INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Disentangling the Contribution of Non-native Speech in Automated Pronunciation Assessment

Shuju Shi[1], *Kaiqi Fu*[2], *Yiwei Gu*[2], *Xiaohai Tian*[1], *Shaojun Gao*[2], *Wei Li*[2], *Zejun Ma*[2]

[1]ByteDance Research, Singapore
[2]ByteDance Research, P.R. China

shuju.shi@bytedance.com, kaiqi.fu@gmail.com, {guyiwei.92, xiaohai.tian}@bytedance.com, gaoshaojun123@163.com, {liwei.speech, mazejun}@bytedance.com

Abstract

This study explores the impact of using non-native speech data in acoustic model training for pronunciation assessment systems. The goal is to determine how introducing non-native data in acoustic model training can influence alignment accuracy and assessment performance. Acoustic models are trained using different combinations of native and non-native speech data, and the Goodness of Pronunciation (GOP) metric is used to evaluate performance. Results show that models trained with manually labeled non-native data yield the highest assessment performance and alignment accuracy. Models trained with mixed non-native and native data perform best when considering the GOP distribution on both non-native and native speech. Additionally, models trained with native data are more robust to alignment variations. These findings highlight the importance of carefully selecting and incorporating non-native data in acoustic model training for pronunciation assessment systems.
Index Terms: pronunciation assessment, non-native speech, second language acquisition, acoustic model

1. Introduction

With the advancement of automatic speech recognition (ASR) technology, there has been a growing interest in developing pronunciation assessment which is an essential component in Computer-Aided Pronunciation Training (CAPT). The fundamental idea of non-native speech assessment has been measuring the deviation of learners' production from that of native speakers thus earlier studies have been using acoustic models trained using native data of the target language [1, 2, 3] to evaluate learners' pronunciation quality. The Goodness of Pronunciation (GOP) metric is widely used for phone-level pronunciation assessment. First proposed in [1], GOP is defined as the duration-normalized posterior probability of a reference phone given the corresponding acoustic segment. Since its introduction, various studies have extended GOP by improving acoustic models and refining calculation methods [2, 3, 4].

It has long been acknowledged that there are phonetic differences between native and non-native speech [5, 6, 7]. Incorporating non-native speech in acoustic model training has been shown to benefit automatic speech recognition (ASR) performance and mispronunciation detection accuracy on L2 data [8, 9, 10, 11, 12]. Recently, studies on automated pronunciation assessment have also started to incorporate non-native data in acoustic model training [13, 14, 15, 16]. It is important to note that there is a fundamental difference in the purpose of incorporating non-native speech in non-native ASR and non-native assessment or mispronunciation error detection. In the former case, the system aims to be either as adaptive as possible with the nonnative data or as invariant to accent variations as possible to achieve the best ASR accuracy. In contrast, in the latter case, there is a trade-off between higher recognition accuracy on non-native speech and objective assessment of non-native production. [13] and [14] trained acoustic models on a mix of 960-hour LibriSpeech corpus and 1,000-hour non-native speech by L1 Mandarin teenage learners. [15] mentioned about incorporating two sets of L2 data, one being 1,696 hours and the other being 6,591 hours, for different conditions of acoustic model training. Finally, [16] mentioned two non-native datasets, one being 10 hours and the other 4,000 hours, for different conditions of acoustic model training. However, the contribution of non-native speech to the improvement of pronunciation assessment systems has not been clearly addressed, except for the postulate that it could result in better alignment results, which in turn helps to yield more accurate assessment results [2].

In this paper, we aim to fill this gap by investigating the contribution of non-native speech in pronunciation assessment from two different perspectives: alignment accuracy and assessment performance. Specifically, we examine how the proportion of non-native data in the acoustic model training affects the alignment accuracy of non-native speech and how it subsequently impacts the performance of the pronunciation assessment system. Additionally, we investigate the influence of transcriptions used by comparing the performance of systems trained with human-annotated phone sequences and those trained with reference phone sequences. We also address the low-resource condition and investigate the optimal way to utilize non-native data in automated pronunciation assessment when only a limited amount of non-native data (10 hours) is available. Given that the acoustic models in this work will be based on Deep Neural Networks (DNNs), we adopt GOP as defined in [2].

2. Methodology

2.1. Dataset

Four datasets are used in this study, two open-sourced ones, LibriSpeech [17] and SpeechOcean762 [18], and two in-house datasets, L2-InHouse-Phone and L2-InHouse-Align. LibriSpeech consists of around 1000 hours of read speech based on public domain audiobooks. The training set (around 960 hours) and clean data from the dev and test sets (around 5.4 hours each) are used. Speechocean762 is a corpus designed for pronunciation assessment which includes a total of 5,000 read English utterances from 250 L1 Mandarin learners, half of which are children and the other half are adults. For pronunciation accuracy, each utterance is rated by five experts at three different levels, i.e., phone-level with a range of 0 to 1, word-level and utterance-level with a range of 1 to 10. For all three levels, a higher rating indicates better pronunciation and vice versa. Following the convention provided by the corpus, final scores at

10.21437/Interspeech.2023-380

phone-level are set as the average of all five experts and final scores at the other two levels are the median values of all five experts. The L2-InHouse-Phone dataset is an internal dataset collected at the authors' institution which includes around 1,450 hours (around 700,000 utterances) of read English speech by adult L1 Mandarin learners. The phone sequence of each utterance is manually annotated by linguistic experts (with advanced degrees in linguistics) after collection. The L2-InHouse-Align dataset is also an internal dataset which contains around 4 hours (3,300 utterances) of read English speech by adult L1 Mandarin learners. The segmentation boundaries of each phone are manually checked by linguistic experts (with advanced degrees in linguistics) using Praat [19]. The LibriSpeech (train, dev-clean) dataset and the L2-InHouse-Phone dataset are used for acoustic model training. The SpeechOcean762 dataset and LibriSpeech (test-clean) dataset are used for pronunciation assessment and the L2-InHouse-Align dataset is used for alignment accuracy evaluation.

2.2. Acoustic models

Triphone Gaussian Mixture Model and Hidden Markov Model (GMM-HMM) is trained using Kaldi [20] and forced-alignement is applied to obtain the frame-level senone (tied triphone states) labels of the acoustic data used for further acoustic model training. The input features for the GMM-HMM training are 13-dimension Mel-frequency cepstral coefficients (MFCCs), the change in coefficients, the change in delta values and pitch, which adds up to 40-dimension inputs. The deep feedforward sequential memory network and Hidden Markov Models (HMM), i.e., DFSMN-HMM, is adopted as the architecture of acoustic model training for alignment and assessment [21]. The DFSMN architecture consists of 2 convolution layers and 24 FSMN layers followed by two fully connected (FC) layers, a feedforward layer and a bottleneck layer. The number of units by the softmax output layers are determined by the senone labels derived from forced-alignment with corresponding GMM-HMM systems. The input features are 39-dimension Mel-frequency cepstral coefficients (MFCCs).

The DFSMN-HMM acoustic models are used in two ways: (1) to force-align the speech to get phone-level time stamps, and (2) to calculate Goodness of Pronunciation (GOP) scores for pronunciation assessment. In the forced-alignment procedure, posteriors from the DFSMN-HMM and state transition probability of the corresponding GMM-HMM are used to get the time stamps for the reference phones. The GOP score [2] for a phone p is calculated as follows:

$$GOP(p) = \mathcal{LPP}(p) - max_{q \in Q}\mathcal{LPP}(q) \quad (1)$$

where p is the phone in consideration and Q is the whole phone set. $\mathcal{LPP}(p)$ is the log phone posterior and is computed as $logp(p|\mathbf{o}; t_s, t_e)$, where t_s and t_e are the start and end frame indexes of phone p, and $\mathbf{o}$ are the corresponding acoustic observations. It is then further normalized to be in range of 0 to 1, for convenience of further analysis.

$$GOP_{norm} = \frac{GOP(p) - min(GOP)}{max(GOP) - GOP(p)} \quad (2)$$

3. Experiments and Results

3.1. Experimental settings

Two GMM-HMMs are trained to get the frame-level senone labels for the acoustic data and Table 1 shows the experimental settings for each of them. The 960 hours of L1 English are the train set from LibriSpeech and the 1,430 hours of L2 English are from the L2-InHouse-Phone dataset. Phone-labeling (PL) indicates whether the transcriptions used for the L2 dataset are based on human-annotated phone sequences ('YES') or reference forms from the CMU dictionary ('NO') [22].

Table 1: *Experimental settings for the GMM-HMM training. L1 and L2 indicate native English data and non-native English data, respectively. PL represents phone labeling conditions.*

	Data		PL
	L1	L2	
GMM-HMM_1	960h	1,430h	YES
GMM-HMM_2	960h	1,430h	NO

Table 2 shows various ways of data combinations for the DFSMN acoustic model training. The first five model conditions, from DFSMN_1 to DFSMN_5 aim to investigate how different ways of mixing native and nonnative English data could influence the alignment and assessment on nonnative speech hence the overall size of the training data are kept as the same, i.e., 960 hours, for fair comparisons among the five conditions. Models DFSMN_6 and DFSMN_7 are trained to examine the low-resource condition when only limited nonnative data are available. Models DFSMN_8 and DFSMN_9 are trained with the maximum number of training data available for this study with the expectation that this will lead to the highest performance.

Table 2: *Experimental settings for the DFSMN training. The Column Initial-Align shows the GMM-HMM used for the initial alignment results.*

	L1	L2		Initial-Align
		Duration	PL	
DFSMN_1	960h	0h	-	GMM-HMM_1
DFSMN_2	0h	960h	YES	GMM-HMM_1
DFSMN_3	0h	960h	NO	GMM-HMM_2
DFSMN_4	480h	480h	YES	GMM-HMM_1
DFSMN_5	480h	480h	NO	GMM-HMM_2
DFSMN_6	960h	10h	YES	GMM-HMM_1
DFSMN_7	960h	10h	NO	GMM-HMM_2
DFSMN_8	*960h*	*1,430h*	*YES*	*GMM-HMM_1*
DFSMN_9	*960h*	*1,430h*	*NO*	*GMM-HMM_2*

The selection of the checkpoint for each of the DFSMN models under each condition is based on a common development set which includes the LibriSpeech dev-clean data (around 5 hours) and around 20 hours of L2 data from the L2-InHouse-Phone dataset. The highest frame-level accuracy of senone labels for each of the DFSMN models on the common development set are given in Table 3.

As shown in Table 3, the DFSMN_5 model, trained on a combination of 480 hours of LibriSpeech data and 480 hours of L2-InHouse-Phone data with reference phone transcriptions, exhibits the highest performance among the first seven models, DFSMN_1 to DFSMN_7. The results show that models trained on a mixture of native and nonnative English data demonstrate greater accuracy than models trained on either dataset in isolation. Furthermore, models trained exclusively on nonnative data exhibit higher accuracy than those trained solely on native data, which is expected as the development set comprises

Table 3: *The highest frame-level classification accuracy of senones for each of the DFSMN models.*

Model	# of senones	Accuracy
DFSMN_1	5,712	0.581
DFSMN_2	5,712	0.616
DFSMN_3	5,664	0.632
DFSMN_4	5,712	0.695
DFSMN_5	5,664	**0.717**
DFSMN_6	5,712	0.635
DFSMN_7	5,664	0.639
DFSMN_8	*5712*	*0.732*
DFSMN_9	*5664*	*0.757*

both native and nonnative data, and there is a greater proportion of nonnative data in the corpus. Notably, the use of reference transcriptions, rather than human-annotated phone sequences, in the training data enhances the accuracy of models trained on nonnative data. This phenomenon may be attributed to the fact that learners from the same L1 background often share similar pronunciation error patterns, which in turn yields more stable acoustic distributions of sounds, despite deviating from native English production.

3.2. Alignment accuracy

To evaluate the accuracy of alignment, we used the L2-InHouse-Align dataset and measured the absolute differences between the force-aligned results and the human-annotated results. The average phone length in the L2-InHouse-Align dataset, as annotated by humans, was 114 ms. The metric used for evaluation was based on the absolute difference of time stamps at the phone-level. We analyzed the distribution of absolute differences at three levels, namely less than 25ms, 50ms, and 100ms, and report the percentage of phones falling in each range in Table 4. This evaluation approach will help us compare the performance of different models and determine their effectiveness in accurately aligning phone sequences.

Table 4: *Alignment accuracy by the DFSMN-HMM models on the L2-InHouse-Align dataset.*

Model	<25ms	<50ms	<100ms
DFSMN_1	0.327	0.61	0.829
DFSMN_2	**0.376**	**0.684**	**0.911**
DFSMN_3	0.351	0.642	0.878
DFSMN_4	0.373	0.682	0.910
DFSMN_5	0.355	0.647	0.885
DFSMN_6	0.351	0.651	0.884
DFSMN_7	0.356	0.644	0.877
DFSMN_8	*0.379*	*0.685*	*0.912*
DFSMN_9	*0.352*	*0.642*	*0.877*

Table 4 reveals that, among the first seven models from DFSMN_1 to DFSMN_7, the DFSMN_2 model, which is trained exclusively on 960 hours of nonnative data, attains the highest alignment accuracy across all three levels. The DFSMN_4 model, trained on a mixture of 480 hours of native data and 480 hours of non-native data, achieves slightly lower but comparable results to DFSMN_2. Notably, using native data exclusively, as done in DFSMN_1, results in the least alignment accuracy. In contrast to the frame-level classification accuracy presented in Table 3, where using reference phone sequences outperforms the human-annotated phone sequences when other factors are constant, the alignment results demonstrate that utilizing human-annotated phone sequences of nonnative data leads to better alignment accuracy than using reference phone sequences. Using human-annotated phone sequences of L2 data can result in more accurate alignment, indicating that acoustic models trained on L2 data with such annotations are better equipped to capture the nuances of L2 speech. In low-resource conditions, the results indicate that incorporating just 10 hours of nonnative data with either human-annotated phone sequences or reference phone sequences into acoustic model training can significantly improve alignment accuracy compared to using only native data. Furthermore, the results obtained with this approach are comparable to those obtained when all or half of the training data are nonnative data with reference phone sequences.

3.3. Pronunciation assessment performance

We evaluate the performance of pronunciation assessment on the SpeechOcean762 dataset. Since the training data for the acoustic models only consist of speech by adult learners, we partition the SpeechOcean762 dataset into two subsets: one exclusively including speech by adult learners and the other solely by children learners. The assessment results for both subsets, as well as for the entire dataset, are presented in Table 5, encompassing the performance of nine models, from DFSMN_1 to DFSMN_9. Furthermore, we examine different combinations of alignment and GOP calculation models to explore their impact on the assessment performance, and the outcomes are also presented in Table 5.

Table 5 presents the evaluation results of our models on the SpeechOcean762 dataset. The Column Alignment and Column GOP indicate the models used for forced-alignment and GOP calculation, respectively. The three columns, Adult, Child, and Both, represent results based solely on nonnative speech from adult learners, child learners, and the entire dataset, respectively. The evaluation results are reported in terms of Pearson correlation coefficients (PCC) between the GOP scores and the corresponding human-annotated scores at three levels: phone, word, and sentence. The GOP scores for phones are calculated using Equation 1 and then normalized based on Equation 2. For words and sentences, the GOP scores are the average of the scores of all phones within the corresponding word/sentence. In general, using the DFSMN_2 model for both alignment and GOP calculation gives the best result for assessment followed by the DFSMN_4 model, which is consistent with the alignment result in Table 4. When comparing the results between different learner groups, the results on adult learners outperforms that on the child learners consistently. This is not surprising since the data used for acoustic model training are by solely adult learners. Table 5 includes results obtained by using the DFSMN_2 model for forced-alignment and the remaining models for GOP calculation, which enables the disentangling of the effect of alignment accuracy and GOP assessment. Overall, the tendency remains similar to the results obtained when matching models are used for both alignment and GOP calculation. A more accurate but mismatching alignment model improves the pronunciation assessment performance of the DFSMN_1 model (which uses only native data for acoustic model training), and results in a small degradation in performance for the acoustic models with a mixture of native and non-native data where the proportion of native data is dominant or human-annotated phone sequences

Table 5: *Pronunciation assessment result, in terms of Pearson correlation coefficients (PCC), on the SpeechOcean762 dataset at three different levels, i.e., phone-level, word-level and sentence-level. The Column Alignment indicates models used for forced alignment and the Column GOP indicates the models used for GOP calculation.*

Alignment	GOP	Adult			Child			Both		
		Sentence	Word	Phone	Sentence	Word	Phone	Sentence	Word	Phone
DFSMN_1	DFSMN_1	0.624	0.397	0.407	0.441	0.284	0.319	0.543	0.360	0.373
DFSMN_2	**DFSMN_2**	**0.670**	**0.530**	**0.559**	**0.484**	**0.397**	**0.424**	**0.585**	**0.489**	**0.512**
DFSMN_3	DFSMN_3	0.581	0.430	0.455	0.438	0.302	0.324	0.514	0.39	0.409
DFSMN_4	DFSMN_4	0.664	0.488	0.515	0.491	0.362	0.407	0.587	0.452	0.477
DFSMN_5	DFSMN_5	0.628	0.440	0.459	0.462	0.329	0.356	0.545	0.407	0.422
DFSMN_6	DFSMN_6	0.637	0.445	0.456	0.468	0.337	0.371	0.563	0.413	0.425
DFSMN_7	DFSMN_7	0.645	0.458	0.46	0.456	0.347	0.369	0.556	0.425	0.427
DFSMN_8	*DFSMN_8*	*0.682*	*0.511*	*0.537*	*0.524*	*0.408*	*0.443*	*0.601*	*0.478*	*0.502*
DFSMN_9	*DFSMN_9*	*0.617*	*0.443*	*0.461*	*0.470*	*0.336*	*0.355*	*0.536*	*0.408*	*0.420*
DFSMN_2	DFSMN_1	0.625	0.410	0.414	0.433	0.322	0.329	0.534	0.382	0.382
DFSMN_2	DFSMN_3	0.541	0.411	0.430	0.422	0.332	0.334	0.477	0.386	0.395
DFSMN_2	DFSMN_4	0.656	0.488	0.507	0.462	0.366	0.384	0.568	0.452	0.465
DFSMN_2	DFSMN_5	0.589	0.435	0.443	0.417	0.318	0.329	0.505	0.399	0.403
DFSMN_2	DFSMN_6	0.625	0.438	0.448	0.437	0.350	0.357	0.537	0.411	0.416
DFSMN_2	DFSMN_7	0.610	0.438	0.445	0.431	0.349	0.354	0.527	0.412	0.413

are used for the non-native data. However, the performance of the pronunciation assessment is more severely degraded for the DFSMN_3 and DFSMN_5 when non-native data is used with reference phone sequences. In general, the results suggest that models trained with a greater proportion of native data are more robust to alignment variations.

For the low-resource conditions, when comparing the model performance of the DFSMN_1, DFSMN_6 and DFSMN_7 models, it shows that incorporating 10 hours of nonnatvie speech data leads to significant improvement in assessment accuracy across all three levels. The difference between the DFSMN_6 and DFSMN_7 models is negligible, suggesting that the use of human-annotated phone sequences does not have a substantial impact when native data dominates the acoustic model training. In contrast, comparing DFSMN_2 vs. DFSMN_3, DFSMN_4 vs. DFSMN_5, and DFSMN_8 vs. DFSMN_9 highlights the importance of human-annotated phone sequences when native data is less dominant in the training data. Notably, the PCC values between the DFSMN_7 and DFSMN_9 models are comparable at phone and word levels, but the DFSMN_7 model outperforms the DFSMN_9 model at sentence level, despite the DFSMN_9 model being trained on 1430 hours of nonnative data compared to only 10 hours for the DFSMN_7 model. These findings underscore the value of carefully considering the balance of native and nonative data, as well as the role of human-annotated phone sequences, in training effective pronunciation assessment models under low-resource conditions.

Table 6: *Pronunciation assessment result on the LibriSpeech test-clean set at the phone level.*

# Phones	Model	% GOP <1
190,765	**DFSMN_1**	**0.038**
190,622	DFSMN_2	0.216
190,411	DFSMN_3	0.240
190,736	DFSMN_4	0.045
190,486	DFSMN_5	0.240
190,744	**DFSMN_6**	**0.038**
190,532	**DFSMN_7**	**0.038**

Table 6 shows how the L2 pronunciation assessment models perform on the LibriSpeech test-clean dataset. The table presents the number of phones and the percentage of phones with a GOP value below 1, indicating incorrect pronunciation. The results show that two factors are crucial for assessing L1 speech: the proportion of nonnative data in acoustic model training and the use of human-annotated phone-labeling data. Using only native data yields the best performance, with little degradation seen from adding 10 hours of nonnative data. The model's performance is slightly worse when nonnative data with human-annotated phone sequences accounts for half of the training data. However, model performance significantly degrades when only nonnative data is used, and when half of the data are L2 data with reference phone sequences.

4. Conclusion

This study aims to enhance our understanding of the influence of non-native data on pronunciation assessment systems by examining alignment accuracy and assessment performance. It also investigates the impact of the proportion of non-native data in acoustic model training and the type of transcriptions used, including human-annotated phone sequences and reference phone sequences. Results reveal that using non-native data with human-annotated phone sequences during acoustic model training leads to the highest accuracy in alignment and pronunciation assessment of non-native speech. Mixing half native data and half non-native data with human-annotated phone sequences in training can achieve slightly worse but comparable results to the use of solely non-native data with human-annotated phone sequences. Additionally, the mixing condition performs better on the pronunciation assessment of native data. In low-resource conditions, adding 10 hours of non-native data, regardless of the type of transcriptions used, significantly improves alignment accuracy and assessment performance compared to using only native data for acoustic model training. These findings suggest that incorporating non-native data in acoustic model training can improve pronunciation assessment systems and that the proportion of non-native data and the type of transcriptions used are crucial factors to consider when developing such systems.

5. References

[1] 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.

[2] 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.

[3] 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." in *INTERSPEECH*, 2019, pp. 954–958.

[4] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," *arXiv preprint arXiv:2008.08647*, 2020.

[5] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, vol. 92, pp. 233–277, 1995.

[6] P. K. Kuhl, B. T. Conboy, S. Coffey-Corina, D. Padden, M. Rivera-Gaxiola, and T. Nelson, "Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e)," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 979–1000, 2008.

[7] U. Gut, *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Peter Lang, 2009, vol. 9.

[8] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[9] D. Vergyri, L. Lamel, and J.-L. Gauvain, "Automatic speech recognition of multiple accented english data," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] S. S. Juan, L. Besacier, B. Lecouteux, and T.-P. Tan, "Merging of native and non-native speech for low-resource accented asr," in *Statistical Language and Speech Processing: Third International Conference, SLSP 2015, Budapest, Hungary, November 24-26, 2015, Proceedings 3*. Springer, 2015, pp. 255–266.

[11] A. Lee and J. Glass, "Mispronunciation detection without nonnative training data," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] H. Hu, X. Yang, Z. Raeesy, J. Guo, G. Keskin, H. Arsikere, A. Rastrow, A. Stolcke, and R. Maas, "Redat: Accent-invariant representation for end-to-end asr by domain adversarial training with relabeling," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6408–6412.

[13] B. Lin, L. Wang, X. Feng, and J. Zhang, "Automatic scoring at multi-granularity for l2 pronunciation." in *Interspeech*, 2020, pp. 3022–3026.

[14] B. Lin and L. Wang, "Gated fusion of handcrafted and deep features for robust automatic pronunciation assessment," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1399–1404.

[15] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7262–7266.

[16] W. Liu, K. Fu, X. Tian, S. Shi, W. Li, Z. Ma, and T. Lee, "Leveraging phone-level linguistic-acoustic similarity for utterance-level pronunciation scoring," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[18] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[19] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2022.

[20] 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," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[21] S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-fsmn for large vocabulary continuous speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5869–5873.

[22] R. Weide *et al.*, "The carnegie mellon pronouncing dictionary," *release 0.7b, http://www.speech.cs.cmu.edu/cgi-bin/cmudict*, 2015.