A Two-pass Framework of Mispronunciation Detection & Diagnosis for Computer-aided Pronunciation Training

Xiaojun Qian*, Helen Meng* and Frank Soong†
* The Chinese University of Hong Kong, Hong Kong SAR of China
E-mail: xjqian@se.cuhk.edu.hk, hmmeng@se.cuhk.edu.hk
† Microsoft Research Asia, Beijing, China
E-mail: frankkps@microsoft.com

***Abstract*—This paper presents a two-pass framework of mispronunciation detection and diagnosis (MD&D) – detection followed by diagnosis, without the need of explicit error pattern modeling, so that the main efforts can be devoted to improving acoustic modeling by discriminative training (or by applying alternative models like neural nets). The framework instantiates a set of anti-phones and a filler model in addition to the original phone model set, and crafts a general and compact phone error detection network. The detection network guarantees full coverage of all possible error patterns while maximally exploits the constraint offered by the text prompt. Specifically, it includes anti-phones to detect substitutions, filler model to detect insertions, and skips to detect deletions, so there is no prior assumptions on the possible form of error patterns. The subsequent diagnosis step expands the detected insertions and substitutions into phone networks, after which another recognition pass reveals the true identities of the detected errors. The crux of the trick is to bring down the modeling and recognition granularity down in the detection pass. Discriminative training (DT) of the detection and diagnosis models by minimizing the two expected full-sequence phone-level errors in the respective passes brings down the overall phone-level MD&D error by a relative of 40%. In particular, visualization of models in the framework shows that discriminative training effectively separates the canonical phones and their anti-phones.**

I. Introduction

Automatic speech recognition (ASR) technologies hold much promise for an online computer-aided pronunciation training (CAPT) platform that can supplement teachers instructions with round-the-clock accessibility and individualized feedback. As one of the most useful features of CAPT platforms, mispronunciation detection refers to locating a phone that is incorrectly articulated, and involves a binary decision. Mispronunciation diagnosis proceeds further with the identification of the actual phonetic production.

Predominant approaches to MD&D extensively exploit the existing ASR framework in an off-the-shelf manner, e.g. adopting the forced-alignment, adapting the input features, expanding the pronunciation dictionary, or post-processing the ASR scores. Most of the works on phone scoring [1] and mispronunciation detection [2] using ASR do not proceed to mispronunciation diagnosis [3] which is pedagogically necessary for generating feedback [4] to second-language learners. Those with mispronunciation diagnosis rely heavily on explicit error pattern modeling by prior linguistic knowledge (e.g. [5]) or in a data-driven fashion (e.g. [6]). Error pattern modeling allows to generate a rich set of possible phonetic error patterns for a given text prompt. Under the de-facto ASR framework, forced-alignment with the the pronunciation dictionary populated by such error patterns helps reveal the true phonetic identities where mispronunciations are present. The major consideration behind such paradigm of mispronunciation diagnosis is to avoid shifting the problem towards the more intractable free-phone recognition, by properly constraining the search space in ASR. However, explicit error pattern modeling is not satisfactory when no prior knowledge is available for a given L1-L2 language pair or when it is too costly to perform error pattern derivation with [7] or without [8] labeled non-native speech. The other risks include failing to include potential error patterns in the dictionary or overfitting to idiosyncratic error patterns.

We present an out-of-the-box thinking towards mispronunciation detection and diagnosis on read speech without explicit error pattern modeling. Imposing no priors on the possible forms of error patterns means to cover them all, which amounts to exponentially many variants in the search space. Without compromising the performance of a far-from-perfect acoustic model in the intractable search space, the trick is to reduce the search space by pairing each canonical phone (in the context of a given text prompt) with an anti-phone which covers the complementary acoustic space. Recognition in this augmented network or 'sausage' detects phonetic substitutions. Similarly, insertions and deletions can also be detected by introducing a filler model and carefully designing the network topology. Once the insertions and substitutions are detected, a follow-up step 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 discriminative training to minimize the expected phone errors, which is consistent with the evaluation metrics in the detection and diagnosis framework.

The rest of the paper is organized as follows: Section 2 introduces the phonetically-labeled corpus used for acoustic

model training and evaluation, as well as the unlabeled big data for analysis; Section 3 explains the two-pass framework; Section 4 introduces the experimental setup, establishes the baseline acoustic models, compares the discriminatively-trained acoustic models with those of the baseline and illustrates the performance improvement, all on the labeled corpus; and Section 5 concludes the paper.

II. CORPRA

The Chinese University - Chinese Learners of English (CU-CHLOE) Cantonese subset consists of 15 hours of prompted speech recordings from 50 male and 50 female speakers. Phonetic transcription is provided and cross-checked by trained linguists using the CMU ARPABET plus the schwa [ax]. The corpus is divided into training and test sets by speakers. The 7.3-hour training data contains 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.

III. THE TWO-PASS FRAMEWORK

A phone substitution is the acoustic deviation from the canonical production, so it falls in the complementary acoustic space of the canonical phone. Therefore, to detect phone substitutions, we model the anti-phones, apart from the canonical phones. The concept is simple: in addition to fitting the labeled 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 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) or *filler model* which covers all the non-silence phones. The UPM is padded between each successive phones as an optional phone loop. The detection network for the example word "THE" is shown in Fig. 1.

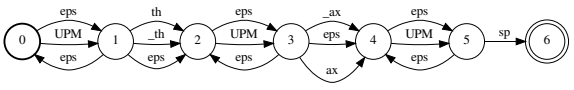

Fig. 1. Detection network for the word "THE" [th ax]. 'eps' stands for a non-emitting skip, and UPM is short for *universal phone model* which is also known as *filler model*. The anti-phones share a '_' prefix. 'sp' stands for the word-end short pause.

A recognition pass in the detection network leads to transcriptions like: [_th ax UPM]. Aligning the canonical transcription with the recognized transcription indicates the detection of the following errors: [th] is substituted, and there is a phone inserted at the end of the word.

One possible pitfall of such design of detection network, anti-phones and UPM is that the anti-phones may compete with the UPM 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.

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 each detected substitution of a canonical phone, it is expanded by all the other possible canonical phones, and for each detected UPM, it is expanded by all the possible canonical phones. Suppose the transcription from the detection is [_th ax], the expanded network for diagnosis is shown in Fig. 2.

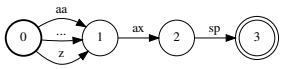

Fig. 2. A diagnosis network for the word "THE" once a substituted [th] is detected. Note that [th] does not appear in the arcs.

Compared to the standard free-phone recognition, complexity of recognition in the diagnosis network is greatly reduced. Even if in the worst case when all the phones are considered mispronounced and there are multiple phone insertions, the complexity is still manageable as the maximum length of the resulting transcription is a number known in advance. Unlike in standard free phone recognition, pruning is not necessary any longer and the search can be exact.

IV. EXPERIMENTS

A. Baseline

We extract standard 13-dim MFCC features with delta and delta-delta. Cepstral mean normalization is performed on a per utterance basis. There are two sets of HMMs for detection and diagnosis respectively. The detection HMMs for the canonical phones and the diagnosis HMMs are built following a standard 'mix-up' recipe with maximally 32 mixtures per state. 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. Similarly, the UPM is estimated using all the non-silence phone segments.

Performance evaluation in the detection pass requires a transcription which includes anti-phones and UPMs if substitution-/insertion-type of phone mispronunciations are present. To convert the manual transcription, we align the manual transcription with the canonical transcription and replace insertions by UPMs and substitutions by their corresponding anti-phones.

TABLE I
The rates of insertions, deletions and substitutions, as well as the PER of the baseline detection HMMs.

ins. (%)	del. (%)	sub. (%)	PER (%)
22.41	7.21	16.42	46.03

The results including the rates of insertions, deletions and substitutions (normalized via dividing the individual counts by the total number of non-silence phones), as well as the PER, all for the baseline detection HMMs on the test set are shown in Table I. There is an excessive number of insertions of which 92% are UPMs. This is because the acoustic space characterized by the UPM and the anti-phones overlap due to the way they are built. We solve this problem by adding log-likelihood penalty to UPMs during decoding to penalize them. The results on the test set are shown in Table II. Attaching greater and greater penalties to UPMs can significantly reduce the chance of insertions while at the same time keeps deletions below a proper level. A good balancing point is around 30.

TABLE II
The rates of insertions, deletions and substitutions, as well as the PER of the baseline detection HMMs.

penalty	0.3	3	30	300
ins. (%)	21.49	14.11	2.56	**1.81**
del. (%)	**7.24**	7.83	8.49	8.63
sub. (%)	16.31	15.43	**15.05**	15.35
PER (%)	45.04	37.38	26.10	**25.79**

Based on the transcription out of the detection pass with the log-likelihood penalty of UPM being 30, we expand them into diagnosis networks per utterance and perform recognition using the diagnosis HMMs. The resulting PER is 27.65%. An oracle experiment is also conducted based on the transcription of 'perfect' detection, which gives a PER of 6.94%. There is a huge gap between the PER of the two experiments on diagnosis, which also reflects the poor performance of the baseline detection HMMs.

B. Discriminative Training of the HMMs for Detection

We generate detection lattices on the training set using the baseline detection HMMs and the detection network with the penalty attached to UPM being 30. To control the sizes of the lattices, a 16-best token passing algorithm is employed with beam pruning. The HMMs are MPE-trained (minimum phone error [9]) for 8 iterations with a likelihood smoothing factor of 0.03. The recognition experiments are done using a UPM penalty of 30, which gives a PER of 14.93% – relative reduction by 42.8%. We visualize the first two cepstral coefficients of the central state of the canonical model of $[t]$ and that of its anti-model before and after DT in Fig. 3 and 4. The two plots are the respective GMMs' contour lines.

The anti-model of $[t]$ has been tuned to automatically capture the modes surrounding the canonical $[t]$ that does not quite belong to the canonical $[t]$. Another such canonical/anti-phone pair on the schwa $[ax]$ is shown in Fig. 5 and 6. For the pair on $[ax]$, 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 is not effective any longer, as shown in Table III. To investigate, we visualize the first two

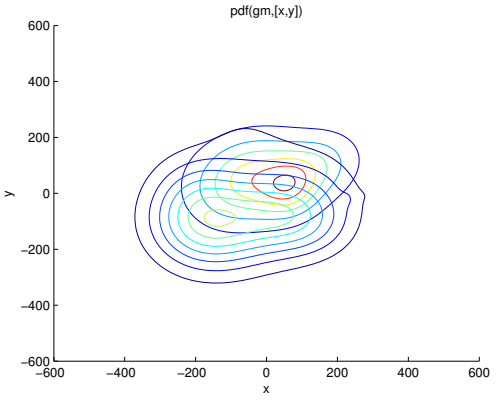

Fig. 3. *The canonical/anti-models of [t], before DT.*

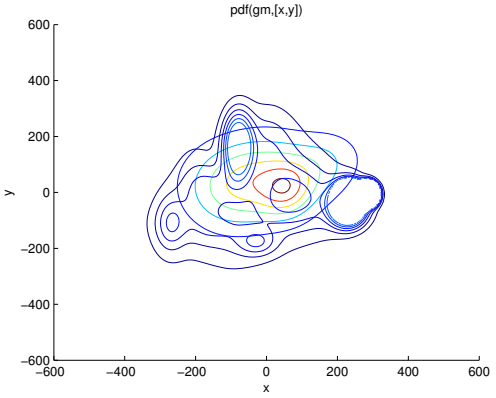

Fig. 4. *The canonical/anti-models of [t], after DT.*

cepstral coefficients of the central state of the models, before and after DT, in Fig. 7 and Fig. 8.

TABLE III
PER of MPE-trained detection HMMs with different UPM penalty.

penalty	0.3	3	30
PER (%)	14.91	14.92	14.93

As conjectured, there is a huge overlap between the acoustic space covered by the UPM and that by the anti-phone of $[s]$ before DT. DT effectively separates the two spaces covered by the respective models, which renders the penalty to UPM unnecessary. The other plausible view is that, since we attach such penalty when generating lattices, this is conceptually similar to applying a margin term to the UPM, as done in "Boosted MMI" [10]. Hence, the UPM is penalized during discriminative training.

C. Discriminative Training of the HMMs for Diagnosis

We generate diagnosis lattices on the training set using the baseline diagnosis HMMs and the diagnosis network which is based on 'perfect' detection of errors. To control the sizes of the lattices, a 32-best token passing algorithm is employed, and no beam pruning is applied. The HMMs are MPE-trained with a smoothing factors of 0.03. The oracle experiment based on

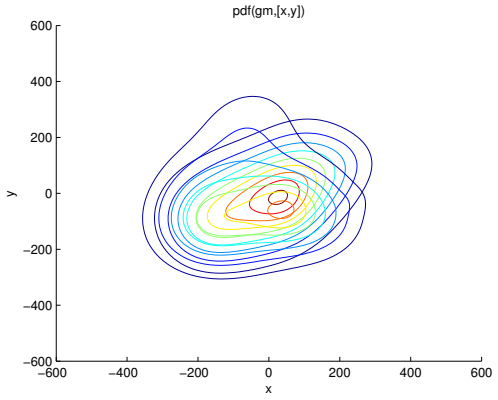

Fig. 5. *The canonical/anti-models of [ax], before DT.*

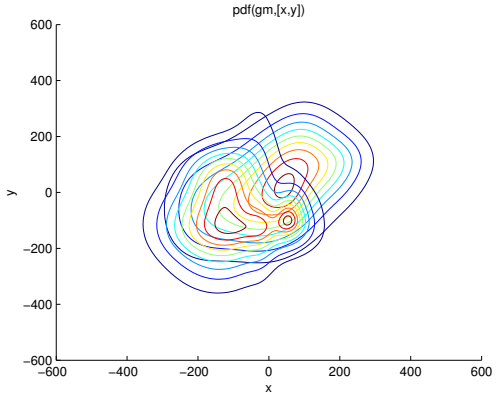

Fig. 6. *The canonical/anti-models of [ax], after DT.*

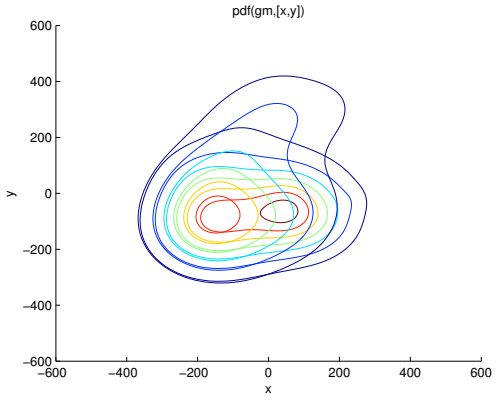

Fig. 7. *The UPM and the anti-model of [s], before DT.*

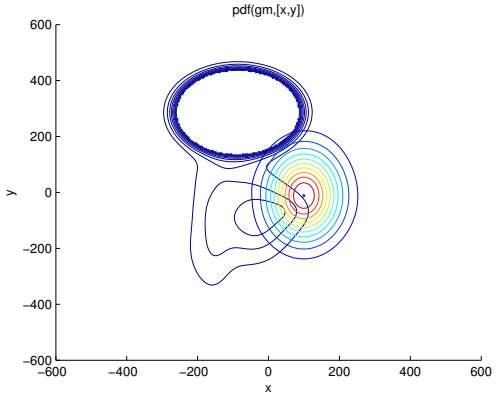

Fig. 8. *The UPM and the anti-model of [s], after DT.*

the transcription of 'perfect' detection is re-rerun, which yields a PER of 5.22% – 24.8% relative reduction compared to the 6.94% baseline. To test the discriminatively-trained detection and diagnosis HMMs jointly in the two-pass framework, we take the detection transcription by the discriminatively-trained detection HMMs and expand the resulting detection transcription into diagnosis networks. The best diagnosis HMMs gives a PER of 16.48%. Compared to the 27.65% baseline, DT of the diagnosis HMMs provides a relative reduction of 40.4%.

V. Conclusions

To achieve mispronunciation diagnosis, previous works tend to put efforts on error pattern modeling. However, error pattern modeling faces its own trade-off between under-generation and over-generation, depending on the sharpness of the acoustic model. In this paper, we demonstrate a two-pass mispronunciation detection and diagnosis framework without the need of error pattern modeling so that the main efforts can be devoted to improving the acoustic models. The framework provides full coverage of all possible pronunciation error patterns, while maximally utilizes the constraint offered by the text prompt to achieve high performance mispronunciation detection and diagnosis. The crux of the trick is to bring down the modeling/recognition granularity down in two passes – starting from a coarse model to a finer one, if the original goal turns out to be too ambitious.

References

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

[2] S. Wei, G. Hu, Y. Hu, R. H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communications*, vol. 51, no. 10, pp. 896–905, 2009

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

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

[5] D. Herron, W. Menzel, E. Atwell, R. Bisani, F. Daneluzzi, R. Morton, J. A. Schmidt, E. K. Verlag, "Automatic localization and diagnosis of pronunciation errors for second-language learners of English," *Proc. of Eurospeech*, 1999.

[6] Y.-B. Wang, L.-S. Lee, "Towards unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning," *Proc. of ICASSP*, 2013.

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

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

[9] D. Povey, "Discriminative training for large vocabulary speech recognition," *PhD. thesis*, Cambridge University, 2003.

[10] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, "Boosted MMI for model and feature-space discriminative training," *Proc. of ICASSP*, 2008.