INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Multilingual Speech Evaluation: Case Studies on English, Malay and Tamil

Huayun Zhang, Ke Shi, Nancy F. Chen

Institute for Infocomm Research, A*STAR, Singapore
zhang_huayun, shi_ke, nfychen@i2r.a-star.edu.sg

Abstract

Speech evaluation is an essential component in computer-assisted language learning (CALL). While speech evaluation on English has been popular, automatic speech scoring on low resource languages remains challenging. Work in this area has focused on monolingual specific designs and handcrafted features stemming from resource-rich languages like English. Such approaches are often difficult to generalize to other languages, especially if we also want to consider suprasegmental qualities such as rhythm. In this work, we examine three different languages that possess distinct rhythm patterns: English (stress-timed), Malay (syllable-timed), and Tamil (mora-timed). We exploit robust feature representations inspired by music processing and vector representation learning. Empirical validations show consistent gains for all three languages when predicting pronunciation, rhythm and intonation performance.

Index Terms:computer-assisted language learning (CALL), low-resource spoken language processing, multilingual, speech evaluation.

1. Introduction

Learning online has become the new norm after the COVID-19 pandemic. Increasing demand for online language learning has made Computer-Aided Pronunciation Training (CAPT) even more popular than before. Automatic speech evaluation, a key component in CAPT, aims to score speaking proficiency according to the standardized assessment criteria [1, 2, 3].

Traditional speech assessment studies usually focus on extracting features from the Hidden Markov Model (HMM) based automatic speech recognizer (ASR) (e.g. [4, 5, 6]). [4] explored posterior probabilities and duration related features in automatic scoring. As a variation of likelihood ratio, Goodness of Pronunciation (GOP) [5] is one of the the most widely adopted feature in speech evaluation task [6]. With the development of deep neural network (DNN), GOP was further optimized to predict better phone segmentation and posterior estimation [7, 8, 9]. Weighted GOP was also proposed to improve its discriminative ability in non-native speech [10]. Recently, an ASR-free approach was proposed, which drives features from the marginal distribution of speech signals [11].

Linguistic features have also attracted research interests. As in [12], a prompt-aware feature was proposed for spontaneous speech evaluation. A context-aware GOP was proposed in[13] to incorporate phone transition factor and phone duration factor into the calculation of GOP. [14] leveraged Bidirectional Long Short Term Memory (Bi-LSTM) to learn high-level abstraction features by encoding both time-sequence and time aggregated information from speech. [15] proposed to encode lexical information and acoustic information in separate neural networks. All these studies are focused on languages with rich resources. Extracting cross language effective features, especially on low resource languages, is still challenging.

In terms of modeling approaches, early scoring strategies were based on statistical models such as Gaussian process [16]. Recent studies employed more deep learning approaches. [17] proposed to utilize Convolution Neural Network (CNN) along with a multi-layer perceptron classifier. [18] explored three deep learning based acoustic models including Tandem GMM-HMM, DNN, CNN, and found they provide substantial improvement in scoring performance. Long short-term memory recurrent network (LSTM) was adopted in pronunciation assessment [19, 20]. More recently, attention mechanism has also been applied [21, 15, 12] to speech evaluation. These studies have presented promising improvement on speech evaluation performance in the language specific tasks. However, multilingual speech scoring task was not explored.

In this paper, we propose a unified framework for fine-grained multilingual speech evaluation on assessing pronunciation, rhythm, and intonation by leveraging robust feature representations. Specifically, we investigate rhythm-aware tempo features and multilingual vector representations. Moreover, multi-task learning strategy is employed to further improve the evaluation performance on the low-resource languages.

2. Methods

2.1. Setup

All the experiments in this paper follow the system diagram in Figure 1: Given a speech utterance from a student, forced alignment and phone decoding is applied to an acoustic model to obtain phoneme level timing and the acoustic log likelihood. Various features are derived to train scoring model to predict fine-grained scores to quantify oral language skills (i.e. pronunciation, rhythm, and intonation).

Performance Metrics: The Pearson Correlation Coefficient (PCC) and Mean Square Error (MSE) between model prediction and the average score of teachers.

2.2. Acoustic Model

Mel-filter bank feature (29 dimensional) was adopted in acoustic modeling. Optimized on various ASR tasks, this model is designed as a combination of different neural networks: At the bottom levels, a stack of specially designed convolution neural network (CNN) running on 2D windows across time and frequency is trained to extract robust intermediate representation from the filter bank outputs. Time Delay Neural Network (TDNN) and LSTM are placed on top of these CNN. The output of the final LSTM is linked to 1500 output senones via a fully connected layer. This model was trained on WSJ [22], Switchboard [23] and Fisher [5], using lattice-free Maximum

http://dx.doi.org/10.21437/Interspeech.2021-1258

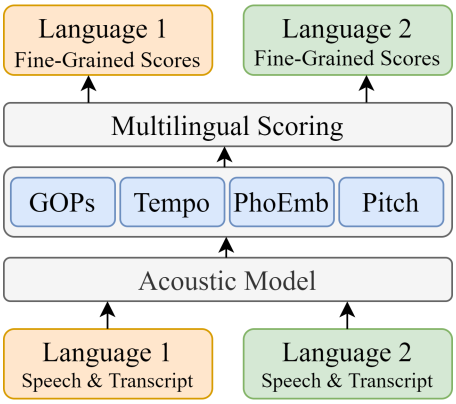

Figure 1: *Speech evaluation framework, fine-grained scores include the metrics of pronunciation, fluency, and intonation.*

Mutual Information (MMI) criterion with a sub-sampling factor of 3 [24]. To minimize the acoustic mismatch on L2 speech evaluation tasks, acoustic adaptation was implemented using the speechocean762 training data as described in Section 3.

Malay and Tamil Acoustic Models: Malay and Tamil acoustic models have the same configuration and were trained on 600 hours' Malay speech (1,500 native speakers from Singapore and Malaysia) and 200 hours' Tamil speech (700 native speakers from Singapore) respectively.

2.3. Feature Representation

2.3.1. Acoustic Posterior Probability

Goodness of Pronunciation (GOP) [5] has been widely adopted in speech evaluation. Phone level GOP is the time average of log posterior probability over the phone duration:

$$\mathrm{GOP}(p) = \frac{1}{T} log(P(p|\mathbf{O})) \tag{1}$$

where $\mathbf{O} = [\mathbf{o}_1, ..., \mathbf{o}_T]^T$ is a speech segment of phone p in the alignment.

$$P(p|\mathbf{O}) = \frac{P(p)P(\mathbf{O}|p)}{\sum_{q \in Q} P(q)P(\mathbf{O}|q)} \tag{2}$$

where Q is the collection of all possible paths on the observation $\mathbf{O}$. $P(p)$ and $P(q)$ are the priors of p and q.

2.3.2. Tempo / Duration

In musical terminology, tempo is the pace of a given piece [25]. It is usually measured in beats per minute. We borrow this concept to measure the phonological timing patterns in speech. In this study, speech tempo is defined as a combination of speaking rate and normalized phone duration:

$$\mathrm{T}(p) = \mathrm{concat}(1/\tau, (\tau - \mu)/\sigma) \tag{3}$$

where τ is the duration of current canonical phone in the alignment, (μ, σ) are the normal distribution parameters of τ in the sentence. The instant phone tempo is spliced with its context into a tempo vector:

$$\mathbf{T}(p) = \mathrm{concat}(\mathrm{T}(p_{i-k}), ..., \mathrm{T}(p_i), ..., \mathrm{T}(p_{i+k})) \tag{4}$$

where k is the number of neighboring phones considered in each direction.

For comparison, phone duration feature was also tested. Similar to the tempo feature, phone duration and duration difference between successive phones are spliced with its neighbors into a duration vector.

2.3.3. Multilingual Phonemic Embedding

In this work, we propose to use a multilingual vector representation to characterize the spoken utterances from English, Malay and Tamil. Representing the speech signal with phonetic features from other languages have shown to be useful in many tasks, including speech recognition [26], spoken term detection [27], speech summarization [28], and spoken language identification [29]. However, such multilingual representations have not been applied to speech evaluation. Therefore, in this work, we adapt distributed linguistic representation and apply well tested approaches in NLP to characterize the multilingual phonemic space. A phoneme embedding matrix was estimated by Google's Word2Vec [30] on a multilingual training corpus (words were mapped to phoneme strings). Items in language-specific phoneme tables $Q_{English}$,Q_{Malay} and Q_{Tamil} are prefixed with corresponding language ID and merged together to form a multilingual phoneme table,$Q_{Multilingual}$. Each canonical phone in alignment is assigned a unique one-hot index from this multilingual phoneme table. By multiplying the embedding matrix, this one-hot phoneme index is transformed into a D dimensional vector. $D = 32$ in this study.

2.3.4. Pitch

Pitch provides acoustic cues for a speaker's intonation, confidence and expressiveness. In this study, a feature vector including raw log pitch, normalized log pitch, delta log pitch and wrapped NCCF (Normalized Cross Correlation Function) was extracted from Kaldi [31]. Frame-wise pitch vectors are averaged in each canonical phone's duration.

2.3.5. Feature Assembly

For each canonical phone, relevant features are concatenated in a sequence.

$$\mathbf{V}(p) = \{\mathrm{GOP}(p), \mathrm{Tempo}(p), \mathrm{PhoEmb}, \mathrm{Pitch}\} \tag{5}$$

At the end of each sentence, an utterance ending symbol is appended.

2.4. Scoring Module

The scoring module aims to map the meta features described in Section 2.3 to the fine-grained proficiency scores assigned by human raters. RNN is investigated as the deep learning backbone, in particular, stacked Bi-directional long short term memory (Bi-LSTM) is utilized in sequential modeling. Finally, the hidden representations encoded by RNN are fed into a linear layer followed by a $\tanh$ activation function to predict the scores.

The human scores are re-scaled into the range of $(-1, +1)$ before training, and the model predictions are scaled back before comparing with human scores. MSE is chosen as the loss function in training. In this study, both monolingual and bilingual models have approximately the same numbers of parameters (around 2M).

2.5. Low Resource Speech Evaluation

Collecting and labeling L2 speech data is time consuming and labor intensive, so data preparation is often the bottleneck when developing speech evaluation models for low resource languages. From our experience, we speculate that for the same type of fine-grained metrics, be it pronunciation, rhythm, or intonation, there is certain language agnostic information that teachers use during scoring. This type of language agnostic information could help mitigate the adverse effects stemming from the scarcity of linguistic resources.

In Singapore, while Malay and Tamil are two of the four official languages and there is strong support to develop spoken language technology to help students learn their ethnic mother tongues, 67% Malay households and 70% Indian households speak English as their main language. Only 3% of the population speaks Tamil at home[32]. Therefore, we employ two strategies to tackle the low-resource challenge. The first is model adaptation: Tamil model could be adapted from a well trained English or Malay model. The second is data augmentation: Multilingual tasks are learned simultaneously by sharing most model parameters and a language-specific linear layer before the output tanh.

In this work, we used Malay and Tamil for multilingual speech evaluation as the teaching scoring data are more homogeneous; both Malay and Tamil were scored by teachers certified by the Ministry of Education in Singapore, whereas the English data was scored outside Singapore.

3. Experiments

3.1. Speech Corpora

Experiments were conducted in three languages: English, Malay, and Tamil.

English: Speechocean762[1], a recently released data set for speech evaluation was investigated in this study. It consists of 5000 English utterances collected from 250 nonnative speakers. Half of the data, i.e. 2500 utterances and 125 speakers are reserved as the test data. There is no speaker overlap between testing and training. Mandarin Chinese is the first language for all speakers. Half of the speakers are children. For each utterance, five raters' scores are provided at 3 levels: phoneme level, word level and sentence level. Sentence level scores were investigated in this study. The average inter-rater PCC are 0.754, 0.767and 0.753 on pronunciation, rhythm, and intonation respectively.

Malay: Our Malay corpus contains 14,088 utterances and 230 Singapore speakers between 9-16 years old. The average inter-rater PCC are 0.547, 0.571, 0.545 on pronunciation, fluency, and intonation respectively.

Tamil: Our Tamil corpus consists of 5,215 utterances collected from 100 Singapore speakers between 9-16 years old. Each utterance was scored by four raters. The average inter-rater PCC are 0.568, 0.619, and 0.582 on pronunciation, fluency, and intonation respectively.

Score Annotations from Teachers: English utterances from Speechocean762 are scored by human-raters independently using a 10-point scale (1 is the lowest, 10 is the highest). Malay and Tamil utterances were scored by human-raters independently using a 5-point scale. (1 is the lowest, 5 is the highest). For Malay and Tamil, the average rating scores were used as ground truth scores. For each corpus, multiple inter-rater PCC were calculated between the scores of one rater and the average scores of the rest of all raters [21]. By averaging all inter-rater PCC, the upper bound of the scoring performance (Human performance) was obtained (see the bottom lines in Table 2-3). For Speechocean762, the median scores were adopted following the example score files coming with the database. Multiple inter-rater PCC were calculated between the scores of one rater and the median scores of the rest of all raters as shown in Table 1.

3.2. Results

All model and feature configurations are compared with PCC and MSE metrics following setup in previous work [15, 21].

English Experimental Results: The results for English are shown in Table 1[2]. It is expected that GOP feature performs acceptably on the speech evaluation task. The scoring performance was improved by replacing normal duration feature with tempo feature, which reduced MSE by 3.3%, 1.5%, 3.7% relatively and improved PCC by 2.3%,0.7%, 1.8% relatively on the three oral proficiency measures, i.e. pronunciation, rhythm, and intonation, respectively. The phoneme embedding feature boosted the performance further by 0.4%, 4.8%, 1.0% relative MSE decrements and 0.8%, 2.4%, 0.4% relative PCC improvements on the three oral proficiency measures.

Malay Experimental Results: The results for Malay are shown in Table 2. Replacing duration feature with tempo feature brought 1.1%, 1.2% and 0.9% relative MSE reductions and 1.7%, 2.1% and 1.9% relative PCC improvements on the three oral proficiency measures respectively. By using phoneme embedding feature, scoring performance was further improved by 1.3%, 1.5% and 0.9% relatively for MSE and 2.5%, 2.4% and 1.9% relatively for PCC on the three oral proficiency measures respectively. Pitch feature improved the PCC performance by another 1.2%,4.3% and 9.6% relatively. In addition, multi-task learning strategy benefited to Malay speech scoring task, especially for pronunciation and fluency proficiency measures.

Tamil Experimental Results: Table 3 shows the results for Tamil. Similar to English and Malay, the tempo feature performs better than duration. The multilingual embedding feature brought improvements on pronunciation while the pitch feature brought improvements on fluency and intonation. As data scarcity is a major bottleneck for Tamil speech evaluation, two cross language training strategies were further investigated: acoustic adaptation and data augmentation (multi-task learning). The results in Table 3 show that both methods are effective. Especially, multi-task learning reduced MSE by 2.4%, 2.6%, 3.5% relatively and improved PCC performance by 3.1%, 2.2%, and 3.6% relatively on the three oral proficiency measures respectively.

[1] http://www.openslr.org/101/

[2] Speechocean762's scoring categories are accuracy, fluency and prosody, which primarily evaluate the pronunciation, rhythm and intonation according to http://www.openslr.org/101/. We adapt the terminology to be more consistent with other datasets to make it easier for comparing performance.

Table 1: *Results on English: MSE and PCC scores of different model and feature configurations.*

Feature	**Pronunciation**		**Rhythm**		**Intonation**	
	MSE	**PCC**	**MSE**	**PCC**	**MSE**	**PCC**
GOP	1.428	0.639	1.065	0.694	1.009	0.713
GOP+Dur	1.392	0.647	1.021	0.707	1.018	0.710
GOP+Tempo	1.346	0.662	1.006	0.712	0.980	0.723
GOP+Tempo+PhoEmb	**1.341**	**0.667**	**0.958**	**0.729**	**0.970**	**0.726**
GOP+Tempo+PhoEmb+Pitch	1.368	0.654	1.037	0.702	1.064	0.699
Human	-	0.754	-	0.767	-	0.753

Table 2: *Results on Malay: MSE and PCC scores of different feature configurations.*

Model	**Feature**	**Pronunciation**		**Fluency**		**Intonation**	
		MSE	**PCC**	**MSE**	**PCC**	**MSE**	**PCC**
Monolingual	*GOP*	0.537	0.473	0.580	0.487	0.707	0.415
	GOP+Dur	0.536	0.474	0.579	0.487	0.703	0.421
	GOP+Tempo	0.530	0.482	0.572	0.497	0.697	0.429
	GOP+Tempo+PhoEmb	0.523	0.494	0.563	0.509	0.691	0.437
	GOP+Tempo+PhoEmb+Pitch	0.518	0.500	0.546	0.531	0.658	0.479
Malay, Tamil Bilingual	*GOP+Tempo+PhoEmb+Pitch*	**0.506**	**0.518**	**0.538**	**0.540**	**0.656**	**0.482**
Human		-	0.547	-	0.571	-	0.545

Table 3: *Results on Tamil: MSE and PCC scores of different feature configurations and training strategies.*

Model	**Feature**	**Pronunciation**		**Fluency**		**Intonation**	
		MSE	**PCC**	**MSE**	**PCC**	**MSE**	**PCC**
Monolingual	*GOP*	0.344	0.490	0.354	0.584	0.259	0.552
	GOP+Dur	0.334	0.513	0.348	0.594	0.254	0.563
	GOP+Tempo	0.333	0.516	0.342	0.603	0.250	0.573
	GOP+Tempo+PhoEmb	0.330	0.522	0.348	0.594	0.260	0.552
	GOP+Tempo+PhoEmb+Pitch	0.332	0.518	0.347	0.595	0.258	0.555
Monolingual (Adaptation)	*GOP+Tempo+PhoEmb+Pitch*	0.334	0.519	0.339	0.608	0.254	0.562
Tamil, Malay Bilingual	*GOP+Tempo+PhoEmb+Pitch*	**0.324**	**0.534**	**0.338**	**0.608**	**0.249**	**0.575**
Human		-	0.568	-	0.619	-	0.582

4. Discussion

Speech Tempo: There have been studies on speech measurements that compare different rhythmic patterns across languages [33, 34, 35]. We adopted speech tempo as the three languages we are investigating are known to possess distinct rhythm patterns: stress-timed for English, syllable-timed for Malay and mora-timed for Tamil. We empirically show that speech tempo features are straightforward to use and effective in speech evaluation modeling, showing consistent improvements compared to traditional duration features for the PCC metric.

Multilingual Phoneme-Aware Scoring: By introducing a multilingual phoneme embedding feature, data from the same phoneme is trained in a phoneme-specific subspace, while data from different phonemes would be trained in separate subspaces. The variability in the scoring model is attributed to both the pronunciation variation that is phoneme independent and the pronunciation variation that is phoneme dependent. Phoneme aware modeling decouples these two variations and provides a better prediction on the language proficiency.

Cross Lingual Modeling: On low resource tasks such as Tamil, we attempted to improve model performance by leveraging from Malay. Both model adaptation and data augmentation have been shown to be effective. Consistent improvement was observed, suggesting language-agnostic information could potentially help speech evaluation scoring. Especially data augmentation improved data efficiency and alleviated training data over-fitting on low resource speech scoring tasks. We did explore using English data and models to help Tamil, but only observed minimal gains. We suspect this is because the English data was scored from a different standard than Malay and Tamil (the latter two is based on needs of Singapore Education). How to further exploit English resources to help lower-resource languages for speech evaluation is a theme of on-going work.

Pitch: In the Malay task, We observed obvious contribution by using pitch feature though similar trends were not observed for other languages such as English. We leave further analysis and investigations on how to more appropriately exploit pitch-related features for future work.

5. Conclusion

We systematically compared different feature configurations on multilingual speech evaluation tasks, focusing on sentence level fine-grained metrics. Tempo feature and multilingual phoneme embedding features were introduced. Consistent improvements were observed in experiments by adopting tempo-aware and phoneme-aware features in evaluation modeling. While Malay and Tamil are from different language, cross-lingual experiments showed that data and models in other languages could help improve speech evaluation performance. In the future, we will explore unsupervised error pattern discovery to diagnose speaker-specific pronunciation problems [36].

6. References

[1] M. Levy and G. Stockwell, *CALL dimensions: options and issues in computer assisted language learning*. Routledge, 07 2006.

[2] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009, spoken Language Technology for Education.

[3] N. F. Chen and H. Li, "Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1–7.

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] J. van Doremalen, C. Cucchiarini, and H. Strik, "Using non-native error patterns to improve pronunciation verification," *Proc. Interspeech*, 2010.

[11] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "Asr-free pronunciation assessment," *Proc. Interspeech*, 2020.

[12] Y. Qian, R. Ubale, M. Mulholland, K. Evanini, and X. Wang, "A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 979–986.

[13] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," *Proc. Interspeech*, 2020.

[14] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 338–345.

[15] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6234–6238.

[16] R. C. van Dalen, K. M. Knill, and M. J. Gales, "Automatically grading learners' english using a gaussian process." in *SLaTE*, 2015, pp. 7–12.

[17] A. Lee *et al.*, "Language-independent methods for computer-assisted pronunciation training," Ph.D. dissertation, Massachusetts Institute of Technology, 2016.

[18] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 6140–6144.

[19] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 338–345.

[20] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving mispronunciation detection for non-native learners with multisource information and lstm-based deep models," in *Proc. Interspeech 2017*, 2017, pp. 2759–2763.

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

[22] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26*, 1992.

[23] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, pp. 517–520 vol.1.

[24] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," in *Proc. Interspeech 2018*, 2018, pp. 12–16.

[25] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[26] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on speech and audio processing*, vol. 4, no. 1, p. 31, 1996.

[27] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 nist spoken term detection evaluation," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*. Causal Productions Pty Ltd, 2007, pp. 2385–2388.

[28] N. F. Chen, B. Ma, and H. Li, "Minimal-resource phonetic language models to summarize untranscribed speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8357–8361.

[29] F. Pellegrino and R. André-Obrecht, "Automatic language identification: an alternative approach to phonetic modelling," *Signal Processing*, vol. 80, no. 7, pp. 1231–1244, 2000.

[30] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," in *roceedings of the International Conference on Learning Representations (ICLR 2013)*, 01 2013, pp. 1–12.

[31] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2494–2498.

[32] Department of Statistics, Ministry of Trade and Industry, Republic of Singapore, "General household survey," *Press Release*, 2015.

[33] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 75, no. 1, pp. AD3–AD30, 2000.

[34] V. Dellwo, P. Karnowski, and I. Szigeti, *Rhythm and speech rate: A variation coefficient for deltaC*. Peter Lang, 01 2006, pp. 231–241.

[35] L. E. Ling, E. Grabe, and F. Nolan, "Q uantitative characterizations of speech rhythm: Syllable-timing in singapore english," *Language and speech*, vol. 43, no. 4, pp. 377–401, 2000.

[36] A. Lee, N. F. Chen, and J. Glass, "Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6145–6149.