INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Mispronunciation Detection Leveraging Maximum Performance Criterion Training of Acoustic Models and Decision Functions

Yao-Chi Hsu, Ming-Han Yang, Hsiao-Tsung Hung, Berlin Chen
Department of Computer Science and Information Engineering,
National Taiwan Normal University, Taiwan
Email: {ychsu, mh_yang, alexhung, berlin}@ntnu.edu.tw

Abstract

Mispronunciation detection is part and parcel of a computer assisted pronunciation training (CAPT) system, facilitating second-language (L2) learners to pinpoint erroneous pronunciations in a given utterance so as to improve their spoken proficiency. This paper presents a continuation of such a general line of research and the major contributions are two-fold. First, we present an effective training approach that estimates the deep neural network based acoustic models involved in the mispronunciation detection process by optimizing an objective directly linked to the ultimate evaluation metric. Second, along the same vein, two disparate logistic sigmoid based decision functions with either phone- or senone-dependent parameterization are also inferred and used for enhanced mispronunciation detection. A series of experiments on a Mandarin mispronunciation detection task seem to show the performance merits of the proposed method.

Index Terms: computer assisted pronunciation training, mispronunciation detection, discriminative training, deep neural networks

1. Introduction

The recent significant progress being made in the field of automatic speech recognition (ASR) has led to its growing applications in computer assisted pronunciation learning (CAPT). Paramount to the success of a CAPT system is the accuracy of the mispronunciation detection module, which manages to pinpoint erroneous pronunciations in the utterance of a second-language (L2) learner in response to a text prompt.

A common practice for mispronunciation detection is to extract decision features (attributes) [1] from the prediction output of acoustic models which normally are estimated based on certain criteria that maximize the ASR performance. Although hidden Markov models with Gaussian mixture models accounting for state (or senone) emission probabilities (denoted by GMM-HMM) used to be the predominating approach for building the acoustic models involved in the mispronunciation detection process, a recent school of thought is to leverage various state-of-the-art deep neural network (DNN) architectures in place of GMM for modeling the state emission probabilities in HMM (denoted by DNN-HMM) [2-4], which shows excellent promise for improving empirical performance [5-7]. As for decision feature extraction, log-likelihood, log-posterior probability and segment duration-based scores, among others [8], are frequently used in evaluating phone- [9] or word-level [10] pronunciation quality, while log-posterior probability based scores and its prominent extension, namely goodness of pronunciation (GOP) [11, 12], are the most prevalent and have been shown to correlate well with human assessments. Yet there still are a wide array of studies that capitalize on various acoustic and prosodic cues, confidence measures and speaking-style information, to name just a few, for use in mispronunciation detection. Interested readers may also refer to [13-17] for comprehensive and enjoyable overviews of state-of-the-art methods that have been successfully developed and applied to various mispronunciation detection tasks.

Our work in this paper continues this general line of research and has at least the following two major contributions: First, we present an effective learning approach that estimates the deep neural network based acoustic models involved in the GOP-based mispronunciation detection process by optimizing an objective directly linked to the ultimate evaluation metric of mispronunciation detection. Second, along the same vein, two disparate logistic sigmoid based decision functions with either phone- or senone-dependent parameterization are also estimated and employed for mispronunciation detection.

2. GOP-based Mispronunciation Detection

The general task of mispronunciation detection is to determine whether there exist mispronounced phone segments in the utterance of an L2 learner with regard the canonical phone-level pronunciations indicated by a text prompt. Given an utterance u composed of N_u phone segments, the GOP score for a phone segment $\mathbf{O}_{u,n}$, aligned to a canonical (actual) phone label $q_{u,n}$, can be defined as follows by assuming all phones share the same prior probability [11, 12, 18] :

$$\begin{aligned} \mathrm{GOP}(u,n) &= \frac{1}{T_{u,n}} \log P(q_{u,n} \mid \mathbf{O}_{u,n}) \\ &\approx \frac{1}{T_{u,n}} \log \frac{P(\mathbf{O}_{u,n} \mid q_{u,n})}{\sum_{\tilde{q} \in Q_{u,n}} P(\mathbf{O}_{u,n} \mid \tilde{q})} \end{aligned} \tag{1}$$

where $T_{u,n}$ is the duration of $\mathbf{O}_{u,n}$, $Q_{u,n}$ represents the set of acoustic models for all possible phone labels corresponding to $\mathbf{O}_{u,n}$. Alternatively, we may use the maximum operation [6, 11] to approximate the summarization operation in Eq. (1) for the sake of computational simplicity, which leads to the following formula:

$$\mathrm{GOP}(u,n) \approx \frac{1}{T_{u,n}} \log \frac{P(\mathbf{O}_{u,n} \mid q_{u,n})}{\max_{\tilde{q} \in Q_{u,n}} P(\mathbf{O}_{u,n} \mid \tilde{q})} \tag{2}$$

http://dx.doi.org/10.21437/Interspeech.2016-1602

The probabilities $P(\mathbf{O}_{u,n}\,|\,q_{u,n})$ and $P(\mathbf{O}_{u,n}\,|\,\widetilde{q})$ involved in Eqs. (1) and (2) can be calculated with either the GMM-HMM or the DNN-HMM based acoustic models, whereas the latter has demonstrated superior performance over the former in a wide range of ASR and mispronunciation detection tasks [2, 3, 5, 7].

The GOP-based score for an arbitrary phone segment $\mathbf{O}_{u,n}$, in turn, is taken as a decision feature and fed into a decision function, such as the logistic sigmoid function:

$$\mathrm{D}(u,n)=\frac{1}{1+\exp\left[\alpha(\mathrm{GOP}(u,n)+\beta)\right]} \tag{3}$$

where α and β are tunable parameters controling the shape of the decision function; the higher the value of the output, the more likely that the phone segment is mispronounced. As such, we can use the output score of the decision function in relation to a pre-established threshold to determine whether the phone segment is correctly pronounced or mispronounced.

Further, in an attempt to obtain a finer-grained inspection of the pronunciation quality of a phone segment $\mathbf{O}_{u,n}$, we can align $\mathbf{O}_{u,n}$ into a sequence of senone segments $\mathbf{O}_{u,n,i}$ in accordance with its canonical phone label $q_{u,n}$, where each senone segment may consist of one to several consecutive speech frames that belong to the same senone identity. By doing so, we can calculate the GOP score $\mathrm{GOP}(u,n,i)$ and subsequently the senone-level decision score $\widetilde{\mathrm{D}}(u,n,i)$ for each senone segment $\mathbf{O}_{u,n,i}$ involved in $\mathbf{O}_{u,n}$, using formulas defined similarly to Eqs. (1)-(3). Afterwards, an ensemble of the output scores of all senone-level decision functions for $\mathbf{O}_{u,n}$ can be taken as a more elaborate measure to determine whether $\mathbf{O}_{u,n}$ is mispronounced or not:

$$\mathrm{D}(u,n)=\frac{1}{S_{u,n}}\sum_{i=1}^{S_{u,n}}\widetilde{\mathrm{D}}(u,n,i) \tag{4}$$

where is $\mathrm{S}_{u,n}$ the total number of senone segments corresponding to $\mathbf{O}_{u,n}$.

3. Maximum F1-Score Criterion Training

In the conventional setting for GOP-based mispronunciation detection, the underlying acoustic models are normally trained with criteria that maximize the ASR performance, such as maximum likelihood (ML) estimation, minimum cross-entropy (MC) estimation and the state-level minimum Bayes risk (sMBR) estimation [3, 19-21] , to name just a few, while the parameters of the decision function are often determined empirically. In this paper, we explore to learn the DNN-HMM based acoustic models, as well as the decision function, with a discriminative objective that is directly linked to the ultimate evaluation metric of mispronunciation detection. Here we take the F1-score for investigation, since it was frequently adopted as the evaluation metric in previous work on mispronunciation detection [22-24]. Further, in this paper, the parameters of the decision function is set to be either phone- or senone-dependent when the phone-level (*cf.* Eq. (3)) or finer-grained senone-level decision functions (*cf.* Eq. (4)), respectively, are used for mispronunciation detection.

In the context of mispronunciation detection, the training objective for the maximum F1-score criterion (MFC) can be defined as follows:

$$\begin{aligned}\Xi(\boldsymbol{\theta})&=\frac{2C_{\mathrm{D}\cap\mathrm{H}}}{C_{\mathrm{D}}+C_{\mathrm{H}}}\\&=\frac{2\cdot\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{I}(\mathrm{D}(u,n))\cdot\mathrm{H}(u,n)}{[\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{I}(\mathrm{D}(u,n))]+C_{\mathrm{H}}}\end{aligned} \tag{5}$$

where $\boldsymbol{\theta}$ denotes the set of parameters pertaining to both the DNN-HMM based acoustic models and the decision functions, U is the total number of training utterances, N_u is the total number of phone segments in an utterance u, $\mathrm{C_D}$ is the total number of phone segments in the training set that were identified as being mispronounced by the current mispronunciation detection module, $\mathrm{C_H}$ is the total number of phone segments in training set that were identified as being mispronounced by the majority vote of human assessors, $\mathrm{C_{D\cap H}}$ is the total number of phone segments in the training set that are identified as being mispronounced simultaneously by both the current mispronunciation detection module and the majority vote of human assessors, and the indicator function $\mathrm{I}(\mathrm{D}(u,n))$ can be further expressed by

$$\mathrm{I}(\mathrm{D}(u,n))=\begin{cases}1 & \text{if } \mathrm{D}(u,n)\geq\tau\\ 0 & \text{otherwise}\end{cases} \tag{6}$$

where τ a pre-specified threshold. As a matter of convenience, the training objective defined in Eq. (5) can be further approximated by

$$\Xi(\boldsymbol{\theta})\approx\frac{2\cdot\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n)\cdot\mathrm{H}(u,n)}{[\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n)]+C_{\mathrm{H}}} \tag{7}$$

The training objective depicted in Eq. (7) can be viewed as a soft version of the training objective defined in Eq. (5), which in turn can be optimized using a stochastic gradient ascent algorithm, in conjunction with the chain rule for differentiation, to iteratively update the parameter set of both the DNN-HMM acoustic models and the decision function. Below we briefly highlight the procedure for maximum F1-score criterion training:

1) Train the DNN-HMM based acoustic models on the native-speaker (denoted by L1) portion of training data with the minimum cross-entropy (MC) estimation.
2) On top of the DNN-HMM based acoustic models estimated from Step 1, try to compute the decision scores of all phone segments of the training utterances (some of them contain mispronunciations) that belong to the L2 learners, where the decision function can be instantiated with either Eq. (3) or Eq. (4) and the parameters of the decision functions are empirically determined and set to be identical for all phones or senones.
3) Use the training objective introduced in Eq. (7) to iteratively update the parameters of the DNN-HMM

Table 1: *The statistical information of the speech corpus used in the mispronunciation detection experiments.*

		Duration (hours)	# Speakers	# Phone Tokens	# Errors
Training Set	L1	6.68	44	73,074	NA
	L2	15.79	63	118,754	26,434
Development Set	L1	1.40	10	14,216	NA
	L2	1.46	6	11,214	2,699
Test Set	L1	3.20	26	32,568	NA
	L2	7.49	44	55,190	14,247

Table 2: *Different structures of the DNN module in DNN-HMM.*

	# Layers	# Neurons per Layer
DNN(A)-HMM	4	1,024
DNN(B)-HMM	4	2,048
DNN(C)-HMM	6	1,024

Table 3: *ASR Experimental Results.*

	Syllable Error Rate (%)	Phone Error Rate (%)
GMM-HMM	50.9	34.3
DNN(A)-HMM	41.2	27.7
DNN(B)-HMM	40.1	27.0
DNN(C)-HMM	40.7	27.2
DNN(B)-HMM+sMBR	37.9	24.9

based acoustic models and the parameters of the phone- (Eq. 3) or senone-level (Eq. 4) decision function, with the stochastic gradient ascent algorithm and the chain rule for differentiation. Note that the estimated parameters of the decision function can be either phone (or sensone)-independent or phone (or senone)-dependent.

The notion of leveraging evaluation metric-related training criteria for training the GMM-HMM based acoustic models has recently attracted much attention in the CAPT research with some success [15, 17, 24, 25]. However, as far as we are aware, this notion has never been extensively explored for jointly training the DMM-HMM based acoustic models and decision functions.

4. Experimental Setup

4.1. Corpus and Acoustic Modeling

The dataset employed in this study is a Mandarin annotated spoken (MAS) corpus compiled by the Center of Learning Technology for Chinese, National Taiwan Normal University, between 2012 and 2014 [26]. The corpus was split into three subsets: training set, development set and test set. All these subsets are composed of speech utterances (containing one to several syllables) pronounced by native speakers (L1) and L2 learners. Each utterance of an L2 learner may contain mispronunciations, which were carefully annotated by at most four human assessors with a majority vote. Table 1 briefly highlights the statistics of the speech corpus.

Table 4: *Mispronunciation detection results achieved by using either the phone- or senone-level decision function and with or without the MFC training criterion. The acoustic models are DNN(B)-HMM.*

	Recall	Precision	F1 Score
Phone-level	0.681	0.537	0.600
Senone-level	0.675	0.545	0.603
+MFC (Both)	0.696	0.626	0.659
+MFC (AM)	0.697	0.621	0.657
+MFC (DF)	0.688	0.581	0.630

Table 5: *Mispronunciation detection results achieved by using either the phone- or senone-level decision function and with or without the MFC training criterion. The acoustic models are DNN(B)-HMM+sMBR.*

	Recall	Precision	F1 Score
Phone-level	0.671	0.551	0.605
Senone-level	0.652	0.555	0.599
+MFC (Both)	0.743	0.587	0.656
+MFC (AM)	0.738	0.586	0.653
+MFC (DF)	0.698	0.570	0.627

The ASR system was built on top of the Kaldi toolkit [27]. Each GMM-HMM based acoustic model consists of 3 states, where each state has at least 16 Gaussian mixtures. On the other hand, different structures for building the DNN-HMM based acoustic models are investigated in this paper, as shown in Table 2. For the DNN-HMM based acoustic models, the activation function of the hidden layers in the DNN module is the sigmoid function, and the activation function of the output layer is the softmax function. The ASR (free-syllable decoding without language model constraints) results on the test set (only the L1-speraker portion), using either GMM-HMM or DNN-HMM, are shown in Table 3 in terms of syllable error rate (SER) and phone error rate (PER). Since DNN(B)-HMM (w/o sMBR training) achieves the best performance and is far better than GMM-HMM, in the following experiments, the acoustic models are built on top of DNN(B)-HMM, unless otherwise stated.

4.2. Performance Evaluation

The default evaluation metric employed in this paper is the F1-score, which is a harmonic mean of precision and recall, defined as

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{Precision} = \frac{C_{\text{D}\cap\text{H}}}{C_{\text{D}}} \tag{9}$$

$$\text{Recall} = \frac{C_{\text{D}\cap\text{H}}}{C_{\text{H}}} \tag{10}$$

where C_{D}, C_{H} and $C_{\text{D}\cap\text{H}}$ were previously introduced in Section 3, but are instead counted on the test set for performance evaluation here.

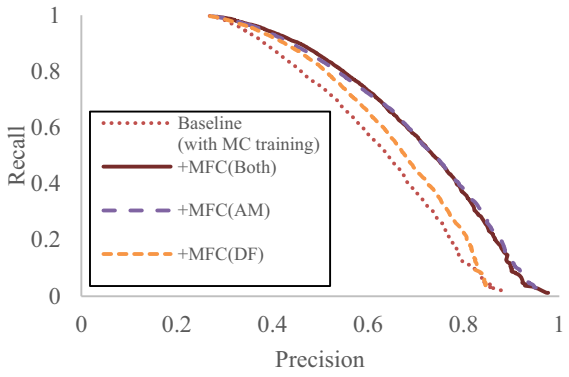

Figure 1: *Recall-precision curves for different training settings shown in Table 4 (with the senone-level decision functions).*

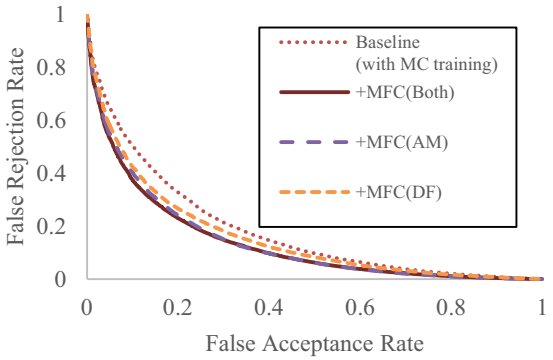

Figure 2: *ROC curves for different training settings shown in Table 4 (with the senone-level decision functions).*

5. Experimental Result

At the outset, we report on the results obtained by using either the phone-level decision functions (*cf.* Eq. (3)) or the senone-level decision functions (*cf.* Eq. (4) for mispronunciation detection. The acoustic models used here are DNN(B)-HMM trained with the minimum cross-entropy (MC) estimation, while the parameters of the decision functions are empirically tuned at optimum values based on the development set. As can be seen from the first two rows of Table 4, mispronunciation detection with the proposed senone-level decision function offers slight improvement over that using the phone-level decision function in terms of the F1-score. In particular, the precision value is increased by about 1.5% relative; this gain comes at the expense of a relatively lower recall value.

In the second set of experiments, we evaluate the utility of leveraging the MFC training criterion for estimating the parameters of the acoustic models (MFC(AM)), the decision function (MFC(DF)), or both of them (MFC(Both)), taking the senone-level decision functions for illustration. Notice here that the acoustic models are pre-trained with the MC estimation. The corresponding results are shown in the last three rows of Table 4, where two noteworthy observations can be drawn. First, all the three MFC training settings can significantly boost the mispronunciation detection performance with respect to the F1-score, as well as the recall and precision values. Especially, the F1-score is increased by about 10% relative when using the MFC(Both) training setting, indicating the effectiveness of using the MFC-based discriminative training for the mispronunciation detection task. Second, using MFC to train the acoustic models alone (MFC(AM)) seems to deliver much more performance gains than using MFC to estimate the decision functions alone (MFC(DF)), corroborating the crucial role of acoustic modeling in mispronunciation detection.

In addition, we also investigate the performance levels of using the acoustic models estimated with the conventional discriminative training criterion for ASR (i.e. sMBR; denoted by DNN(B)-HMM+sMBR), as well as its combination with the MFC training criterion. The corresponding results are depicted in Table 5. Comparing the result shown in Tables 4 and 5, we can see that even though sMBR can considerably improve the ASR performance in terms of SER and PER (*cf.* Table 3), it does not provide any additional gain for mispronunciation detection that employs either the MC-estimated acoustic models or the acoustic models further trained with the MFC criterion. The performance trends exhibited in Table 5 are quite in parallel with those shown in Table 4.

Finally, Figures 1 and 2, respectively, depict the recall-precision curves and the Receiver Operating Characteristic (ROC) curves for the aforementioned different training settings (all with the senone-level decision functions) illustrated in Table 4. Visual inspections of these two figures, again, confirm the obvious advantage of MFC. We also have observed similar trends when using some other popular metrics for performance evaluation; however, due to the space limit, we omit the details here.

6. Conclusions

In this paper, we have explored an effective maximum performance training approach for estimating the deep neural network based acoustic models, as well as the logistic sigmoid based decision functions, involved in the GOP-based mispronunciation detection process. This approach optimizes an objective that is closely related to the ultimate evaluation metric of mispronunciation detection. Furthermore, both phone- and senone-level decision functions with either phone (or sensone) dependent or independent parameterization were also investigated. Experimental evidence indeed supports the effectiveness of the proposed method. As to future work, we plan to investigate integrating more acoustic and prosodic features, as well as other different kinds of speaking-style information cues, into the process of mispronunciation detection. We also plan to develop different evaluation metric-related training criteria, in conjunction with more sophisticated DNN-HMM structures and decision functions, for use in mispronunciation detection.

7. Acknowledgements

This research is supported in part by the "Aim for the Top University Project" of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, and by the Ministry of Science and Technology, Taiwan, under Grants MOST 103-2221-E-003-016-MY2, MOST 104-2221-E-003-018-MY3, MOST 104-2911-I-003-301.

8. References

[1] W. Li, S. M. Siniscalchi, N. F. Chen and C.-H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback eith DNN-based speech attribute modeling," in Proc. ICASSP, 2016.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Processing Magazine, 29(6), pp. 82-97, 2012.

[3] D. Yu and L. Deng, "Automatic speech recognition - a deep learning approach," Springer, 2014.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, 521(7553):436–444, 2015.

[5] X. Qian, H. Meng and F. K. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," in Proc. Interspeech, 2012.

[6] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in Proc.Interspeech, 2013.

[7] W. Hu, Y. Qian, and F. Soong, "A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training," in Proc. ICASSP, pp. 3230–3234, 2013.

[8] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in Proc. Eurospeech, pp. 645–648, 1997.

[9] W. Lo, S. Zhang, and H. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in Proc. Interspeech, pp. 765–768, 2010.

[10] L. Y. Chen and J. S. R. Jang, "Automatic pronunciation scoring with score combination by learning to rank and class-normalized DP-based quantization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(11), pp. 787–797, 2015.

[11] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Communication, 30(2-3), pp. 95–108, 2000.

[12] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. H. Wang, "Automatic mispronunciation detection for Mandarin," in Proc. ICASSP, pp. 5077–5080, 2008.

[13] S. Wei, G. Hu, Y. Hu, and R. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," Speech Communication, 51(10), pp. 896–905, 2009.

[14] W. Hu, Y. Qian, , F. K. Soong, and Y. Wang, " Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," Speech Communication, 67, pp. 154–166, 2015.

[15] H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum F1-score discriminative training criterion for automatic mispronunciation detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(4), pp. 787–797, 2015.

[16] Y. B. Wang and L. S. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(3), pp. 564–579, 2015.

[17] X. J. Qian, H. Meng, and F. Soong, "A two-pass framework for mispronunciation detection & diagnosis in computer-aided pronunciation training," to appear in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016.

[18] W. Hu, Y. Qian, and F. K. Soong, "An Improved DNN-based Approach to Mispronunciation Detection and Diagnosis of L2 Learners' Speech" in Proc. SLaTE, 2015.

[19] V. Goel and W.J. Byrne, "Minimum Bayes-risk automatic speech recognition," Computer Speech & Language, 14(2), 115–135, 2000.

[20] M. Gibson and T. Hain, "Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition," in Proc. Interspeech, pp. 2406–2409, 2006.

[21] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in Proc. ICASSP, pp. 3761–3764, 2009.

[22] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in Proc. ICASSP, pp. 8227–8231, 2013.

[23] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and utilization of MLLR speaker adaptation technique for learners' pronunciation evaluation," in Proc. Interspeech, 2009, pp. 608–611.

[24] H. Huang, J. Wang, and H. Abudureyimu, "Maximum F1-score discriminative training for automatic mispronunciation detection in computer-assisted language learning," in Proc. Interspeech, pp. 815–818, 2012.

[25] X. Qian , F. Soong, and H. Meng "Discriminatively trained acoustic models for improving mispronunciation detection and diagnosis in computer aided pronunciation training (CAPT)," in Proc. Interspeech, 2010.

[26] Y. Hsiung, B. Chen, and Y. Sung, "Development of Mandarin annotated spoken corpus (MAS Corpus) and the learner corpus analysis," in Proc. WoALF, 2014.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motl´ıcek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in Proc. IEEE ASRU, 2011.