SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education
20-21 September 2019, Graz, Austria

Noise robust goodness of pronunciation measures using teacher's utterance

Sweekar Sudhakara[1], *Manoj Kumar Ramanathi*[1], *Chiranjeevi Yarra*[1],
Anurag Das[2], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
[2]Computer Science and Engineering, Texas A&M University - College Station, USA

{sweekars[1],manojkumar[1],chiranjeeviy[1],prasantg[1]}@iisc.ac.in, das.anurag2012@gmail.com[2]

Abstract

In the applications of computer-aided pronunciation training (CAPT), evaluation of second language learner's pronunciation is an important task. For this task, goodness of pronunciation (GoP) is shown to be effective and is typically computed under clean speech conditions. However, in real scenarios, CAPT systems often need to deal with noisy conditions, which could degrade the effectiveness of GoP. We analyze the variations in GoP performance under noisy conditions by adding three types of noises namely, babble, white and f-16 at 20 dB, 10 dB and 0 dB signal-to-noise ratio (SNR) conditions. We hypothesize that the use of phonemes uttered by a teacher would make GoP score more robust and mimic the human rating closely, based on which we propose a modification to the typical lexicon based GoP (LGoP). The proposed scheme is referred as teacher utterance based GoP (TGoP). In addition, GoP of learner's and teacher's utterances are combined to propose a GoP like (GL) score based on the difference between the two. Correlation coefficient between the GoPs and the teacher's ratings is used as the performance metric. Experiments conducted on the speech data collected from Indian English learners reveal that, although the performance of different GoP schemes drops with additive noise, TGoP performs better than LGoP in both clean and noisy conditions. In low SNR conditions, GL performs better than both TGoP and LGoP.

Index Terms: Goodness of pronunciation, Computer-aided pronunciation training, Lexicon based GoP, Teacher utterance based GoP, GoP like score, Noise analysis for GoP.

1. Introduction

English is commonly known as the lingua franca of business [1]. With the growing significance of learning English, computer-aided pronunciation training (CAPT) [2] could help non-native English learner's in terms of its availability and interactivity. In these related applications, the learner's utterance is automatically evaluated and feedback on their mispronunciation is provided either at phoneme level [3], word level [4] or sentence level [5]. Generally, the pronunciation evaluation in these applications is based on an assumption that the language learner shares similar acoustic properties as that of a native English speaker when the learner achieves a good pronunciation quality. Based on this, the pronunciation quality was initially evaluated by computing the likelihood [6,7] of the phonemes in a learner's utterance using the acoustic model trained with native English speech. Later posterior probability based method known as goodness of pronunciation (GoP) [3] was introduced, which is defined as the probability of acoustic observations within the uttered phoneme given its respective phoneme model. The latter is the most commonly used method in CAPT [8] due to its effectiveness.

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

GoP was defined by Witt et al. [3], where it was computed using Gaussian mixture model-hidden Markov model (GMM-HMM) based native phoneme models and was also implemented by Luo et al. [9] and Wang et al. [10] with slight modifications in the formulation. GoP was later introduced for the deep neural network (DNN)-HMM based acoustic models by Wenping et al. [11–13] and Huang et al. [14] which resulted in a significant performance improvement compared to that using GMM-HMM based acoustic models.

The existing GoP formulations have been mostly implemented on clean speech data [15]. However, in real scenarios, CAPT systems often need to handle noisy condition [16]. For example, the real scenarios could include babble noise, improper setting of the microphone, etc. Under these conditions, the existing GoP methods may fail to perform reliably. However, very few works have addressed the computation of GoP under noisy conditions. These include applying denoising to the noisy signal before computing GoP [16]. But, the mismatches in the speech acoustics considered in the phoneme model and denoised signal could affect the quality of the GoP. Thus, it is required to analyze the performance of GoP under different noisy conditions considering various types of noises at different signal-to-noise ratio (SNR) conditions.

Further, in most of the existing works, the GoP scores are obtained for the phonemes present in the learner's utterance. However, the uttered phonemes could be erroneous due to the typical phoneme errors. Thus, the GoP score computed considering incorrectly uttered phoneme could degrade the correlation coefficient. We observe that this degradation can be reduced by computing GoP considering phonemes in the teacher's utterance as a reference. Xiao et al. [17] proposed a feature by augmenting the phoneme posterior probabilities of learner's and teacher's utterance for pronunciation assessment in a supervised manner. However, no work in the literature proposed incorporating teachers' utterance into GoP computation and studied its effectiveness in clean as well as noisy conditions.

In order to analyze these, in this work, we propose to compute the GoP considering phonemes in the teacher's utterance defined as the teacher's utterance based GoP (TGoP). Further, we propose a GoP like (GL) score considering GoP scores computed from learner's and teacher's utterances. For which, a mapping function is proposed to combine both the GoP scores of teacher and learner to a GL score. Also, we analyze the performance of the proposed TGoP and GL score along with typical lexicon based GoP (LGoP) scores obtained from existing works under additive noise conditions. We conduct experiments on speech data collected from Indian English language learners considering the phoneme models trained with native English corpus named LibriSpeech (LS) [18]. We compute GoP scores based on the works proposed by Witt et al. [3], Wenping et al. [11,13] and Sudhakara et al. [19] as baseline schemes. We perform the experiments under additive noise conditions with three noises namely, babble, white and cockpit (f-16) at clean

10.21437/SLaTE.2019-13

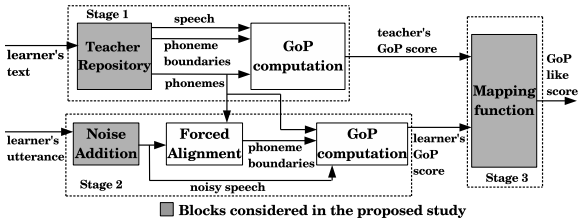

Figure 1: *Block diagram describing the steps involved in the proposed study.*

and 20 dB, 10 dB and 0 dB SNR conditions. These experiments revealed that the correlation coefficient computed between GoP scores and teacher's ratings reduce significantly from clean to 0 dB SNR under all noise conditions. The highest absolute improvements in the correlation coefficient from baseline schemes are found to be 0.030 and 0.266 among all clean and noisy conditions with TGoP and GL score respectively.

2. 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 learners from six different native languages – Malayalam, Kannada, Telugu, Tamil, Hindi and Gujarati. There are a total of 4, 5, 3, 2, 1 and 1 speakers from each of these languages respectively. All the learners were either undergraduate or postgraduate students whose age ranged from 19 to 25 years. Each learner read 800 stimuli, thus, a total of 12800 utterances are present in the corpus. A spoken English teacher manually rated each utterance on a scale of 5 to 1, where the rating 5, 4, 3, 2 and 1 indicate that there is negligible, low, average, considerable and high native language influence in the learner's utterance respectively. The teacher 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, to know the consistency of the teacher, we randomly repeat 1200 utterances. The teacher is found to have more than 70% consistency in the ratings of repeated stimuli. Further, we obtain the teacher's utterances by collecting the recordings from the teacher for all the unique sentences spoken by the learners. We use three noises namely, babble, white Gaussian, f-16 from the NOISEX-92 database [20].

3. Proposed study

Figure 1 shows the three major stages involved in the GoP score computation of the proposed study. The first stage computes the GoP score for the teacher's utterance in two steps and for this purpose, a repository is created which consists of teacher's recordings corresponding to the learner's sentences. In the first step, we obtain the audio and uttered phonemes along with its time-aligned boundaries belonging to a teacher's utterance from the repository. With these, in the second step, we compute the GoP score for the teacher's utterance. The second stage computes a GoP score for learner's utterance in three steps. In the first step, we obtain a noisy signal for a given learner's utterance under additive noise conditions. In the second step, we perform forced alignment on the learner's utterance to obtain time-aligned boundaries with the DNN-HMM phoneme models considering the phoneme in the teacher's utterance. In the third step, we compute a GoP score for the learner's utterance using aligned boundaries and the phonemes in the teacher's utterance. The third stage computes GL score by applying a mapping function on the GoP scores of teacher's and learner's utterance.

3.1. Lexicon based GoP (LGoP)

In general, GoP is defined for a phoneme p over the segment containing acoustic observation $\mathbf{O} = \{O_t, \forall\ 1 \leq t \leq T\}$, where T is the total number of frames in the phoneme segment. The boundaries for the phoneme segments are obtained by forced-aligning an utterance with its respective transcription and a native English lexicon.

Witt et al. [3] formulated the GoP as absolute log of posterior probability of a phoneme $\mathcal{P}(p|\mathbf{O})$ normalized by duration:

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

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. The same authors [3] approximated the GoP as:

$$GoP(p) = \frac{1}{T}\left|\log\frac{\mathcal{P}(\mathbf{O}|p)}{\max_{q\in Q}\mathcal{P}(\mathbf{O}|q)}\right| \quad (2)$$

Further Wenping et al. [11] modified the GoP as:

$$GoP(p) = \frac{\mathcal{P}(\mathbf{O}|p)\mathcal{P}(p)}{\sum_{q\in Q}\mathcal{P}(\mathbf{O}|q)\mathcal{P}(q)} \quad (3)$$

Log-likelihood ratio based GoP [11] was proposed as :

$$GoP(p) = \frac{1}{T}\left[\sum_{t=1}^{T}\log\mathcal{P}(O_t|p) - \max_{\{q\in Q, q\neq p\}}\sum_{t=1}^{T}\log\mathcal{P}(O_t|q)\right] \quad (4)$$

Further GoP was formulated [13] in terms of sub-phonemic posteriors $\mathcal{P}(s_t|O_t)$ and its priors $\mathcal{P}(s_t)$ as:

$$GoP(p) = \frac{1}{T}\sum_{t=1}^{T}\log\frac{\mathcal{P}(s_t|O_t^{(p)})}{\mathcal{P}(s_t)} \quad (5)$$

Sudhakara et al. [19] formulated GoP incorporating sub-phonemic transition probability $\mathcal{P}(s_t|s_{t-1})$ as:

$$GoP(p) = \frac{1}{T}\left[\sum_{t=1}^{T}\log\mathcal{P}(s_t|O_t^{(p)}) + \sum_{t=2}^{T}\log\mathcal{P}(s_t|s_{t-1}) + (T-1)\log n\right] \quad (6)$$

where n is the total number of sub-phonemes.

Generally, the GoP scores are defined at the phoneme level [3]. However, in general, a score is represented for a whole utterance. The score for an utterance is computed by averaging the scores across all the words in the sentence, where the word level score is obtained by averaging the scores across all the phonemes in the word [11].

3.2. Teacher's utterance based GoP (TGoP)

The GoP computed based on forced-alignment using the lexicon could result in lower performance. This is because the phoneme transcriptions obtained from forced-alignment might have phoneme errors even though the phoneme transcriptions are selected from native English lexicon [21]. For a given word, the lexicon contains one or multiple utterances. For example, the word "The", contains the following two different utterances "DH **IH**" and "DH **AH**". The former and latter versions are used when the word "The" appears before a word starting with vowel and consonant sounds respectively in a sentence. Thus, while uttering this word, the L2 learners have to carefully choose one of the utterance. Otherwise, it may result in phoneme errors i.e., mispronunciation. Similarly, such context-based utterance variations exist for many English words including "Project", "Live", "Lead", "Bow" and "Tear". In the forced-

alignment, it is non-trivial to obtain the phoneme transcriptions without such phoneme errors, when the lexicon contains multiple utterances. Under these conditions, existing lexicon based GoP computation results in inconsistent scores. We illustrate this using Figure 2.

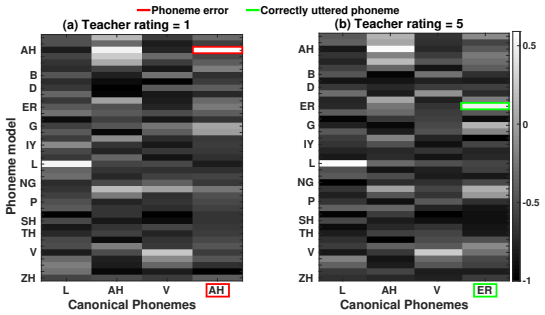

Figure 2: *Comparison of two learner's based on their pronunciation quality.*

Figure 2 shows the GoP scores computed using Equation 6 for the word "Lover" belonging to two utterances, which are assigned with ratings one and five respectively. The word has the following two utterances in the lexicon "L AH V **AH**" and "L AH V **ER**". Between the two utterances, the correct canonical utterance "L AH V **ER**" is observed in the teacher's utterance. Considering both the utterances, the phonemes estimated from forced-alignment are "L AH V AH" and "L AH V ER" respectively. Considering the respective phoneme transcriptions, the GoP scores of both the utterances are found to be 4.768 and 5.094 respectively. From these scores, it is observed that both the scores are closer, however, their respective ratings are far apart. Hence, it could result in performance degradation. In order to circumvent this, we propose to compute GoP score considering the phonemes in the teacher's utterance. Based on this, the GoP score is affected only for the utterance in Figure 2 (a) and it is found to reduce from 4.768 to 3.063. With this new value, the GoP scores are relatively farther compared to when lexicon based forced-aligned phonemes are considered. Thus, we believe that the GoP score computed using aligned boundaries obtained from the forced-alignment considering phonemes in the teacher's utterance could improve the performance compared to that computed based on the forced-alignment using the native English lexicon.

3.3. GoP like (GL) score computation

Typically, GoP is computed based on native English models. It is known that when there is a difference in the English accent between the teacher and the learner, then there could be differences in their speech acoustics. Hence, better performance is expected when the models are trained with the teacher's data if the teacher and learner belong to the same English accent. However, it is costly and cumbersome to obtain a large amount of data from the teacher for learning the phoneme models. In order to overcome this, we propose an approach by considering relative variations in the GoP score of the learner's utterance with respect to the GoP score of the teacher's utterance. Let the GoP score of the learner's and teacher's utterances be $GoP_l(p)$ and $GoP_t(p)$ respectively for the phoneme p. Considering these, we propose a mapping function which outputs a score based on the relative closeness of $GoP_l(p)$ with respect to $GoP_t(p)$ and we define this score as the GoP like (GL) score. We propose the function for GoP like score, for phoneme p as:

$$GL(p) = 1 - \tanh\left(k \times \left|\left(GoP_t(p) - GoP_l(p)\right) / GoP_t(p)\right|\right) \quad (7)$$

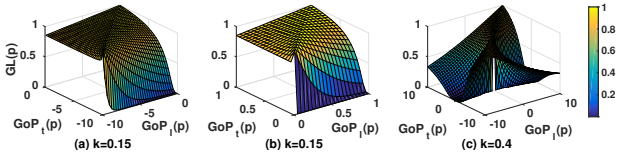

Figure 3: *Plot of $GL(p)$ for various ranges of $GoP_t(p)$ & $GoP_l(p)$*

where k ($k > 0$) is a parameter chosen to control the strictness of scoring. For the experiments in this work, we empirically found the k= 0.15 and 0.40 for Equation 1, 2, 3 & 4 and Equation 5 & 6 respectively. Inspired by the activation function of a neural network, we have used tanh in defining $GL(p)$. It is clear that $0 \leq GL(p) \leq 1$. $GL(p)$ takes value 1 when $GoP_l(p) = GoP_t(p)$ and vice-versa.

The range of GoP scores for Equation 1 & 2 is $(-\infty\ ,\ 0]$, for Equation 3 is [0 , 1] and for Equation 4, 5 & 6 is $(-\infty\ ,\ \infty)$. Figure 3 shows the variation of $GL(p)$ as a function of $GoP_t(p)$, $GoP_l(p)$ and k. From Figure 3, it can be observed that the function $GL(p)$ is close to 1 when $GoP_t(p) \approx GoP_l(p)$. From Figure 3(c), it can be observed that increasing the value of k leads to lesser $GL(p)$ score.

3.4. Noise selection

Recently, CAPT systems due to its reliability and cost-effective nature [22] are being implemented in classrooms where a large number of students utilize the service at the same time. In this environment, the learner's voice is corrupted with the surrounding students who are actively speaking. In general, this unwanted voice in the background could be characterized using babble noise [16]. The babble noise considered in the speech systems, encounters when a group of people are talking or babbling together among which the target speaker's voice is present. It is influenced by factors like the number of speakers and the surrounding environment. Further, CAPT systems are designed to be used through laptops and mobile devices, where the quality of the inbuilt microphones may vary across devices in terms of recording quality [23]. Besides that, the recording quality could also be degraded due to improper microphone settings. These together could be characterized by white Gaussian noise, which has a flat spectral density. Under these conditions, the analysis of GoP performance could be useful.

In addition, typically, planes are equipped with communication radio systems, which is used by pilots to communicate with air traffic control (ATC). Thus, it is crucial for the pilots to convey the information without any mispronunciation at crucial junctures to avoid wastage of time. In these conditions, the analysis of GoP could be useful. However, pilots' voice is corrupted by the noise in the cockpit due to the engine and the wind striking the body of the plane. Due to the significance of GoP analysis under these noise conditions, we propose to study variations under additive noise conditions considering babble, white Gaussian and f-16 cockpit noise at different SNRs. Though in real scenarios, the noise might not be additive, the analysis in this work is primarily performed to obtain insights in a controlled manner.

4. Experimental results

4.1. Experimental setup

We consider the GoP formulations in Equation 1, 2, 3, 4, 5 and 6 described in Section 3.1 for the analysis and refer them as E1, E2, E3, E4, E5 and E6 respectively. We consider the Pearson correlation coefficient [24] between the GoP scores and the

teacher's ratings as the measure. We consider the speech data under clean and noisy conditions with additive noises namely, babble, white Gaussian, f-16 at 0 dB, 10 dB and 20 dB SNRs. We use a DNN-HMM based acoustic model trained with LS Corpus [18]. We use Kaldi automatic speech recognition toolkit [25] to train the DNN-HMM acoustic model considering the architecture as provided in Dan's recipe [26]. We obtain time-aligned boundaries and phoneme transcriptions for the repository by applying forced-alignment on the speech data belonging to the teacher.

4.2. Results and discussion

Table 1 shows the correlation coefficients computed between the teacher's ratings and the scores obtained from each of the six GoPs under clean condition. The correlation coefficients are computed based on the scores separately obtained from LGoP, TGoP and GL score. From the table, it is observed that the correlation coefficients obtained with TGoP are higher than those with LGoP in all six equations. The highest absolute improvement among all equations is found to be 0.03. This indicates the benefit of TGoP in the pronunciation assessment task. This supports our hypothesis that the phoneme errors due to estimated utterance from multiple entries in the lexicon could cause performance degradation.

Table 1: *Correlation coefficient between the scores obtained from LGoP, TGoP & GL score and the teacher's ratings.*

	E1	E2	E3	E4	E5	E6
LGoP	0.4423	0.4450	0.4223	0.4504	0.5658	0.6245
TGoP	**0.4702**	**0.4726**	**0.4488**	**0.4806**	**0.5808**	**0.6399**
GL	0.4587	0.4582	0.4106	0.3201	0.5234	0.5681

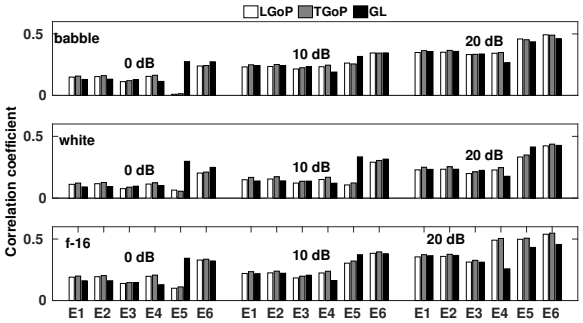

Figure 4: *Correlation coefficient between the LGoP, TGoP and GL score and the teacher's ratings for babble, white & f-16 noise at 0 dB, 10 dB & 20 dB SNR.*

Figure 4 shows the correlation coefficients computed between the teacher's ratings and LGoP, TGoP and GL score for all the six equations under all three noises at all three SNRs. From the figure, it is observed that among all three scores, the correlation coefficients obtained from both TGoP and GL scores are higher than that from LGoP for Equation 3, 5 & 6 and comparable to that from LGoP for Equation 1, 2 & 4 under all three noises at three SNRs. Further, the correlation coefficients gradually decrease from 20 dB SNR to 0 dB SNR conditions under all three noises for all three types of scores (LGoP, TGoP and GL). The highest decrement in the correlation coefficients with that obtained in clean condition is found to be 0.394 under 0 dB SNR. This indicates that the degradation in the performance of GoP is significantly high with additive noise. It is interesting to observe that the correlation coefficients obtained under f-16 noise are the highest and white noise the least respectively among all noises, scores, GoP equations and SNRs. This indicates that the GoP based pronunciation assessment is least affected by the noise under cockpit conditions and it is maximally affected by the noise due to microphone variabilities.

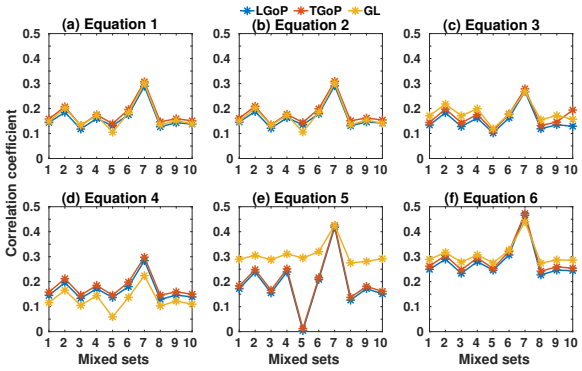

Figure 5: *Correlation coefficient between the LGoP, TGoP and GL score and the teacher's ratings for different sets of combination of noises.*

In general, the background noise in the applications of CAPT can be varied. In order to analyze the GoP performance under these variable conditions, we compute the correlation coefficients on the ten mixed sets derived such that its size is the same as the speech data considered (12800 utterances) in the experiment. The utterances in the mixed set 1 comprises equal amount of randomly chosen recordings from clean speech data as well as noisy speech under all three noises at all three SNRs. The utterances in the mixed sets 2, 3 & 4 are respectively from babble, white and f-16 under clean and all three SNRs. Similarly, the mixed sets 5, 6 & 7 are derived respectively for 0 dB, 10 dB and 20 dB SNR under all three noises. Further, the mixed sets 8, 9 & 10 are derived from all three SNRs under the following combination noises – babble & white, white & f-16 and babble & f-16 respectively. Figure 5 shows that the correlation coefficients obtained using TGoP is higher than those with LGoP in all the sets and all the equations. Further, the correlation coefficients with GL score are higher than or comparable to those with LGoP in most of the sets and most of the equations. This indicates the benefit of teacher utterance based computations under varying noise conditions. Additionally, Equation 6 exhibits an overall better performance compared to all the other equations and this could be because it considers transition probabilities and senone state posteriors in its formulation.

5. Conclusion

We study the variations in the performance of goodness of pronunciation (GoP) under noisy speech conditions to address its effectiveness under real scenarios. We propose to compute the TGoP and GL score for learner's utterance considering phonemes in the teacher's utterance unlike the phonemes estimated in the learner's utterance using forced-alignment process. Experiments are conducted on the speech data collected from Indian learners under additive noise conditions at three SNRs considering three noises. These reveal that the performance obtained using the phonemes from the teacher's utterance is higher than those using forced-alignment. Further investigations are required to obtain better strategies to improve performance under noisy conditions as well as for GL score. Future works also include incorporation of pronunciations based on text-to-speech (TTS) systems when the teacher's utterance is unavailable.

6. References

[1] D. Graddol, "Why global English may mean the end of English as a Foreign Language." ULIS, 2008.

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

[3] 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.

[4] S. Robertson, C. Munteanu, and G. Penn, "Pronunciation Error Detection for New Language Learners." in *INTERSPEECH*, 2016, pp. 2691–2695.

[5] R. Srikanth and J. Salsman, "Automatic Pronunciation Scoring And Mispronunciation Detection Using CMUSphinx," in *Proceedings of the Workshop on Speech and Language Processing Tools in Education*, 2012, pp. 61–68.

[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] 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.

[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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] S. Witt and S. Young, "Performance measures for phone-level pronunciation teaching in CALL," in *Proc. of the Workshop on Speech Technology in Language Learning*, 1998, pp. 99–102.

[16] Y. Luan, M. Suzuki, Y. Yamauchi, N. Minematsu, S. Kato, and K. Hirose, "Performance improvement of automatic pronunciation assessment in a noisy classroom," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 428–431.

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

[18] 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.

[19] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities," *accepted in INTERSPEECH*, 2019.

[20] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[21] R. Weide, "The CMU pronunciation dictionary, release 0.6," 1998.

[22] M. Eskenazi, "Using a computer in foreign language pronunciation training: What advantages?" *Calico Journal*, pp. 447–469, 1999.

[23] Y. Tsubota, M. Dantsuji, and T. Kawahara, "Practical use of autonomous English pronunciation learning system for Japanese students," in *InSTIL/ICALL Symposium*, 2004.

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

[25] D. Povey *et al.*, "The Kaldi speech recognition toolkit," *IEEE workshop on automatic speech recognition and understanding (ASRU)*, 2011.

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