INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities

Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra, Prasanta Kumar Ghosh
Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India
{sweekars,manojkumar,chiranjeeviy,prasantg}@iisc.ac.in

Abstract

Goodness of pronunciation (GoP) is typically formulated with Gaussian mixture model-hidden Markov model (GMM-HMM) based acoustic models considering HMM state transition probabilities (STPs) and GMM likelihoods of context dependent phonemes. On the other hand, deep neural network (DNN)-HMM based acoustic models employed sub-phonemic (senone) posteriors instead of GMM likelihoods along with STPs. However, each senone is shared across many states; thus, there is no one-to-one correspondence between them. In order to circumvent this, most of the existing works have proposed modifications to the GoP formulation considering only posteriors neglecting the STPs. In this work, we derive a formulation for the GoP and it results in the formulation involving both senone posteriors and STPs. Further, we illustrate the steps to implement the proposed GoP formulation in Kaldi, a state-of-the-art automatic speech recognition toolkit. Experiments are conducted on English data collected from Indian speakers using acoustic models trained with native English data from LibriSpeech and Fisher-English corpora. The highest improvement in the correlation coefficient between the scores from the formulations and the expert ratings is found to be 14.89% (relative) better with the proposed approach compared to the best of the existing formulations that don't include STPs.

Index Terms: Goodness of pronunciation, Pronunciation evaluation, DNN-HMM acoustic model, Computer-aided pronunciation training.

1. Introduction

English is a commonly used language for business [1] and cross-cultural communications. In the process of learning English, non-native English learners could benefit from Computer-aided pronunciation training (CAPT) [2], instead of relying on a handful of available human teachers [3]. CAPT helps the non-native learners by automatically evaluating their pronunciation. There is a large number of research works on CAPT to evaluate the non-native learner's pronunciation. Most of these works assumed that acoustic properties in the learner's pronunciation are similar to a native English speaker's acoustics when their pronunciation quality is high and vice-versa. Considering this, for each phoneme's in a learner's utterance, a representative score was proposed based on two approaches – 1) likelihood of uttered phoneme given the phoneme model trained with native English speakers' acoustics [4–6], 2) posterior probability of the phoneme models given uttered phoneme speech acoustics [7], called as goodness of pronunciation (GoP). Between the two, the score computed based on GoP has been shown to be effective in most of the CAPT related works [8].

Witt et al. [7] defined GoP and computed a score from the formulated GoP using Gaussian mixture model-hidden Markov model (GMM-HMM) based native acoustic models. Following this, most of the works made improvements either by proposing variants to the GoP based formulation or by improving quality of the native acoustic models. In the works related to the former, Zhang et al. [9] used scaled log-posteriors in place of posteriors to compute the score. Luo et al. [10] formulated the GoP using the selected state sequence obtained from forward-backward algorithm. Wang et al. [11] used the GoP formulation proposed by Witt et al. [7] with error pattern detectors in phoneme mispronunciation diagnosis task. On the other hand, in the works related to improving the quality of native acoustic model, several techniques were used including discriminative training algorithms such as maximum mutual information estimation (MMIE) [12], minimum classification error (MCE) [13], minimum phone error (MPE) and minimum word error (MWE) [14]. However, the improvements with these were found to be limited [15, 16].

Recent works showed that the pronunciation evaluation based on the score computed using deep neural network (DNN)-HMM based acoustic models has a significant performance improvement compared to that using GMM-HMM based acoustic models [17]. This could be because of the better modelling strategies in DNN-HMM, which results in significant improvement in the word error rate (WER) compared to those obtained with GMM-HMM models [18]. Following this, most of the works used DNN-HMM acoustic models in the score computation. However, the DNN-HMM models involved with sub-phonemic (senone) posterior probabilities and those cannot be mapped directly with HMM state transition probabilities because each senone is shared across many states [19]. Due to this, DNN-HMM based formulations introduced variants to the GoP without considering transition probabilities [17, 20–22]. Wenping et al. computed the scores using senone posterior probability [17]. They also proposed another score by including the senone prior probabilities [20]. Further, the score as well as the features computed based on these score were used in the mispronunciation detection [23] considering transfer learning approach and pronunciation evaluation [22].

Most of the existing DNN-HMM based works heuristically neglect the transition probabilities. However, the effect of these probabilities are not explored in the score computation in the pronunciation evaluation. In order to address these, in this work, we derive a formulation for GoP under DNN-HMM based setup using both the senone posterior probabilities and transition probabilities. In addition, we show the feasibility of the proposed formulation even when each senone is shared across many states by implementing it in Kaldi [24], a state-of-the-art open resource automatic speech recognition toolkit. Experiments are conducted on the data collected from the Indian learner's considering correlation coefficient between the scores from GoP formulations and the human expert ratings as the performance measure. In the experimentation, we consider two native DNN-HMM acoustic models trained with the speech data from LibriSpeech (LS) [25] and Fisher-English (FE) [26] corpus. For the comparison, we consider three GoP formula-

We thank the Department of Science & Technology, Government of India and the Pratiksha Trust for their support.

http://dx.doi.org/10.21437/Interspeech.2019-2363

tions suggested in the work proposed by Wenping et al. [17, 20] and Huang et al. [22] as the baseline, which do not include transition probabilities in their formulation. The correlation coefficients obtained with the proposed GoP formulations are found to be 0.637 and 0.409, which are 2.91% and 14.89% higher than the best among all the three baselines using the LS and FE based native models respectively.

2. Proposed approach

2.1. DNN-HMM modeling

In a Context-Dependent (CD) DNN-HMM [27] based acoustic model, the DNN output layer represents the posterior probability, $\mathcal{P}(s|\mathbf{O})$ of each senone, (s) given the uttered acoustic observation sequence $\mathbf{O}$. The total number of output nodes in DNN is equal to the total number of senones. Each HMM represents a left-to-right three state model. HMM states encode acoustic characteristics of senones. Typically, in the left-to-right HMM, each state is connected to itself by a self-loop transition probability and to the next state by a cross-state transition probability. It is observed that in a DNN-HMM acoustic model, due to state sharing of HMMs, each senone can be associated with many state transition probabilities. Hence, it is non-trivial in DNN-HMM acoustic model to obtain a one-to-one correspondence from senone to state transition probability.

2.2. Basic GoP and its formulation

As proposed by Witt et al. [7], the GoP was defined for each phoneme p as follows:

$$GoP(p) = \frac{1}{T}\Big|\log \mathcal{P}(p|\mathbf{O})\Big| \tag{1}$$

which is the duration normalized absolute log of posterior probability of phoneme p given the acoustic observation sequence $\mathbf{O} = \{O_t, \forall\ 1 \leq t \leq T\}$ belonging to a speech segment of the phoneme p, where, T is the total number of frames in the phoneme segment. They showed the effectiveness of the GoP for pronunciation evaluation formulated with GMM-HMM acoustic model as follows:

$$\frac{1}{T}\Big|\log \mathcal{P}(p|\mathbf{O})\Big| = \frac{1}{T}\left|log\frac{\mathcal{P}(\mathbf{O}|p)\mathcal{P}(p)}{\sum_{q\in Q}\mathcal{P}(\mathbf{O}|q)\mathcal{P}(q)}\right| \tag{2}$$

where Q is the complete phoneme set, $\mathcal{P}(p)$ is the prior of phoneme p and $\mathcal{P}(\mathbf{O}|p)$ is the likelihood of acoustic segment $\mathbf{O}$ given phoneme p. Typically, this equation is formulated based on forward-backward algorithm [20].

However, in the recent past, DNN-HMM acoustic models were shown to be effective than the GMM-HMM models in speech recognition. Considering this and the effectiveness of the GoP in the pronunciation evaluation, different approximated formulations were proposed to implement the GoP using DNN-HMM models. This could be because of the complexity involved in formulating likelihood at the phoneme level as in Equation 2, in terms of senone posterior and transition probabilities. Instead, in this work, we derive a formulation for GoP as defined in Equation 1 in terms of senone posterior and transition probabilities and show that it can be implemented in Kaldi toolkit. We also provide a python wrapper to compute the score from the GoP formulation considering DNN-HMM models from the toolkit[1]. Availability of this open-source implementation would contribute to the CAPT related research.

[1] https://github.com/sweekarsud/Goodness-of-Pronunciation

2.3. Proposed formulation

Similar to the existing works on the GoP formulation using DNN-HMM models, we obtain the senone sequence $\mathbf{S} = \{s_t, \forall\ 1 \leq t \leq T\}$ in a phoneme segment p with observation sequence O using forward-backward algorithm [17]. Assuming that the senone sequence $\mathbf{s}$ is known, the posterior probability $\mathcal{P}(p|\mathbf{O})$ in the GoP in Equation 1 can be written in terms of the senone sequence $\mathbf{s}$ and acoustic sequence $\mathbf{O}$ as follows:

$$\mathcal{P}(p|\mathbf{O}) = \mathcal{P}(\mathbf{s}|\mathbf{O}) = \mathcal{P}(s_1, s_2, ..., s_T|O_1, O_2, ..., O_T) \tag{3}$$

In the typical left-to-right HMM, the current state only depends on the previous state and the current observation is associated only with the current state. Considering this left-to-right assumption [28], Equation 3 can be simplified as follows:

$$\begin{aligned}\mathcal{P}(p|\mathbf{O}) &= \overbrace{\mathcal{P}(s_1|O_1, O_2, ..., O_T)}^{\mathcal{P}(s_1|O_1)}\overbrace{\mathcal{P}(s_2|O_1, O_2, ..., O_T, s_1)}^{\mathcal{P}(s_2|O_2,\, s_1)} \\ &\quad \mathcal{P}(s_3, s_4, ..., s_T|O_1, O_2..., O_T, s_1, s_2) \\ &= \mathcal{P}(s_1|O_1)\prod_{t=2}^{T}\mathcal{P}(s_t|O_t, s_{t-1})\end{aligned} \tag{4}$$

The product term in Equation 4, $\mathcal{P}(s_t|O_t, s_{t-1})$ can be written as:

$$\mathcal{P}(s_t|O_t, s_{t-1}) = \frac{\mathcal{P}(s_t, O_t|s_{t-1})}{\mathcal{P}(O_t|s_{t-1})} \tag{5}$$

The numerator in Equation 5 can be rewritten as $\mathcal{P}(s_t|s_{t-1})\mathcal{P}(O_t|s_t, s_{t-1})$. Considering the assumption of the current acoustic observation (O_t) is associated with only the current state, $\mathcal{P}(O_t|s_t, s_{t-1})$ and $\mathcal{P}(O_t|s_{t-1})$ (the denominator in Equation 5) can be written as $\mathcal{P}(O_t|s_t)$ and $\mathcal{P}(O_t)$ respectively. Following this, we apply Bayes' rule which results in Equation 6.

$$\mathcal{P}(s_t|O_t, s_{t-1}) = \mathcal{P}(s_t|s_{t-1})\frac{\mathcal{P}(O_t|s_t)}{\mathcal{P}(O_t)} = \mathcal{P}(s_t|s_{t-1})\frac{\mathcal{P}(s_t|O_t)}{\mathcal{P}(s_t)} \tag{6}$$

Substituting Equation 6 in Equation 4 gives the following equation:

$$\mathcal{P}(p|\mathbf{O}) = \frac{\prod_{t=1}^{T}\mathcal{P}(s_t|O_t)\prod_{t=2}^{T}\mathcal{P}(s_t|s_{t-1})}{\prod_{t=2}^{T}\mathcal{P}(s_t)} \tag{7}$$

In Equation 6 and 7, the term $\mathcal{P}(s_t)$ is the prior probability of senone. Assuming all senones are equally likely and incorporating Equation 7 in Equation 1 results in as:

$$\begin{aligned}GoP(p) = \frac{1}{T}\Bigg[&\sum_{t=1}^{T}\log \mathcal{P}(s_t|O_t) \\ &+ \sum_{t=2}^{T}\log \mathcal{P}(s_t|s_{t-1}) + (T-1)\log n\Bigg]\end{aligned} \tag{8}$$

where n is the total number of senones in the DNN-HMM acoustic model. From the equation, it is observed that the GoP formulation involves both the senone posterior probabilities and transition probabilities.

2.4. Relation with existing DNN-HMM based works

In the literature, there are three major works that formulate the GoP based on DNN-HMM acoustic model. From each of these

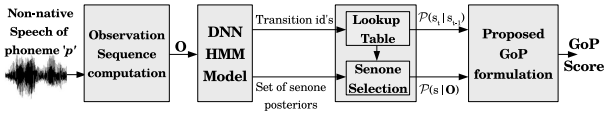

Figure 1: *Block Diagram of Proposed GOP method Implementation*

works, we consider the equation that has the highest performance and discuss the relation between them and the proposed GoP formulation. We also compare the effectiveness of the proposed GoP formulation with them.
Wenping et al. [17] formulated the GoP as:

$$GoP(p) = \frac{1}{T}\left[\sum_{t=1}^{T}\log \mathcal{P}(O_t|p) - \max_{\{q\in Q, q\neq p\}}\sum_{t=1}^{T}\log \mathcal{P}(O_t|q)\right] \quad (9)$$

The term $\mathcal{P}(O_t|p)$ is the frame based likelihood of acoustic observation O_t given phoneme model p. The second term in the equation is heuristically simplified from the denominator in Equation 2. Moreover, it is easy to observe that in Equation 9 there is no transition probability involved in the formulation unlike the proposed GoP. The same authors in another work [20] have proposed the GoP formulation considering senone posteriors and priors as:

$$GoP(p) = \frac{1}{T}\sum_{t=1}^{T}\log \frac{\mathcal{P}(s_t|O_t)}{\mathcal{P}(s_t)} \quad (10)$$

Further, Huang et al. [22] considered the GoP formulation as below proposed by Wenping et al. [21]:

$$GoP(p) = \frac{1}{T}\sum_{t=1}^{T}\log \mathcal{P}(s_t|O_t) \quad (11)$$

Comparing 10 and 11, Equation 10 is related to the proposed GoP formulation without considering transition probabilities and no assumption on senone prior was considered as in Equation 8. Further, Equation 11 is obtained by removing senone priors and transition probabilities.

2.5. Implementation aspects

In order to implement the proposed GoP formulation using Kaldi toolkit, we follow the four steps illustrated in Figure 1. In the first step, we compute 40 dimensional mel frequency cepstral coefficients (MFCC) and 100 dimensional i-vector from non-native speech signal of a phoneme segment and consider it as the acoustic observation sequence. In the second step, we obtain posterior probabilities for all set of senones in the acoustic model with DNN based script[2]. Further, for the spoken utterance, we obtain senone sequence that are encoded in a sequence of fine grained HMM states, typically known as transition-id's [29], using forward-backward algorithm with aligner script[3]. In the third step, we decode the senone sequence and its respective transition probabilities from the transition-id sequence using a look-up table which contains the mapping between transition id's and senones. We obtain the look-up table using show transition script[4]. Following this, we obtain posterior probabilities for the decoded senone sequence selected from the senone posterior probabilities. In the fourth step, we obtain the proposed GoP formulation in Equation 8 using the transition and selected senone probabilities sequence.

[2]https://github.com/kaldi-asr/kaldi/blob/master/src/nnet2bin/nnet-am-compute.cc

[3]https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/online/nnet2/align.sh

[4]https://github.com/kaldi-asr/kaldi/blob/master/src/bin/show-transitions.cc

3. Database

In this work, we consider a read English corpus collected from 16 Indian learners who were in the spoken English training at the time of the recording. Due to the language diversity in India, we consider the learners from six different native languages – Malayalam (MAL), Kannada (KAN), Telugu (TEL), Tamil (TAM), Hindi (HIN) and Gujarati (GUJ). There are a total of 4 (3+1), 5 (1+4), 3 (2+1), 2 (2+1), 1 (0+1) and 1 (0+1) speakers (male + female) from each of these languages respectively. All the learners were either undergraduate or postgraduate students whose age ranges from 19 to 25. Each learner read 800 stimuli out of which 415 are single word stimuli and 385 are multiple word stimuli. Thus, a total of 12800 utterances are present in the corpus. A spoken English expert manually rated each utterance on a scale of 5 to 1, where the rating 5, 4, 3, 2 and 1 indicates there is negligible, low, average, considerable and high native language influence in the learners pronunciation respectively. The expert is a spoken English trainer with an experience of 25 years. In all the ratings of 12800 utterances, 2585, 2656, 2957, 2364 and 2238 utterances are assigned with rating 1, 2, 3, 4 and 5 respectively. Further, in order to know the consistency of the expert, we randomly repeat 1200 utterances. The expert is found to have more than 70% consistency in the ratings of repeated stimuli.

Further, for learning native acoustic models, we use speech data from LS and FE corpora. The LS corpus contains 960 hours of data in the train set. The LS data is read speech recorded from 2238 native English speakers. On the other hand, the FE data is telephonic conversational speech recorded from 12401 native English speakers. This data contains 1600 hours in the train set.

4. Experimental results

4.1. Experimental setup

We consider the GoP formulations in Equation 9, 10 and 11 taken from the existing works as the baseline formulations, referred to as BL-1, BL-2 and BL-3 respectively. We consider Pearson correlation coefficient [30] between the scores from the GoP formulations and expert ratings as the measure. In general, the scores are defined at phoneme level [7]. However, in this work, we require a score representing the entire utterance containing single and multiple words. Following the work by Wenping et al. [17] we compute the scores for single word utterances by averaging the scores of all phones in the word. Similarly, the scores for multiple word utterances are computed by averaging the scores of all the words in the utterance. We consider two DNN-HMM based acoustic models separately trained with LS and FE data from LS Corpus [25] and FE Corpus [26]. We use Kaldi toolkit [24] to train both the DNN-HMM acoustic models considering the architecture as provided in Dan's (Daniel Povey) recipe [31]. From these models, it is observed that the DNN output dimensions are found to be 5745 and 7864 respectively in the LS and FE based acoustic models.

4.2. Results and discussion

Table 1 shows the correlation coefficient computed between the expert ratings and the scores from GoP formulations obtained from the three baselines and the proposed approach. The correlation coefficients are computed considering male (M), female (F) and all (A) the speakers respectively using two native acous-

tic models trained with LS and FE data respectively. From the table, it is observed that the correlation coefficient obtained from the proposed formulation is higher than that from all the three baselines for both the acoustic models across male, female and all speakers. This indicates that the proposed GoP formulation is better than all the three baseline formulations.

Table 1: *Correlation coefficient between the scores obtained from the GoP formulations and the expert ratings with different acoustic models considering male (M), female (F) & all (A) the speakers*

	BL-1		BL-2		BL-3		PA	
	LS	FE	LS	FE	LS	FE	LS	FE
M	0.468	0.305	0.623	0.358	0.637	0.401	**0.653**	**0.452**
F	0.434	0.266	0.593	0.306	0.605	0.343	**0.624**	**0.396**
A	0.453	0.273	0.606	0.316	0.619	0.356	**0.637**	**0.409**

The correlation coefficient is higher for the proposed approach and this could be because the proposed approach considers transition probabilities and senone state posteriors. The lower correlation coefficient obtained by Baseline 2 & 3 compared to the proposed approach could be because both Baseline 2 & 3 ignores the transition probabilities. This indicates the importance of transition probabilities in GoP formulation. Additionally, a higher correlation coefficient for Baseline 3 compared to the other two baselines could be because Baseline 3 ignores senone priors unlike Baseline 2 which considers senone priors. It is also observed that Baseline 1 has got the least correlation coefficient among the three baselines and the proposed approach and this indicates that using the senone posterior based GoP formulation yields better results than using the likelihood based GoP formulation. From the table, it is observed that the correlation coefficients are higher for all the three baselines and proposed method in the case of acoustic model trained with LS than FE data. This could be because LS corpus is recorded in read speech condition which matches with data considered in this work which is also recorded in read speech condition. However, FE corpus is recorded in telephonic conversational condition and hence a lower correlation coefficient is observed. It is interesting to observe that the correlation coefficient is higher for all the male speakers' speech.

Table 2: *Correlation coefficient between the scores obtained from the GoP formulations and the expert ratings with different acoustic models for multiple words (MW) and single words (SW)*

	BL-1		BL-2		BL-3		PA	
	LS	FE	LS	FE	LS	FE	LS	FE
MW	0.4687	0.3603	**0.5286**	0.3913	0.5283	0.4002	0.5210	**0.4099**
SW	0.5229	0.4314	0.6111	0.4914	0.6263	0.5015	**0.6272**	**0.5072**

Table 2 shows the correlation coefficient computed between the expert ratings and the scores from GoP formulations obtained from the three baselines and the proposed approach. The correlation coefficients are computed considering multiple words and single word separately using two native acoustic models trained with LS and FE data respectively. From the table, it is observed that the correlation coefficient obtained from the proposed approach is higher than that from all the three baselines for both the acoustic models considering single word. This indicates that the proposed approach is better than all the three baseline schemes for single word. It is also observed that for multiple words, the correlation coefficient of the proposed approach is comparable with Baseline 2 & 3. This indicates the benefit of using the proposed approach for single and multiple words. From the table, it is also observed that the correlation coefficient is greater for single word than multiple words for all the three baselines and the proposed formulation. This could be because, in multiple word level scoring, the scores obtained from all the words are considered with uniform weights; however, that may not be the best strategy.

Table 3: *Correlation coefficient between the scores obtained from the GoP formulations and the expert ratings with different acoustic models calculated across different native languages*

	BL-1		BL-2		BL-3		PA	
	LS	FE	LS	FE	LS	FE	LS	FE
MAL	0.425	0.221	0.585	0.289	0.606	0.334	**0.631**	**0.394**
KAN	0.421	0.241	0.592	0.271	0.605	0.317	**0.621**	**0.368**
TAM	0.442	0.230	0.603	0.281	0.619	0.340	**0.650**	**0.418**
TEL	0.515	0.344	0.663	0.409	0.671	0.436	**0.679**	**0.475**
HIN	0.439	0.312	0.554	0.329	0.563	0.359	**0.584**	**0.412**
GUJ	0.398	0.275	0.551	0.241	0.550	0.271	**0.561**	**0.316**

Table 3 shows the correlation coefficient computed between the expert ratings and the scores from GoP formulations obtained from the three baselines and the proposed approach. The correlation coefficients are computed considering six different native language having sixteen different speakers using two native acoustic models trained with LS and FE data respectively. From the table, it is observed that the correlation coefficient obtained from the proposed approach is higher than that from all the three baselines for both the acoustic models across six different native languages. This indicates that the proposed method is better than all the three baseline formulations irrespective of the native language of the learners. In the table, it is observed that the correlation coefficient for all the six native languages across all the speakers are closer to the correlation coefficients mentioned in Table 1. However, in the case of HIN and GUJ native language speakers it is observed that the correlation coefficient is not similar to the overall correlation coefficient mentioned in Table 1 and this could be because both the native language speakers are female and, as already stated, female speakers have a lower correlation coefficient compared to the male speakers.

5. Conclusion

Considering the basic GoP definition, we derive a formulation for implementing in DNN-HMM based native acoustic model, which involves senone posterior probabilities and HMM state transition probabilities. Unlike the existing works, we observe that the derived formulation results in the product of senone posterior and transition probabilities. Further, to address the non-triviality of one-to-one correspondence between senones and the states, we illustrate the step-by-step implementation of the proposed GoP formulation. Experiments with the non-native English data collected from Indian learners reveal that the scores computed from the proposed GoP formulation better correlates with expert ratings compared to that from three baselines. Further investigations are required to develop better strategies for improving the performance under mismatched speech conditions in the native and non-native data. Future works also include analysis of trade-offs between the improvements and computational efforts involved in the proposed and the baseline approaches using multiple speech corpora.

6. References

[1] B. Seidlhofer, "Common ground and different realities: World Englishes and English as a lingua franca," *World Englishes*, vol. 28, no. 2, pp. 236–245, 2009.

[2] M. C. Pennington, "Computer-Aided Pronunciation Pedagogy: Promise, Limitations, Directions," *Computer Assisted Language Learning*, vol. 12, no. 5, pp. 427–440, 1999.

[3] M. A. Salteh and K. Sadeghi, "Teachers and Students Attitudes Toward Error Correction in L2 Writing," *The Journal of Asia TEFL*, vol. 12, no. 3, pp. 1–31, 2015.

[4] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1997, pp. 1471–1474.

[5] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic Detection Of Phone-level Mispronunciation For Language Learning," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[6] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[7] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[8] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *Proc. ISADEPT*, vol. 6, 2012.

[9] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for Mandarin," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 5077–5080.

[10] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and Utilization of MLLR Speaker Adaptation Technique for Learners' Pronunciation Evaluation," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[11] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting Error patterns with empirical analysis for Computer-Aided Pronunciation Training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5049–5052.

[12] L. R. Bahl, P. F. Brown, P. V. De Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *ICASSP*, 1986, pp. 231–234.

[13] B.-H. Juang, W. Hou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio processing*, vol. 5, no. 3, pp. 257–265, 1997.

[14] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–105.

[15] K. Yan and S. Gong, "Pronunciation Proficiency Evaluation based on Discriminatively Refined Acoustic Models," *International Journal of Information Technology and Computer Science*, vol. 3, no. 2, pp. 17–23, 2011.

[16] X. Qian, F. K. Soong, and H. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT)," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[17] W. Hu, Y. Qian, and F. K. Soong, "A New Dnn-based High Quality Pronunciation Evaluation for Computer-Aided Language Learning (CALL)." in *Interspeech*, 2013, pp. 1886–1890.

[18] J. Pan, C. Liu, Z. Wang, Y. Hu, and H. Jiang, "Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMS in acoustic modeling." in *ISCSLP*, 2012, pp. 301–305.

[19] M.-Y. Hwang and X. Huang, "Shared-distribution hidden Markov models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 414–420, 1993.

[20] W. Hu, Y. Qian, and F. K. Soong, "An Improved DNN-based Approach to Mispronunciation Detection and Diagnosis of L2 Learners' Speech." in *SLaTE*, 2015, pp. 71–76.

[21] 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*, vol. 67, pp. 154–166, 2015.

[22] H. Huang, H. Xu, Y. Hu, and G. Zhou, "A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. 3165–3177, 2017.

[23] Y. Xiao, F. K. Soong, and W. Hu, "Paired Phone-Posteriors Approach to ESL Pronunciation Quality Assessment," pp. 1631–1635, 2018.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[26] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text." in *LREC*, vol. 4, 2004, pp. 69–71.

[27] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2012.

[28] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice Hall, 1993.

[29] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiát, S. Kombrink, P. Motlíček, Y. Qian *et al.*, "Generating exact lattices in the WFST framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4213–4216.

[30] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson Correlation Coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[31] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of DNNs with Natural Gradient and Parameter Averaging," *arXiv preprint arXiv:1410.7455*, 2014.