A Two-Pass Framework of Mispronunciation Detection and Diagnosis for Computer-Aided Pronunciation Training

Xiaojun Qian, *Member, IEEE*, Helen Meng, *Fellow, IEEE*, and Frank Soong, *Fellow, IEEE*

***Abstract*—This paper presents a two-pass framework with discriminative acoustic modeling for mispronunciation detection and diagnoses (MD&D). The first pass of mispronunciation detection does not require explicit phonetic error pattern modeling. The framework instantiates a set of antiphones and a filler model to augment the original phone model for each canonical phone. This guarantees full coverage of all possible error patterns while maximally exploiting the phonetic information derived from the text prompt. The antiphones can be used to detect substitutions. The filler model can detect insertions, and phone skips are allowed to detect deletions. As such, there is no prior assumption on the possible error patterns that can occur. The second pass of mispronunciation diagnosis expands the detected insertions and substitutions into phone networks, and another recognition pass attempts to reveal the phonetic identities of the detected mispronunciation errors. Discriminative training (DT) is applied respectively to the acoustic models of the mispronunciation detection pass and the mispronunciation diagnosis pass. DT effectively separates the acoustic models of the canonical phones and the antiphones. Overall, with DT in both passes of MD&D, the error rate is reduced by 40.4% relative, compared with the maximum likelihood baseline. After DT, the error rates of the respective passes are also lower than those of a strong single-pass baseline with DT by 1.3% and 5.1% relative which are statistically significant.**

***Index Terms*—Computer-aided pronunciation training, mispronunciation detection and diagnosis, discriminative training.**

I. Introduction

Second-language learners are eager to attain high pronunciation proficiency of the target language. However, learning the unfamiliar sounds and prosody of a new language is not easy. High proficiency is attained through extensive practice with guidance from language instructors. But good (especially native) language instruction is often a scarce resource. Automatic speech recognition (ASR) technologies can support an online computer-aided pronunciation training (CAPT) platform that supplements teachers' instructions with round-the-clock accessibility and individualized feedback for second-language learners. CAPT platforms integrate various spoken language technologies, among which an essential technology is mispronunciation detection, which refers to locating a phone that is incorrectly articulated, and involves a binary decision. To provide useful feedback for the learner, another important technology is mispronunciation diagnosis, which identifies the incorrect phone(s) produced in place of the canonical phone.

Predominant approaches to CAPT primarily performs mispronunciation detection by extending ASR technologies, especially through post processing recognition scores (e.g., by thresholding [1] or classification [2]). Most previous work did not focus on mispronunciation diagnosis [3], [4]. Later on, there has been some work on modeling error patterns in the learners' speech using linguistic knowledge [5] or adopting a data-driven approach [6]. Error pattern modeling enables generation of a set of possible phonetic error patterns for a given text prompt. A single-pass forced-alignment using the recognition network expanded by the error patterns enables phonetic error detection and diagnosis of erroneous phonetic production in an integrated fashion. The major consideration behind such paradigm of MD&D is to properly constrain the search space in ASR and thus avoiding a shift towards the free-phone recognition task (which is a more intractable problem). However, explicit error pattern modeling is often infeasible when no prior knowledge is available for a given L1-L2 (first and second) language pair, or when it is too costly to perform error pattern derivation with [7] or without [8] labeled non-native speech. Other risks include insufficient error pattern coverage. Contrastively, inclusion of overly redundant error patterns, which may include rare or idiosyncratic ones, is also a risk, because it may be difficult for the acoustic model to differentiate them. These risks hurt the ability to detect the errors or diagnose them.

This paper presents a two-pass framework for MD&D which attempts to break away from the reliance on explicit error pattern modeling. Imposing no prior knowledge on the possible forms of error patterns implicitly aims for full error pattern coverage. This may lead to exponentially many variants in the search space. Such an intractable search space presents a challenge to MD&D, especially in terms of an acoustic model with insufficiently strong discriminative ability. Hence, the proposed approach aims to reduce the search space by pairing each canonical phone (given the text prompt) with an anti-phone which covers the complementary acoustic space. A first-pass recognition in this network of pairs detects phonetic

Manuscript received June 28, 2015; revised November 15, 2015 and January 18, 2016; accepted January 27, 2016. Date of publication February 08, 2016; date of current version April 29, 2016. This work was supported by the Hong Kong SAR Government Research Grants Council (Project reference no. 415511). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Bin Ma.

X. Qian and H. Meng are with the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong (e-mail: xjqian@se.cuhk.edu.hk; hmmeng@se.cuhk.edu.hk).

F. Soong is with Speech Group, Microsoft Research Asia, Beijing 100080, China (e-mail: frankkps@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2526782

substitutions. Furthermore, this first-pass can also detect insertions and deletions by introducing a filler model and designing the network topology which allows phone skips. Once the insertions and substitutions are detected, a second pass performs free-phone recognition on the segments of detected insertions and substitutions to identify the actual phones. Correspondingly, we refine the two sets of HMM-based acoustic models by individualized discriminative training to minimize the expected full-sequence phone-level errors. In summary, the two-pass framework attacks the MD&D problem in a step-by-step manner and optimize the acoustic models separately to suit the respective passes.

The rest of the paper is organized as follows: Section 2 reviews previous work on mispronunciation detection and diagnosis. Section 3 introduces the phonetically-labeled corpus used for acoustic model training and evaluation. Section 4 explains the two-pass framework. Section 5 details the evaluation metrics. Section 6 presents experimentation that demonstrate the advantage of the two-pass framework compared to the single-pass framework and the performance improvement of discriminatively trained acoustic models over a baseline. Section 7 concludes the paper.

II. Previous Work

The sections starts by grouping previous work on mispronunciation detection according to the set of features employed. Originally, mispronunciation detection is mostly built upon the confidence measures derived from ASR. Later, it is extended to other features when there is prior knowledge on which features are prominent in differentiating correct or incorrect pronunciations. Then, the section proceeds to the single-pass framework where mispronunciation diagnosis is first enabled by explicit modeling of the possible error patterns.

A. Mispronunciation Detection Using Phone-Level Scores

Phone-level scores are usually measured in terms of ASR confidence measures [10], such as the posterior probability. It is implicitly assumed that the score is positively correlated with the probability of correctness per speech segment [11]. Examples of phone-scoring algorithms include: the "Goodness of Pronunciation" (GOP), introduced by Witt & Young [1], which is an approximation to the posterior score (a relative measure which takes into consideration the likelihood of other competing phones in addition to that of the canonical phone) and uses either a frame-based or phone-based likelihood ratio; as well as a log-likelihood score [12], which represents an absolute measure of how closely a pronunciation approximates a given phone model which may vary considerably among different phone models. There is also a variety of modified versions of likelihood scores, posterior scores or likelihood ratios that are pursued in follow-up papers, e.g., [13]. Mispronunciation detection using phone-level scores is often posed as a binary classification problem. This may be simply achieved by thresholding where the thresholds are pre-determined empirically from data [14], [15] to optimize detection performance. Alternatively, an array of phone-level confidence measures which span a "pronunciation space" [2] can be utilized as the features for classification, e.g., the posterior probability of the target phone as well as those for the competing phones. Binary classifiers of various kinds may be used, e.g., SVM in [2] and [9]. Still other options may include the confidence measures that are estimated in alternative ways, e.g., using neural nets [16] or by introducing other sources of information, e.g., manners of articulation [17].

B. Mispronunciation Detection Using Other Features

Acoustic information other than phone-level scores can also be incorporated as the input to classifiers. A considerable amount of feature design and engineering has been done in the literature according to the *a priori* knowledge on identifying the set of features which is more effective in distinguishing confusable phone pairs. For example, segment duration [18], acoustic-phonetic features including log root mean-square (RMS) energy, the first-order derivative of log RMS energy and zero-crossing-rate [19], adaptively-warped cepstrum [20], and low-dimensional sub-space features [21]. When the gold standard or template (i.e., reference utterance from a native speaker for the same prompt) is available, one may also exploit features which measure the differences between the native and non-native speakers' realizations [22]. In a recent "template"-based approach [23], dynamic time warping (DTW) is carried out between a student's utterance and a teacher's utterance. Various word-level and phone-level features are extracted to describe the degree of mis-alignment in the warping path and the distance matrix, and are taken as the input to an SVM for classification.

C. Single-Pass Mispronunciation Detection and Diagnosis Using Extended Recognition Networks

Most of the previous work does not consider mispronunciation diagnosis, however, accurate identification of the pronunciation error can help close the loop of CAPT by providing a corrective feedback. The "*extended recognition network*" (ERN) framework achieves mispronunciation detection and diagnosis at the same time, by explicit incorporation of the common phonetic error patterns in addition to the canonical transcription per prompted word. Thus, it constrains the search space during phone recognition of non-native prompted speech. By running a single recognition pass through the ERN, we are able to achieve both mispronunciation detection and diagnosis, i.e. telling which phone is mispronounced as another phone, in additional to pinpointing what is wrong. Prior knowledge on the native and non-native language pair can be a useful source of information. For instance, non-native speakers may substitute a phoneme in the target language with phonemes from their mother tongue or they may have difficulties in perceiving and/or realizing phonetic contrasts that are not distinctive in their mother tongue. Apart from the substitutions, insertion or deletion of phones are quite common as well. For example, inserting vowels within consonant clusters or after syllable-final consonants, and deleting syllable-final liquids are particularly frequent in English by Japanese learners [24]. Such phonetic confusions, either context-dependent or context-independent, may be captured in manually-authored phonological rules [25]. When phonetically-labelled transcriptions are available, phonological rules which describe transductions from canonical phones to mispronounced phones can be automatically derived from the transcriptions [7]. This is done by aligning and comparing the canonical phonetic transcription with the manually-labelled phonetic transcriptions involving errors. When there are only speech recordings available, rules can be extracted directly from speech [30], [31]. Usually, the automatically derived rules can reveal phonological processes that are missing from manually-authored rules, but some rules may also have rare occurrences that may result from incorrect transcriptions, misread words, etc. Possible solutions include selectively enabling rules using decision trees [26], pruning rules according to their occurrences [7], introducing a cutoff probability to the generated variants [27], and explicitly employing letter-to-phone rules to cover errors due to orthographic interference [28], [29].

III. The Two-Pass Framework

As introduced in Section II, previous work on single-pass MD&D relies on explicit error pattern modeling in the ERN, thus the performance of MD&D is dependent on the quality of ERN which usually requires a trade-off. The two-pass framework attempts to abandon explicit error pattern modeling and aims for relaxation of the search space. However, over-relaxation leads to an intractable search space. The proposed approach aims to reduce the search space by pairing each canonical phone with an anti-phone which covers the complementary acoustic space. Therefore, the first-pass mispronunciation detection tells whether a phone is mispronounced or not. Also, insertions and deletions are taken into account by a careful design of the recognition network. Once the insertions and substitutions are detected, the second-pass mispronunciation diagnosis performs free-phone recognition on the segments of detected insertions and substitutions to identify the actual phone mispronunciations.

A. Mispronunciation Detection

A phone substitution is an acoustic deviation from the canonical production, so it occupies the fraction of the non-silence acoustic space that is complementary to that of the canonical phone. Here, we model this complementary space with anti-phones (in the form of GMM-HMMs) to detect phone substitutions. The concept is simple – in addition to fitting the labelled data belonging to a phone, we directly construct a model to fit data that do not belong to the phone. The recognition network is augmented by pairing each canonical phone with its anti-phone. In this way, each phone is subject to a binary classification. In the case of phone deletions, one needs to allow a phone to be skipped. Phone insertions can possibly happen at every location of a canonical pronunciation, and there can be multiple instances of insertions at a single location. To capture insertions, we introduce a *universal phone model* (UPM, also known as the *filler* model) which covers all the non-silence phones. The UPM is padded between each successive phones as an optional phone loop. The mispronunciation detection network for the word "THE" is shown in Figure 1. A recognition pass in the detection network leads to transcriptions like: [_th ax UPM]. This refers to the alignment between the canonical transcription and the specific recognition transcription, that [th] is substituted (i.e., with the anti-phone [_th]), and a phone (i.e., the UPM) is inserted at the end of the word.

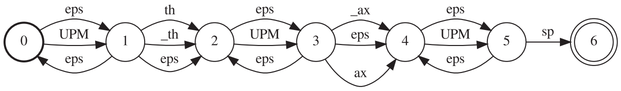


Fig. 1. Mispronunciation detection network for the word "THE" [th ax]. 'eps' stands for a non-emitting skip, UPM is short for *universal phone model* which is also known as *filler model*. The anti-phones share a '_' prefix.

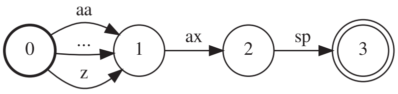


Fig. 2. Mispronunciation diagnosis network for the word "THE" once a substituted [th] is detected. [th] does not appear between state 0 and 1.

One possible pitfall of such a design of a mispronunciation detection network is that the UPM may incur unnecessary insertions, as there is overlap between the the acoustic space spanned by the anti-phones and that by the UPM. So the two models may compete to gain control over a segment of frames which is accessible to both of them. The issue shall be discussed experimentally later.

B. Mispronunciation Diagnosis

Once anti-phones and UPMs are found in the transcription, mispronunciation diagnosis targets revealing the phone identities of the detected phone errors. The diagnosis network is constructed as follows: for the detected anti-phones, it is expanded by all the other possible canonical phones, and for the detected UPM, it is expanded by all possible phones. Suppose the transcription from the detection is [_th ax], the expanded network for diagnosis is shown in Figure 2.

As can be seen, the mispronunciation diagnosis network allows consideration of all possible phonetic error patterns, which offers maximal relaxation over the use of ERNs as described in Section II part C. On the other hand, it is much more constrained than using free-phone recognition. This is because the mispronunciation diagnosis network is built upon the results of the first-pass of mispronunciation detection. Mispronunciation detection has pinned down the number of phones to consider in mispronunciation diagnosis in coarse resolution. Mispronunciation diagnosis serves to identify the detected errors in a higher resolution. In the worst case, should all the phones be considered mispronounced, together with multiple phone insertions, the complexity of mispronunciation diagnosis is still manageable as the length of the resulting transcription is known in advance. Unlike in standard free-phone recognition, pruning is not necessary any longer and the search can be exact.

TABLE I
PHONETIC ALIGNMENT BETWEEN THE CONVERTED MANUAL TRANSCRIPTION AND THE TRANSCRIPTION OUT OF DETECTION FOR THE WORD "WRAPPED"

1	canonical	r	ae	p	t		
2	manual	w	ae		t	ih	d
3	converted	_r	ae		t	UPM	UPM
4	detected	r	ae	UPM		UPM	UPM

TABLE II
THE MANUAL TRANSCRIPTION, TRANSCRIPTION OUT OF MISPRONUNCIATION DETECTION AND TRANSCRIPTION OUT OF MISPRONUNCIATION DIAGNOSIS FOR THE WORD "WRAPPED"

(a)	manual	w	ae		t	ih	d
	detected	_r	ae	p		UPM	UPM
	diagnosed	w	ae	p		ih	d
(b)	manual	w	ae		t	ih	d
	detected	_r	ae		t	UPM	UPM
	diagnosed	w	ae		t	ih	d

IV. PHONE-LEVEL PERFORMANCE METRICS

We propose to evaluate the performance of the two-pass framework using *phone error rate* (PER) which is the normalized aggregation of *insertion*, *deletion* and *substitution*. This is derived from the alignment between the manually-labelled transcription and the recognized transcription using dynamic programming. It is noteworthy that, in the context of MD&D, the notion of insertion, deletion and substitution can also be employed to categorize the type of phone mispronunciation made by the learner. This is derived from the alignment between the canonical transcription (1st row of Table I) and the manually-labelled transcription (2nd row of Table I). Here we point out the difference not to confuse the reader.

Based on our design of the mispronunciation detection network, the recognition transcription out of the mispronunciation detection pass may contain anti-phones and UPMs. Therefore, the manually-labeled transcription for evaluating the PER of mispronunciation detection should be converted to include anti-phones and UPMs. The conversion process aligns the canonical transcription with the manually-labeled transcription and replaces substitutions by anti-phones and insertions by UPMs. For example, as shown in the second and third rows of Table I, [r] is mispronounced as [w] and is indicated by the anti-phone [_r], and there are two inserted phones [ih] and [d] which are marked as UPMs. On the other hand, mispronunciation diagnosis does not require such conversion since it can be viewed as a constrained phone recognition problem. So, hereafter, we will be using PER1 and PER2 to denote the two metrics, used in the context of mispronunciation detection and mispronunciation diagnosis, respectively.

In mispronunciation detection, PER1 can be calculated by aligning the converted manual transcription with the recognition transcription and extracting the mismatches. An example is shown in the 3rd and 4th rows of Table I.

Furthermore, the substitution errors can be divided according to whether the mispronounced phone (appearing as an anti-phone in the converted transcription) is recognized as the canonical phone (false acceptance), or the canonical phone is recognized as an anti-phone (false rejection), leading to the two metrics, namely: *false rejection rate* (FRR) and *false acceptance rate* (FAR). Where substitution errors occur, the FRR is calculated by the number of false rejections divided by the total number of canonical phones, while the FAR is the number of false acceptances divided by the total number of anti-phones.

In mispronunciation diagnosis, PER2 is calculated by aligning the manual transcription with the recognition transcription and extracting the mismatches. An example is shown in the 1st and 3rd rows of part (a) of Table II. It is noted that the accuracy of mispronunciation diagnosis is dependent on the accuracy of mispronunciation detection. Two examples of transcriptions out of mispronunciation diagnosis are shown in Table II. In part (a) of the table, the detection pass gives a flawed result as it falsely accepts a [p] which is actually deleted by the learner, and it also fails to accept the [t] which is correctly produced. The two errors made in the detection pass are propagated to the diagnosis pass which results in erroneous diagnosis of an non-existent [p]. If the detection is more accurate as in part (b) of the table, there is a better chance of accurate diagnosis.

V. CORPUS AND EXPERIMENTAL SETUP

Experiments are set up on the Chinese University - Chinese Learners of English (CU-CHLOE) corpus - Cantonese subset [27]. The recording device is Sennheiser PC155 headset which consists of a noise-canceling uni-directional microphone and built-in sound-card. Recordings are conducted from three sites: a soundproof recording room 1 and two study rooms 2 (without sound-proofing). Speakers are selected based on the criteria that their mother tongue is the Cantonese dialect, they have learned English for at least 10 years and their English pronunciation are deemed intermediate to good by the English teachers in our university. Each speaker is asked to verify the recording quality of their utterances through playback as well as waveform visualization (e.g. to ensure no clipping). The corpus contains recordings from 100 speakers (50 male, 50 female). The data was digitized at 16-bit per sample and a sampling rate of 16 kHz. The average SNR of the recordings are 37.6dB for the sound-proof room and 36.7dB for the study rooms. The recording text prompts include: (i) "The North Wind and the Sun", a classic example of Aesop's Fables; (ii) Minimal pairs, confusable word groups and phonemic sentences that are designed by English teachers in the university. Recordings are phonetically transcribed and cross-checked by three trained linguists. We use all the TIMIT symbols for transcription except for [hv]. For the experiments in this paper, the symbols are normalized, resulting in 40 phones excluding short pause and silence. The normalized phone-set is equivalent to the CMU ARPABET plus the schwa [ax]. There are a total of 63,080 words and the lexicon size is 436 words.

The recordings in the corpus are divided into training and test sets by speakers. Since every speaker is reading the same prompt, there is no OOV in the test set. The 7.3-hour training data contain recordings from 25 male speakers and 24 female speakers. Correspondingly, the 7.8-hour test data contains recordings from the other 25 males speakers and 26 female speakers. There are 30,929 and 32,151 words in the training and test data, respectively. Among them, 48.53% and 51.02%

words are mispronounced, respectively. Since the word prompts are known, we align the the manually-labeled phone-level transcription per word token with the pronunciation extracted from the canonical pronunciation dictionary. This yields a phone-level pronunciation accuracy of 82.5% and 82.2% on the training and test sets.

Standard MFCC features are extracted and cepstral mean normalization is performed on a per-utterance basis. All the HMM model training and recognition experiments in the paper are conducted using HTK (http://htk.eng.cam.ac.uk/).

VI. Experiments

We organize our experimentation as follows: First, we build and test HMMs for the single-pass framework as a reference point. Second, after establishing the baseline HMMs for detection and diagnosis in the two-pass framework, we illustrate the issue of the competition between anti-phones and the UPM. This issue is addressed by the use of a penalty. Third, error propagation from mispronunciation detection to diagnosis is shown by examining the gap between the two-pass pipeline that has "perfect" detection, compared with one that has "imperfect" detection as in real systems. Fourth, we apply discriminative training (DT) to optimize the two sets of HMMs for detection and diagnosis respectively. The results are also compared with those of the discriminatively trained HMMs in the single-pass framework. Finally, we revisit the UPM penalty and error propagation problems as we compare the performance difference before and after DT.

A. Single-Pass Framework Setup and Baseline HMM Results

The HMMs for each phone are trained using standard Baum-Welch on all the phone segments derived from a forced-alignment on the training set, and the number of Gaussian mixtures per HMM state grows iteratively (through the "mix-up" process, i.e., mixture split then re-estimation). The number of Gaussian mixtures per state of a phone is the same across all the states of that phone. This per-phone number is linearly proportional to the occurrences of each phone in the training set, with a minimal and maximal number of 1 and 32.

As a compact representation of the error patterns, the "cheating ERN" contains the canonical pronunciations and the error patterns for every word in the test set (so there will not be recognition errors due to missing error patterns in the ERN. In other words, the ERN contains the error patterns already and correct mispronunciation detection only depends on whether the acoustic models can guide the search to find the correct path in the ERN). The results on the "cheating ERN" are usually among the best compared with those of ERNs derived from other error pattern modeling techniques[1], and can thus serve as a proper reference for the single-pass framework. A recognition pass in the ERN yields the phonetic transcription for each utterance from the test set. A direct comparison between the recognition transcription and the manual transcription gives a PER2 of 31.7%. We also align the recognition transcription from the single-pass framework with the the canonical transcription and convert it to a form which replaces insertions with UPMs and substitutions with anti-phones. The resulting transcription is compared with the converted manual transcription, yielding a PER1 of 27.2% (FRR = 14.1%, FAR = 20.0%).

TABLE III
The Rates of Insertions, Deletions and Substitutions, as Well as the PER1 of the Baseline Detection HMMs in the Two-Pass Framework, Without and With Log-Likelihood Penalty Attached to the UPMs

baseline detection	insertions	deletions	substitutions	PER1
no penalty	22.4%	7.2%	16.4%	46.0%
penalty@30	2.6%	8.50%	15.1%	26.1%

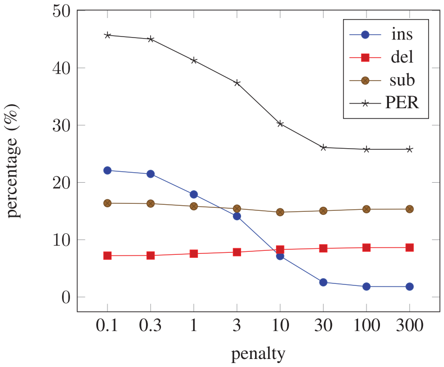


Fig. 3. Percentages of insertions, deletions and substitutions, as well as PER1 of the mispronunciation detection HMMs, with penalty added to UPM in the log-likelihood.

B. Two-Pass Framework Setup and Baseline HMM Results

The HMMs for detection consist of canonical phones, anti-phones and a UPM, while the HMMs for diagnosis contains the canonical phones only. The detection HMMs for each canonical phone and the diagnosis HMMs are obtained following the same steps as in the single-pass framework on the training set. For each canonical phone, an anti-phone HMM is built from all the non-silence phone segments that do not belong to the canonical phone in the training set using Baum Welch and undergoes the same "mix-up" process. Similarly, the UPM is built from all the non-silence phone segments. The number of mixtures per state in the anti-phones and the UPM is 64. The results including the rates of insertions, deletions and substitutions (they are calculated by dividing the individual counts by the total number of non-silence phones), as well as the PER1 are shown in Table III.

According to the second row of Table III, there is an excessive number of insertions of which about 92% are UPMs (an example is shown in the 3rd and 4th rows of Table I). This is because the acoustic space characterized by the UPM and the anti-phones largely overlap. We solve this problem by adding a log-likelihood penalty to the UPMs during decoding. The results on the test set are shown in Figure 3. Attaching increasing penalties to the UPMs can significantly reduce the chance of insertions while at the same time keeps substitutions and deletions below a reasonable level. We empirically select a penalty

[1] As a benchmark, the PER of free-phone recognition of the same acoustic model on a bi-gram phone LM is 41.9%. This performance compares with PER1 = 27.2% for the cheating ERNs.

of 30 where the curves of insertions, deletions and substitutions tend to stabilize (see Figure 3). The PER1 at this point is 26.1% (FRR = 13.8%, FAR = 20.5%) as can be found in the third row of Table III.

Based on the transcription out of the mispronunciation detection pass with the log-likelihood UPM penalty of 30, we generate the mispronunciation diagnosis networks for each utterance and perform recognition using the diagnosis HMMs. The resulting PER2 is 27.7%. An oracle experiment is also conducted based on the transcription of "perfect" detection, which gives a PER2 of 6.9%. There is a huge gap between the PER2 of the two experiments on diagnosis, which also reflects the poor performance of the baseline detection HMMs.

C. Minimum Detection Error Training and Results

To optimize the detection HMMs θ^{det} based on minimum phone error (MPE) training [32], we maximize the accuracy of phone error detection according to the following objective function:

$$\max_{\boldsymbol{\theta}^{\text{det}}} \sum_{r=1}^{R} \frac{\sum_{\tilde{\boldsymbol{s}}^r} p^{\kappa}_{\boldsymbol{\theta}^{\text{det}}}(\mathbf{O}^r|\tilde{\boldsymbol{s}}^r)p(\tilde{\boldsymbol{s}}^r)\mathcal{A}(\tilde{\boldsymbol{s}}^r,\tilde{\mathbf{w}}^r)}{\sum_{\tilde{\boldsymbol{u}}^r} p^{\kappa}_{\boldsymbol{\theta}^{\text{det}}}(\mathbf{O}^r|\tilde{\boldsymbol{u}}^r)p(\tilde{\boldsymbol{u}}^r)}, \qquad (1)$$

where $\mathbf{O}^r$, $\tilde{\boldsymbol{s}}^r$, $p^{\kappa}_{\boldsymbol{\theta}^{\text{det}}}(\mathbf{O}^r|\tilde{\boldsymbol{s}}^r)$ and $p(\tilde{\boldsymbol{s}}^r)$ are the speech frames, phonetic transcription, likelihood and prior probability, all for the rth utterance. A tilde is attached to $\boldsymbol{s}^r$ to indicate that it is from the mispronunciation detection network which may include anti-phones and UPMs. Likewise, $\tilde{\mathbf{w}}$ is the manually-labelled transcription in its converted form. $\mathcal{A}(\tilde{\boldsymbol{s}}^r,\tilde{\mathbf{w}}^r)$ denotes the number of matched phones between $\tilde{\boldsymbol{s}}^r$ and $\tilde{\mathbf{w}}^r$.

We generate mispronunciation detection lattices using the baseline detection HMMs with a UPM of 30. To control the sizes of the lattices, a 16-best token passing algorithm is employed and the beam pruning threshold is set to 400. The model is discriminatively trained for 8 iterations and the PER1 on the test set is 14.9% (FRR = 7.9%, FAR = 7.5%), which reduces PER1 from 26.1% by a relative of 42.8%.

A similar discriminative training procedure is also applied to the HMMs in the single-pass framework, where the lattices are generated from the networks spanned by all the pronunciations found in the training set. A PER1 of 15.1% (FRR = 8.1%, FAR = 9.5%) on the test set out of the "cheating ERN" is obtained by the discriminatively trained HMMs, which is slightly higher than the PER1 of 14.9% in the first-pass of the two-pass framework.

We visualize the first two cepstral coefficients of the central state of the canonical model of the consonant $[dh]$ (which is non-existent in the mother tongue of the Cantonese learners of English) and that of its anti-model before and after DT in Figures 4 and 5. The two plots correspond to the contour lines of the two central state of the GMMs. The anti-model of $[dh]$ has been adjusted to automatically capture the modes surrounding the canonical $[dh]$ in the data that does not belong to the canonical $[dh]$.

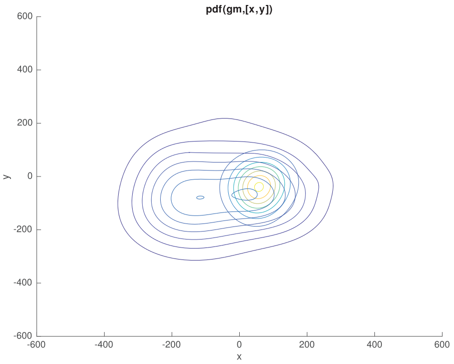


Fig. 4. The contour of the first and second cepstral coefficients of the central state of the GMMs for the canonical/anti-phones of $[dh]$, before discriminative training (DT).

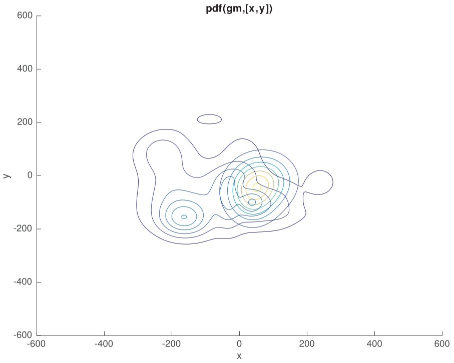


Fig. 5. The contour of the first and second cepstral coefficients of the central state of the GMMs for the canonical/anti-phones of $[dh]$, after DT.

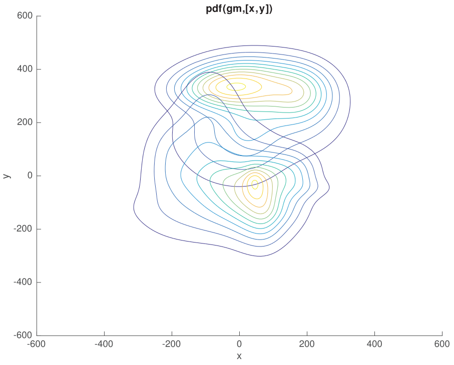


Fig. 6. The contour of the first and second cepstral coefficients of the central state of the GMMs for the canonical/anti-phones of $[iy]$, after DT.

Another such canonical/anti-phone pair on the vowel $[iy]$ is shown in Figures 7 and 6. For the pair on $[iy]$, apart from capturing some extraneous modes, the canonical phone and the anti-phone are separated as much as they can to reduce the overlap between them.

It is also interesting to note that after discriminative training, the penalty to UPM in recognition does not seem to be effective, as shown in Table IV. Since we attach such penalty when generating lattices, this is likely due to the indirect adaption of the acoustic models to the scaled UPM in the lattice during DT. Therefore, the UPM is penalized during the estimation.

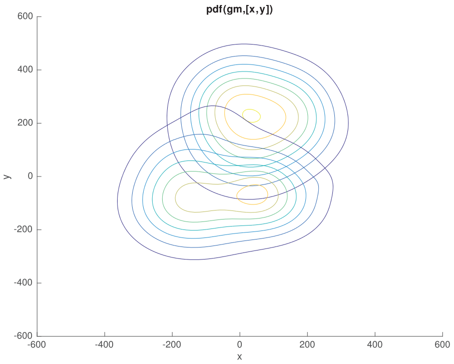


Fig. 7. The contour of the first and second cepstral coefficients of the central state of the GMMs for the canonical/anti-phones of $[iy]$, before DT.

TABLE IV
PER1 of Discriminatively Trained Detection HMMs With Different UPM Penalty

penalty	0.1	0.3	1	3	10	30
PER1	14.92%	14.92%	14.92%	14.91%	14.89%	14.93%

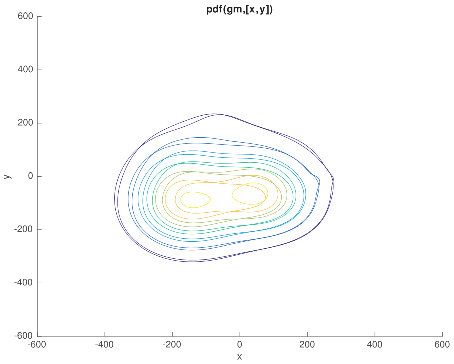


Fig. 8. The contour of the first and second cepstral coefficients of the central state of the GMMs for the UPM and the anti-phone of $[t]$, before DT.

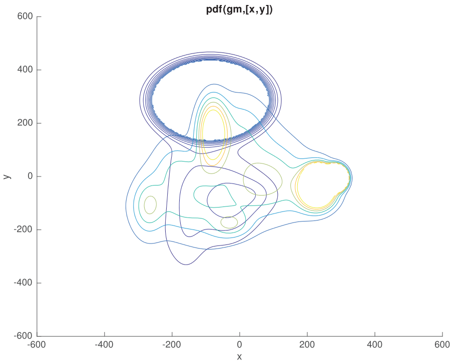


Fig. 9. The contour of the first and second cepstral coefficients of the central state of the GMMs for the UPM and the anti-phone of $[t]$, after DT.

TABLE V
The Comparison Between the Single-Pass Framework and the Two-Pass Framework, Before and After DT. "det." is Short for "detection" and "dia." is Short for "diagnosis"

Mispronunciation detection

framework	before DT		after DT	
	FRR	FAR	FRR	FAR
single-pass	14.1%	**20.0%**	8.1%	9.5%
two-pass	**13.8%**	20.5%	**7.9%**	**7.5%**

Mispronunciation detection and diagnosis

framework	before DT		after DT	
	det. (PER1)	dia. (PER2)	det. (PER1)	dia. (PER2)
single-pass	27.2%	31.7%	15.1%	17.4%
two-pass	**26.1%**	**27.7%**	**14.9%**	**16.5%**
"perfect" det.	0.0%	6.9%	0.0%	5.2%

To investigate, we visualize the first two cepstral coefficients of the central state of the HMMs, before and after DT, in Figure 8 and Figure 9. There is a huge overlap between the acoustic space covered by the UPM and that by the anti-phone of $[t]$ before DT. DT effectively separates the two spaces covered by the respective models so that the penalty to UPM is no longer necessary.

D. Minimum Diagnosis Error Training and Results

Similarly, the diagnosis HMMs θ^{dia} is optimized to maximize the accuracy of phone error diagnosis following the objective function shown below:

$$\max_{\boldsymbol{\theta}^{\text{dia}}} \sum_{r=1}^{R} \frac{\sum_{\boldsymbol{s}^r} p^{\kappa}_{\boldsymbol{\theta}^{\text{dia}}}(\mathbf{O}^r|\boldsymbol{s}^r)p(\boldsymbol{s}^r)\mathcal{A}(\boldsymbol{s}^r,\mathbf{w}^r)}{\sum_{\boldsymbol{u}^r} p^{\kappa}_{\boldsymbol{\theta}^{\text{dia}}}(\mathbf{O}^r|\boldsymbol{u}^r)p(\boldsymbol{u}^r)}, \tag{2}$$

where $\mathcal{A}(\boldsymbol{s}^r,\mathbf{w}^r)$ means the number of matched phones between $\boldsymbol{s}^r$ and $\mathbf{w}^r$.

We generate mispronunciation diagnosis lattices based on "perfect" detection of errors using the baseline diagnosis HMMs. To control the sizes of the lattices, a 32-best token passing algorithm is employed. As the mispronunciation diagnosis lattice does not grow exponentially as the mispronunciation detection lattice does, no beam pruning is applied. The model is discriminatively trained for 8 iterations. The oracle experiment based on the transcription of "perfect" detection is re-rerun on the test set.

The PER2 in the oracle experiment is 5.2% which translates to 24.8% relative reduction compared to the 6.9% result before DT. To test the discriminatively trained detection and diagnosis HMMs jointly in the two-pass framework, we take the detection transcription by the best detection HMMs obtained so far and expand the resulting detection transcription into mispronunciation diagnosis networks. The diagnosis HMMs give a PER2 of 16.5%. Compared to the 27.7% result before DT, DT of the diagnosis HMMs provides a relative reduction of 40.4%.

To compare the result of the single-pass framework with that of the two-pass framework, we calculate the PER2 of the discriminatively trained HMMs using the "cheating ERN". The PER2 is 17.4% which is slightly worse than that of the two-pass framework.

Overall, the results can be summarized in Table V. We conduct one-tailed significant test for proportions on both detection and diagnosis to test the error reduction by the two-pass framework (after DT) over the single-pass framework (after DT), at at significance level of 0.05. There is strong evidence that both passes are statistically significant.

VII. Conclusions

This paper describes a two-pass framework for mispronunciation detection and diagnosis without any prior knowledge of the error patterns, and the discriminative training of acoustic models in the framework. The first-pass of mispronunciation detection is achieved by a carefully designed network which can capture insertions, deletions and substitutions. A recognition pass over the designed network yields a transcription containing possible insertions, deletions and substitutions. Maximum likelihood training of the detection HMMs, especially the anti-phones, fails to make a sharp distinction between the canonical phones and their corresponding anti-phones, with a benchmark PER1 of 46.0%. This is because the universal phone model (UPM), which is used to account for the phone insertions, overlaps with the anti-phones greatly, so the UPM and anti-phone compete to control the speech segments accessible to both of them. The problem can be partially solved by attaching a penalty to the UPM, which leads to a PER1 reduction to 26.1%. Discriminative training (DT) of the mispronunciation detection HMMs reduces the PER1 further to 14.9%. Visualization shows that DT separates not only the canonical phones and their anti-phones but the anti-phones and the UPM as well. The second-pass of mispronunciation diagnosis, as a follow-up to mispronunciation detection, considers all possible phones for detected phone insertions and substitutions and tells their true identities through another recognition pass. The maximum likelihood baseline of the diagnosis HMMs yields a benchmark PER2 of 27.7% based on the detection result given by the baseline detection HMMs. Discriminative training of the diagnosis HMMs is performed on lattices based on "perfect" detection, as the training focuses on discriminating the correct phone and the rest of the competing diagnoses, instead of being biased towards handling insertions and deletions due to errors in detection. Overall, when conducting detection and diagnosis consecutively in the two-pass framework, discriminative training of the two sets of HMMs brings the PER2 down to 16.5%. A possible lower bound to the PER2 is 5.2% which is given by the discriminatively trained diagnosis HMMs based on "perfect" mispronunciation detection in the first pass. Future directions include experimenting on replacing the GMMs by with deep neural networks for further acoustic model replacement.

References

[1] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2, pp. 95–108, 2000.

[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," *Speech Commun.*, vol. 51, no. 10, pp. 896–905, 2009.

[3] P. Bonaventura, D. Herron, and W. Menzel, "Phonetic rules for diagnosis of pronunciation errors," in *Proc. KOVENS*, 2000, pp. 225–230.

[4] A. Neri, C. Cucciarini, and H. Strik, "ASR-based corrective feedback on pronunciation: Does it really work," in *Proc. Interspeech*, 2006.

[5] D. Herron *et al.*, "Automatic localization and diagnosis of pronunciation errors for second-language learners of english," in *Proc. 6th Eur. Conf. Speerch Commun. Technol. (Eurospeech)*, 1999.

[6] Y. B. Wang and L. S. Lee, "Towards unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning," in *Proc. IEEE Int. Conf. Acoust. Speech. Signal Process. (ICASSP)*, 2013, pp. 8232–8236.

[7] W. K. Lo, S. Zhang, and H. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in *Proc. Interspeech*, 2010.

[8] C. Molina, N. B. Yoma, J. Wuth, and H. Vivanco, "ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion," *Speech Commun.*, vol. 51, no. 6, pp. 485–498, 2009.

[9] H. Franco, H. Bratt, R. Rossier, V. R. Gadde, E. Shriberg, V. Abrash, and K. Precoda, "EduSpeak$^{®}$: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Lang. Test.*, vol. 27, no. 3, pp. 401–418, 2010.

[10] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Commun.*, vol. 45, no. 4, pp. 455–470, 2005.

[11] M. Eskenazi, "Detection of foreign speakers' pronunciation errors for second language training-preliminary results," in *Proc. 4th Int. Conf. Spoken Lang. (ICSLP)*, 1996, vol. 3, pp. 1465–1468.

[12] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Proc. Eurospeech,* 1997.

[13] F. de Wet, C. Van der Walt, and T. R. Niesler, "Automatic assessment of oral language proficiency and listening comprehension," *Speech Commun.*, vol. 51, no. 10, pp. 864–874, 2009.

[14] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proc. Eurospeech*, 1999.

[15] A. Ito, Y. L. Lim, M. Suzuki, and S. Makino, "Pronunciation error detection for computer-assisted language learning system based on error rule clustering using a decision tree," *Acoust. Sci. Technol.*, vol. 28, no. 2, pp. 131–133, 2007.

[16] K. C. Sim, "Improving phone verification using state-level posterior features and support vector machine for automatic mispronunciation detection," in *Proc. SLaTE*, 2009.

[17] J. Tepperman and S. Narayanan, "Using articulatory representations to detect segmental errors in nonnative pronunciation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 8–22, Jan. 2008.

[18] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proc. 4th Int. Conf. Spoken Lang. (ICSLP)*, 1996.

[19] K.P., Truong, "Automatic pronunciation error detection in Dutch as a second language: An acoustic-phonetic approach," MA thesis, Utrecht Univ., Utrecht, The Netherlands, 2004.

[20] Z. Ge, S. Sharma, and M. Smith, "Adaptive frequency cepstral coefficients for word mispronunciation detection," in *Proc. 4th Int. Congr. Image Signal Process. (CISP)*, 2011.

[21] H. Li, S. Huang, S. Wang, and B. Xu, "Context-dependent duration modeling with backoff strategy and look-up tables for pronunciation assessment and mispronunciation detection," in *Proc. Interspeech*, 2011.

[22] L. Gu and J. Harris, "SLAP: A system for the detection and correction of pronunciation for second language acquisition," in *Proc. Int. Symp. Circuits Syst. (ISCAS)*, 2003.

[23] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLTW)*, 2012.

[24] G. Kawai and K. Hirose, "A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, 1998.

[25] H. Meng, Y. Y. Lo, L. Wang, and W. Y. Lau, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, 2007.

[26] H. Wang and T. Kawahara, "Effective error prediction using decision tree for ASR grammar network in CALL system," *Proc. IEEE Int. Conf. Acoust. Speech. Signal Process. (ICASSP)*, 2008, pp. 5069–5072.

[27] J. Van Doremalen, C. Cucchiarini, and H. Strik, "Optimizing automatic speech recognition for low-proficient non-native speakers," *EURASIP J. Audio Speech Music Process.*, 2010, vol. 2010, Article ID 2.

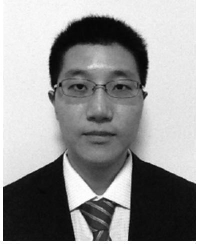
[28] W. Menzel, D. Herron, P. Bonaventura, and R. Morton, "Automatic detection and correction of non-native english pronunciations," in *Proc. INSTILL*, 2000, pp. 49–56.

[29] T. Stanley, K. Hacioglu, and B. Pellom, "Statistical machine translation framework for modeling phonological errors in computer assisted pronunciation training system," in *Proc. SLaTE*, 2011.

[30] A. Lee and J. Glass, "Mispronunciation detection without nonnative training data," in *Proc. Interspeech*, 2015.

[31] N. F. Chen, S. W. Tam, S. Wade, and J. P. Campell, "Characterizing phonetic transformations and acoustic differences across english dialects," *IEEE Trans. Audio Speech Lang. Process.*, vol. 22, no. 1, pp. 110–124, Jan. 2014.

[32] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 2005.



Xiaojun Qian (M'13) received the Ph.D. degree in systems engineering and engineering management from the Chinese University of Hong Kong, Hong Kong, and the B.E. degree in electrical engineering from Fudan University, Shanghai, China, in 2007. From 2007 to 2010, he was with the speech group of Microsoft Research Asia. His research interests include discriminative training, subspace acoustic modeling, and deep learning. He was the recipient of the 2010 Microsoft Research Asia Fellowship Award.

Helen Meng (F'15) received the S.B., S.M., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge, MA, USA. She joined the Chinese University of Hong Kong, in 1998, where she is currently a Professor and the Chairman with the Department of Systems Engineering and Engineering Management. In 1999, she established the Human-Computer Communications Laboratory and serves as the Director. In 2005, she established the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies and serves as the Co-Director. This laboratory was recognized as a Ministry of Education of China (MoE) Key Laboratory in 2008. In 2013, she helped the University establish the Big Data Decision Analytics Research Center, with a generous donation from the Dr. Stanley Ho Medical Development Foundation. She served as an Associate Dean (Research) of the Faculty of Engineering from 2006 to 2010. She has also been awarded the Peng Cheng Visiting Professorship of Tsinghua Graduate School of Shenzhen, and is a Visiting Professor at Tsinghua University, Beijing, China, Fudan University, Shanghai, China, and the Northwestern Polytechnical University. Her research interests include human-computer interaction via multimodal and multilingual spoken language systems, computer-aided language learning systems, as well as trans-lingual speech retrieval technologies. She is the Editor-in-Chief of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. She is also an elected board member of the International Speech Communication Association. She was the General Chair of the International Symposium on Chinese Spoken Language Processing 2012 and the Technical Chair of Interspeech 2014. Her international and local professional services include Steering Committee member on eHealth Record Sharing of the HKSAR Government, and Council Membership of the Research Grants Council, Hong Kong Productivity Council and Open University of Hong Kong. She also serves on the review panels of various agencies, including the Hong Kong SAR Governments Innovation and Technology Commission, Swedish Research Council European Research Infrastructure Initiative, and the National Centres of Competence in Research, Swiss National Science Foundation. She is elected into the IEEE Signal Processing Society Board of Governors in 2013. She is also a Fellow of HKCS and HKIE. She was the recipient of the MoE Higher Education Outstanding Scientific Research Output Awards in Technological Advancements, for the area of Multimodal User Interactions with Multilingual Speech and Language Technologies in 2009. In previous years, she was also the recipient of the CUHK Exemplary Teaching Award, Young Researcher Award and Service Award of the Faculty of Engineering.

Frank Soong (F'10) received the B.S. degree from the National Taiwan University, Taipei, Taiwan, the M.S. degree from the University of Rhode Island, Kingston, RI, USA, and the Ph.D. degrees from Stanford University, Stanford, CA, USA, all in electrical engineering. He is a Principal Researcher and Research Manager of the Speech Group. He joined Bell Labs Research, Murray Hill, NJ, USA, in 1982, worked there for 20 years and retired as a Distinguished Member of Technical Staff in 2001. He published extensively and coauthored more than 200 technical papers in the speech and signal processing fields. His research interests include acoustics and speech processing, including: speech coding, speech and speaker recognition, stochastic modeling of speech signals, efficient search algorithms, discriminative training, dereverberation of audio and speech signals, microphone array processing, acoustic echo cancellation, hands-free noisy speech recognition. He was also responsible for transferring recognition technology from research to AT&T voice-activated cell phones, which were rated by the Mobile Office Magazine as the best among competing products evaluated. He visited Japan twice as a Visiting Researcher: first from 1987 to 1988, to the NTT Electro-Communication Labs, Musashino, Tokyo; then from 2002 to 2004, to the Spoken Language Translation Labs, ATR, Kyoto. In 2004, he joined Microsoft Research Asia (MSRA), Beijing, China to lead the Speech Research Group. He is a Visiting Professor at the Chinese University of Hong Kong (CUHK) and the Co-Director of CUHK-MSRA Joint Research Laboratory, recently promoted to a National Key Laboratory of Ministry of Education, China. He was the Co-Chair of the 1991 IEEE International Arden House Speech Recognition Workshop. He has served the IEEE Speech and Language Processing Technical Committee of the Signal Processing Society, as a committee member and an Associate Editor of the *Transactions of Speech and Audio Processing*. He was the corecipient of the Bell Labs President Gold Award for developing the Bell Labs Automatic Speech Recognition (BLASR) software package.