefly reviewed in the fourth section. The experimental results and conclusions are given in the fifth and sixth sections, respectively.

II. Corpus Preparation

Our investigation is based on the CU-CHLOE corpus [12], which contains recordings of 100 Cantonese-speaking learners of English (50 male and 50 female) reading minimal pairs, confusable words, phonemic sentences and the Aesop's Fable "The North Wind and the Sun".

We split the whole corpus into training and testing sets where the speakers are disjoint, but the text prompts for recording are the same. This means that the training set provides full lexical knowledge of the test set in terms of canonical pronunciations. However, the training set does not offer any knowledge about mispronunciations made by speakers in the test set. So if the errors are repeated, we claim that the error generalizes across speakers and are worthy of modeling. There are indeed errors that are quite idiosyncratic (speaker-dependent) and are constrained to few speakers. Our objective

TABLE I
STATISTICS ON THE TRAINING AND TESTING SET OF THE CU-CHLOE CORPUS.

Sets	# of words	# of pronunciations	# of mispronunciations
training	435	3,794	3,308
testing	435	3,568	3,085

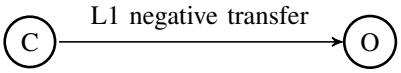

Fig. 1. A graphical model describing the relationship between the canonical pronunciations and the surface observations. C - canonical pronunciations; O - observed pronunciations.

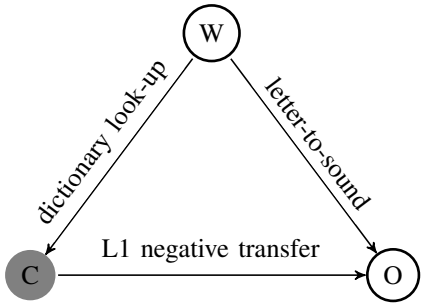

Fig. 2. Directed acyclic graph representing the cause and effect relations among the prompted words, the canonical pronunciations and the observed surface mispronunciations. W - prompted words; C - canonical pronunciations; O - observed pronunciations.

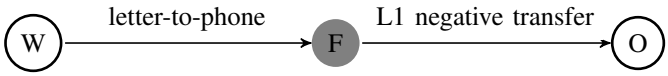

Fig. 3. A new graphical model compactly representing the dependencies among the prompted words, the canonical pronunciations and the mispronunciations. W - prompted words; F - phone sequences in mind; O - observed pronunciations.

is to model with high priority the more "common" errors that are generalizable across speakers.

The statistics of the sets in terms of the number of distinct words, pronunciations and mispronunciations are shown in Table I. There is a total of 1448 overlapping mispronunciations in the training and testing sets. The speech has been annotated by well-trained linguists with the ARPABET phonetic symbols.

III. MODELING THE TEXT-DEPENDENT PRODUCTION OF MISPRONUNCIATIONS

Previous works based on context-senstive phonological rule modeling [6][7][8][9] is basically assuming that the L2 learners will apply the phonological characteristic of their L1 for the L2, and this phenomenon dominates. The joint effect of these rules on a canonical pronunciation can transduce it into a batch of possible mispronunciations. This process can be illustrated as a graphical model in Figure 1.

By inspecting a lot of annotated word pronunciations in the corpus, we find there are mainly three causes of mispronunciations: (1) **L1 negative transfer**, e.g. "the" can be mispronounced as $/d\ ax/$; (2) **Incorrect letter-to-sound conversion**, e.g. "analyst" can be realized as $/ae\ n\ ax\ l\ ay\ s\ t/$ due to its orthographic similarity with "analyze", and we call it "*mispronunciation by analogy*"; (3) **Misread words**, e.g. "cloak" is sometimes mistaken for "clock". Due to the relatively small number of samples of mispronunciations caused by misread words, we neglect factor (3) in our analysis.

Based on these observations, we first construct an intuitive graphical model describing explicitly the cause-effect relations among "prompted word", its "canonical pronunciation(s)" and the "observed pronunciations" as shown in Figure 2. Each directed edge represents the dependency between the two random variables involved. For example, the edge from W to C can be interpreted as the dictionary-lookup or memory recall by the learner, and C here is hidden since we can not observe it throughout the process; Likewise, the edge from C to O indicates the effect of L1 negative transfer, which is the same as the phonological process depicted in Figure 1; The edge from W to O characterizes the letter-to-sound conversion.

We see that the observation O has two possible causes: W and C. Since we do not have the ground truth of whether an observed pronunciation is more likely to be caused by L1 negative transfer or incorrect letter-to-sound conversion, directly estimating the parameters of this graphical model is not easy. To simplify this model, we introduce another latent variable H between W and O, and decompose the edge from W to O, denoted by $\Pr(O|W)$, to:

$$\Pr(O|W) = \sum_{H} \Pr(O|H) \Pr(H|W) \tag{1}$$

Again, $\Pr(O|H)$ explains the course of L1 negative transfer and $\Pr(H|W)$ can be regarded as letter-to-phone conversion by the learner. If we merge the variable H and C for a new latent variable F, a compact equivalent form of the original model is depicted in Figure 3.

This simplification is valid, because it captures the cognitive process of mispronunciation production. Image when a learner is given some text prompt to read aloud, he may immediately generate a sequence of phonemes in his mind. The sequence can be produced by his own knowledge of letter-to-phone sequence conversion if the learner is not familiar with the word, or the sequence can possibly be the result of a "dictionary look-up" or memory recall if he is informed about the word. When the phoneme sequence is articulated, it may be further distorted by the mechanism of L1 negative transfer.

Now, the problem has been casted to estimating $\Pr(O|W)$ instead of estimating the structure of $\Pr(O|C)$ in [10][13], and it looks extremely similar to the grapheme-to-phoneme conversion problem. We reference the state-of-the-art technique of generative joint-sequence model. A brief review will be given in the next section.

IV. The Bisani and Ney's Joint-sequence Model

The approach taken by Bisani and Ney for grapheme-to-phoneme conversion is generative [11]. Given a sequence of letters g, the task of grapheme-to-phoneme conversion can be formalized as getting the N-best phone sequence ϕ such that $p(\phi|g)$ is maximized. By Bayes' decision rule, it is equivalent to maximizing $p(g, \phi)$.

The joint probability $p(g, \phi)$ can be expressed as:

$$p(g, \phi) = \sum_{q \in S(g,\phi)} p(q) \tag{2}$$

where $q = (g, \phi)$, called a graphone, is the pair of a letter sequence and a phoneme sequence of possibly different length. $S(g, \phi)$ is the set of all co-segmentations of g and ϕ:

$$S(g, \phi) = \{q | q_{q_1} \cup \ldots \cup g_{q_K} = g; \phi_{q_1} \cup \ldots \cup \phi_{q_K} = \phi\} \tag{3}$$

Equation (2) is simply saying the joint probability $p(g, \phi)$ is determined by summing over all matching graphone sequences. Hence, $p(g, \phi)$ has been reduced to a probability distribution $p(q)$ over graphone sequences $q = q_1, \ldots, q_K$, where $K = |q|$ is the length of the graphone sequence q.

The graphone sequence can be further modeled using a standard M-gram approximation:

$$p(q_1^K) \approx \prod_{j=1}^{K+1} p(q_j | q_{j-1}, \ldots, q_{j-M+1}) \tag{4}$$

By introducing the symbol h to denote the sequence of preceding joint units $h_j = (q_{j-M+1}, \ldots, q_{j-1})$, $n_{q,h}(q)$ is defined as the number of occurrences of the M-gram $q_{j-M+1}, \ldots, q_j$ in q. Starting from model parameters initialized by assigning a uniform distribution over all graphones satisfying certain manually set length constraints, an EM procedure is employed to re-estimate the model parameter θ, iteratively:

$$p(q; \theta) = \prod_{j=1}^{|q|} p(q_j | h_j; \theta) \tag{5}$$

$$\begin{aligned} e(q, h; \theta) &= \sum_{i=1}^{T} \sum_{q \in S(g_i, \phi_i)} p(q | g_i, \phi_i; \theta) n_{q,h}(q) \\ &= \sum_{i=1}^{T} \sum_{q \in S(g_i, \phi_i)} \frac{p(q; \theta)}{\sum_{q' \in S(g_i, \phi_i)} p(q'; \theta)} n_{q,h}(q) \end{aligned} \tag{6}$$

$$p(q | h; \theta') = \frac{e(q, h; \theta)}{\sum_{q'} e(q', h; \theta)} \tag{7}$$

where T is the number of training samples.

Standard M-gram language modeling techniques including "evidence-trimming" to avoid over-fitting and "model smoothing" to extend the generalizability are later applied when ramping up the lower-order model to a higher-order one.

Given the estimated the model, an N-best search is performed based on the posterior $p(\phi|g)$:

$$p(\phi|g) = \frac{p(g, \phi)}{p(g)} = \frac{\sum_{q \in S(g,\phi)} p(q)}{p(g)} \tag{8}$$

where,

$$p(g) = \sum_{\phi} p(g, \phi) = \sum_{g(q)=g} p(q) \tag{9}$$

V. Experiments

A. Baseline Setup

In [10][13], the context-sensitive phonological rules takes the form:

$$\phi \rightarrow \psi \ / \ \lambda \_\_ \rho \tag{10}$$

This rule is interpreted as follows: ψ in the target language may be pronounced as ψ when following λ and preceding ρ. In [6] and [7], the λ and ρ in the context can include multiple phones, a group of phones (e.g. the set of vowels, denoted by the symbol "V") or no phones, while ϕ and ψ in the rewrite mapping are restricted to a single phone.

The rules $\{r_i\}$ are expressed as Finite State Transducers using the open-source toolkit OpenFST [14]. The Extended Recognition Network (expressed as Finite State Acceptor) comprising the possible pronunciations [9] can be obtained by applying the rules to the canonical pronunciation Φ based on the following expression:

$$\Phi \circ (\bigcup_{i=1}^{N} ((Id*) \cup r_i)) \circ ((Id*) \cup r_1) \circ \ldots \circ ((Id*) \cup r_N) \tag{11}$$

where $\circ$ is the *composition* operation, $\cup$ is the *union* operation, $*$ is the *closure* operation, N is the number of rules and Id is the *identity* FST which transduces every input symbol to the output intact. The expression is simply saying that each rule can be independently applied to any location of the canonical pronunciation if there is a match. The respective outputs by each rule are unified, to which the rules are further applied in a cascade fashion.

The rules in [9] are manually derived from second-language acquisition literature, and are thus knowledge-based. Later, data-driven rule extraction approaches are proposed [10][13]. ϕ and ψ are allowed to incorporate multiple phones to capture interesting patterns from the data, but the context is restricted to one single phone only, and the symbol "#" is used to denote word boundaries.

In [13], to form a basic set of rules, the manually labeled L2 transcriptions in the training set are first aligned with their canonical pronunciations using phonetically-sensitive alignment [8] and then all mismatched phone pairs are extracted with their left and right contextual phones. To alleviate false alarms, these rules are first sorted in descending order according to the number of occurrences in the training set, and then they are pruned incrementally to optimize for their F-measure [13]. Since it is combinatorially hard to determine the set of rules given a particular set size, especially when some of the

TABLE II
PERFORMANCE OF THE KNOWLEDGE-BASED RULE AND THE DIFFERENT SETS OF DATA-DRIVEN RULES PRUNED BY THE NUMBER OF SUPPORTING SAMPLES IN THE TRAINING SET. THE "PRECISION" AND "RECALL" SHOWN BELOW ARE ALL WORD-BASED.

rules	threshold	training		testing	
knowledge		precision	recall	precision	recall
		23.98%	15.96%	23.66%	16.88%
data-driven	3	7.81%	34.70%	7.22%	34.39%
	4	9.07%	30.77%	8.53%	31.05%
	5	12.80%	26.09%	12.34%	27.81%
	6	16.10%	24.33%	15.72%	25.48%
	7	20.65%	21.49%	20.80%	23.21%
	8	21.95%	19.62%	22.35%	21.43%
	9	25.22%	17.50%	25.78%	19.19%
	10	54.45%	16.26%	55.47%	17.76%

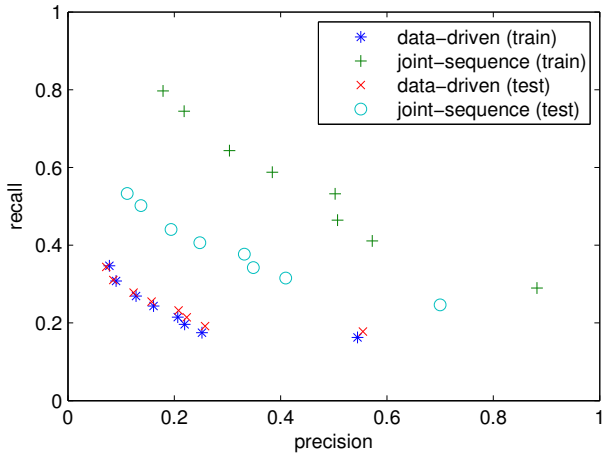

Fig. 4. Precision versus recall plot for the data-driven rules and the joint-sequence models in both training and testing sets.

rules have the same number of occurrences, in this study we prune rules according to different thresholds of occurrences. For the sake of comparison, we define the following measures:

- **precision** - the number of modeled mispronunciations over the number of mispronunciations returned by the rule set;
- **recall** - the number of modeled mispronunciations over the number of mispronunciations found in the evaluation set.

The statistic on these different sets of rules is shown in Table II. One the one hand, the data-driven rules can outperform the knowledge-based rules in terms of both precision and recall (see the row with a threshold of 9 in Table II); On the other hand, data-driven rules offer more flexibility in optimizing for the mispronunciation detection and diagnosis performance [13]. As the data-driven rules are more favorable than knowledge-based rules, the knowledge-based rules are not further analyzed. Only the sets of data-driven rules are setup as the baseline.

Although the recall of the rule sets seems to be low, we point out that among the data-driven rules, even the rule set with the lowest recall (see last row of Table II) can capture 9666 mispronounced tokens out of 14414 mispronounced tokens in the training set (67.06%), and 9197 out of 13624 in the testing set (67.51%), due to the repeated occurrences of common errors.

B. Experiments on the Joint-sequence model

Our implementation of the grapheme-to-phoneme joint-sequence model is based on the Open-source Sequitur G2P toolkit. All pairs of letter sequence (prompted word) and phoneme sequence (annotated pronunciation) in the training set are used to estimate the joint-sequence models.

To compare the performance of the joint-sequence model fairly with the data-driven rules, for each word in the training set, we generate from the joint-sequence model the same number of N-best pronunciations as are returned by the respective sets of data-driven rules. The precision versus recall plot for these two approaches is shown in Figure 4. We see the precision recall curve by the joint-sequence model extends away significantly from the origin. This is explained by the joint-sequence model's capability to model directly the letter-to-sound errors in the data, while in such cases the data-driven rule approach based on alignments between the canonical pronunciation and the mispronunciation would possibly leads to many unjustifiable phonological rules that generalize poorly.

C. Mispronunciation Detection and Diagnosis

To further investigate the effect of predicting mispronunciations by joint-sequence models on the acoustic models, we carry out mispronunciation detection and diagnosis by populating each word's phone lattice with its possible mispronunciations predicted by the joint-sequence model.

We train cross-word, tied-state, Gaussian mixture, triphone HMMs on TIMIT in Maximum Likelihood, and adapt those with the training set using Constrained Maximum Likelihood Linear Regression [15] to compensate for the mismatch between the native and non-native model space.

This model is utilized to align the mispronunciation phone lattices from both the data-driven rule and the joint-sequence model with the L2 speech in the testing set. In general, the recognition performance in terms of the percentage of matching words (having the same phone sequence) between the manual transcription and recognition output is shown in Figure 5. By using the lattice from the joint-sequence model, the percentage of matching word tokens is higher than those yielded by the data-driven rules almost everywhere.

Since most of the "*Diagnostic Errors*" are caused by failing to include the actual mispronunciation in the lattice [10], and the joint-sequence model is designed specifically to tackle this problem, we inspect particularly at the "*Diagnostic Accuracy*" in the testing set, which is defined as the number of correct word diagnosis (correctness in identifying the type of word mispronunciations, e.g. identifying $/r\ ay\ s/ \rightarrow /l\ ay\ s/$) over the number of truly detected word errors. A comparison of the diagnostic accuracy between the joint-sequence model

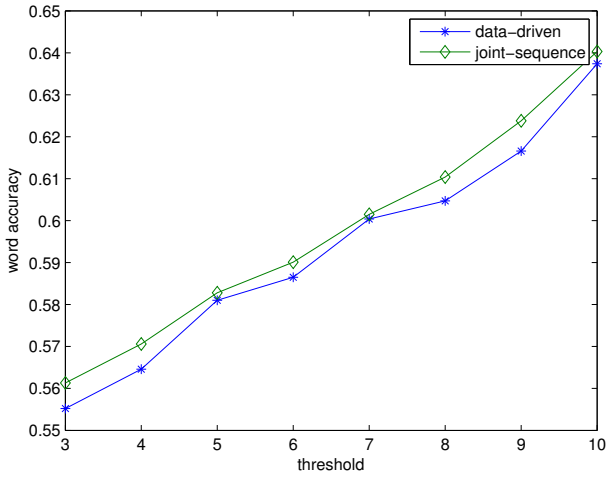

Fig. 5. Performance of the recognition in terms of word accuracy in the testing set. The rules are pruned according to different thresholds of occurrences, and the same number of N-best pronunciations are generated from the joint-sequence model as are returned by the respective sets of data-driven rules.

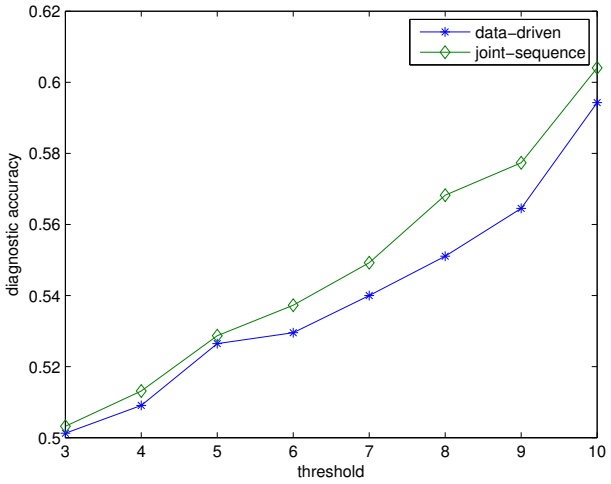

Fig. 6. Comparison of the recognizer's diagnostic accurarcy between the lattice generated by joint-sequence model and the one generated from the data-driven rule.

and the data-driven rule is illustrated in Figure 6, and the joint-sequence model seems achieves higher accuracy over the data-driven rules on the accuracy. This confirms our claim on constructing decoding lattice by possible word mispronunciations more accurately.

VI. Conclusions

In this work, we formalize the sub-problem of mispronunciation modeling in ASR-based text-dependent mispronunciation detection as a grapheme-to-phoneme conversion problem. The state-of-the-art joint-sequence model is applied to predict possible mispronunciation patterns for each word. Experimental results on our L2 speech corpus shows it can populate the extended recognition network with mispronunciation patterns more accurately and compactly. Correspondingly, it also offers mispronunciation diagnosis improvement on our baseline ASR system. Combining it with discriminative training [10] is expected to boost the system performance of mispronunciation detection and diagnosis further.

Acknowledgment

The authors would like to acknowledge Patrick Chu, Sam Wong and Mingxing Li of HCCL at CUHK for their annotations. This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies, and is is conducted under the MSRA-CUHK joint laboratory scheme, and is partially supported by the NSFC/RGC Joint Research Scheme (project no. N_CUHK 414/09).

References

[1] S. Zhang, K. Li, W.K. Lo and H. Meng, *Perception of English Suprasegmental Features by Non-Native Chinese Learners*. In the Proceedings of the Fifth International Conference on Speech Prosody, Doubletree Magnificent Mile, Chicago, USA, 2010.

[2] S. Wei, G. Hu, Y. Hu and R.H. Wang, *A New Method for Mispronunciation Detection using Support Vector Machine Based on Pronunciation Space Models*. In Speech Communication, Vol. 51, pp. 896 - 905, 2009.

[3] S.M. Witt and S.J. Young. *Phone-level pronunciation scoring and assessment for interactive language learning*. In Speech Communication, Vol. 30, pp. 95 - 108, 2000.

[4] F. Zhang, C. Huang, F.K. Soong, M. Chu, R.H. Wang. *Automatic mispronunciation detection for Mandarin*. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 2077 - 2080, 2008.

[5] K. Truong. *Automatic pronunciation error detection in Dutch as a second language: an acoustic-phonetic approach*. Master Thesis, Utrecht University, The Netherlands, 2004.

[6] L. Wang, X. Feng and H. Meng, *Automatic Generation and Pruning of Phonetic Mispronunciations to Support Computer-Aided Pronunciation Training*. In the Proceedings of Interspeech, Brisbane, Australia, 2008.

[7] L. Wang, X. Feng and H. Meng, *Mispronunciation Detection Based on Cross-Language Phonological Comparisons*. In the Proceedings of the IEEE IET International Conference on Audio, Language and Image Processing, Shanghai, China, pp. 307 - 311, 2008.

[8] 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, 2008.

[9] A.M. Harrison, W.K. Lo, X.J. Qian and H. Meng, *Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training*. In the Proceedings of the 2nd ISCA Workshop on Speech and Language Technology in Education (SLaTE), Warrickshire, 2009.

[10] X.J. Qian, F. Soong and H. Meng, *Discriminative Acoustic Model for Improving Mispronunciation Detection and Diagnosis in Computer-Aided Pronunciation Training (CAPT)*. To appear in the Proceedings of Interspeech, Makuhari, Japan, 2010.

[11] M. Bisani and H. Ney, *Joint-sequence Models for Grapheme-to-phoneme Conversion*. In Speech Communication, Vol. 50, pp. 434 - 451, 2008.

[12] H. Meng, Y.Y. Lo, L. Wang and W.Y. Lau, *Deriving Salient Learners' Mispronunciation from Cross-Language Phonological Comparisons*. In the Proceedings of ASRU, Kyoto, Japan, 2007.

[13] W.K. Lo, S. Zhang and H. Meng, *Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System*. To appear in the Proceedings of Interspeech, Makuhari, Japan, 2010.

[14] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut and M. Mohri, *OpenFst: A General and Efficient Weighted Finite-State Transducer Library*. In the Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007), volume 4783 of Lecture Notes in Computer Science, pages 11-23. Springer, 2007. http://www.openfst.org.

[15] M.J.F. Gales, *Maximum Likelihood Linear Transformations for HMM-based Speech Recognition*. In Computer Science and Language, Vol. 12, pp. 75 - 98, 1998.