Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

GoP2Vec: A few shot learning for pronunciation assessment with goodness of pronunciation (GoP) based representations from an i-vector framework and augmentation

Meenakshi Sirigiaju[1], *Chiranjeevi Yarra*[1]

[1]Speech Processing Lab, Language Technologies Research Center (LTRC), IIITH, India
meenakshi.sirigiraju@research.iiit.ac.in, chiranjeevi.yarra@research.iiit.ac.in

Abstract

Automatic pronunciation assessment is a critical component in computer assisted language learning. Typically, modeling pronunciation assessment tasks need labels, which are difficult to obtain as it requires expert annotators. Thus, it is essential to build an accurate model with less annotated data. In this work, an approach is proposed that considers a few speech samples using the i-vector framework. Each sample, first, is lengthened by T factor by concatenating the augmented samples of the same speech. The augmentation is obtained using time-scale modification (TSM), pitch-scale modification (PSM) and both. Next, phoneme-level goodness-of-pronunciation scores of concatenated speech are converted to a vector (GoP2Vec) with the i-vector framework. Experiments on two datasets revealed that the proposed GoP2Vec outperforms the state-of-the-art (SOTA) unsupervised methods and is on par with the SOTA supervised methods when it is used to train a simple neural model with a few samples.

Index Terms: Pronunciation assessment, i-vector, few-shot learning, CALL, data augmentation

1. Introduction

Pronunciation assessment is a critical component of language learning, particularly for second language (L2) learners. Accurate pronunciation is essential for effective communication, as it directly impacts intelligibility and fluency. Traditionally, pronunciation assessment relies on human raters to evaluate speech samples, which can be subjective, time-consuming, and resource-intensive. For language learners, timely and objective feedback on pronunciation is crucial for identifying areas of improvement and accelerating the learning process. However, the scalability of human-based assessment is limited, creating a pressing need for automated systems that can provide consistent and reliable evaluations.

In the literature, pronunciation assessment was explored through both supervised and unsupervised approaches. The most common feature used is the Goodness of Pronunciation (GoP) [1] score, which relies on Automatic speech recognition (ASR) and forced alignment process in ASR. Many works were proposed [2, 3, 4, 5] utilising GoP. Recently, end-to-end (E2E) models have gained significant popularity due to their ability to learn directly from raw audio inputs without requiring extensive feature engineering. For instance, Gong et al. [6] proposed a Transformer-based model (GOPT) that uses multi-task learning to assess pronunciation at multiple granularities—phoneme, word, and utterance levels—leveraging GOP features. Similarly, Chao et al. [7] introduced a hierarchical model for multi-aspect and multi-granular pronunciation assessment, utilizing sub-phoneme embeddings, depth-wise separable convolution, and a score-restraint attention pooling mechanism to capture both local and global contextual cues. Yang et al. [8] combined Multi-Head Self-Attention with Convolutional Neural Networks to extract local and global features, incorporating multi-layer feature fusion and multi-task loss weight optimization to enhance performance.

In addition to end-to-end approaches, fine-tuning pretrained models has also been widely adopted for pronunciation assessment. Liang et al. [9] developed an encoder-decoder framework with a mask-predict strategy to avoid misalignment issues, pre-training the model on ASR datasets and fine-tuning it with expert annotations. Ryu et al. [10] fine-tuned a pretrained Wav2Vec 2.0 model on the TIMIT dataset for phone recognition, optimized jointly for pronunciation assessment and mispronunciation detection and diagnosis tasks. Similarly, Kim et al. [11] fine-tuned Self-Supervised Learning models such as HuBERT and Wav2Vec 2.0 using Connectionist Temporal Classification loss, extracting layer-wise contextual representations for pronunciation scoring.

While neural network (NN)-based pronunciation models have become more popular than traditional methods, as the NN models learn the pronunciation patterns directly from data, they are data-hungry. In the pronunciation assessment modelling, it is challenging to obtain a large amount of speech data collected from L2 learners, followed by annotations from experts. In contrast to NN-based models, the traditional pronunciation assessment models do not require expert annotated ratings in the modelling [1, 2, 3, 4, 5]. As these methods do not utilise annotated ratings, the performance from these models is below that of the recent E2E and finetuning-based approaches.

To address the performance gap with a small amount of annotated speech data, this work proposes an approach to convert the GoP score sequence of an utterance to a vector representation using an i-vector framework and spoken data augmentation strategies. For augmenting a speech sample, a set of methods is chosen such that it would not affect the pronunciation quality of the augmenting speech sample. For augmenting speech, the speech sample is modified with its time scale, pitch scale and both. The augmented samples are randomly selected specific to each modification approach and concatenated to the original sample to increase overall duration. From the resultant concatenated speech, a GoP score sequence is computed using the traditional approach and obtain a vector (GoP2Vec) representation using i-vector modelling. For this, the background Gaussian mixture model (GMM) [12] and total variability matrix are trained with the concatenated samples from the respective augmentation technique. The GoP2Vec is used to train an NN-based classifier with a set of few samples to predict pronunciation quality. Experiments on SpeechOcean762 and voisTUTOR corpora revealed that the proposed approach is on par with the

10.21437/Interspeech.2025-2359

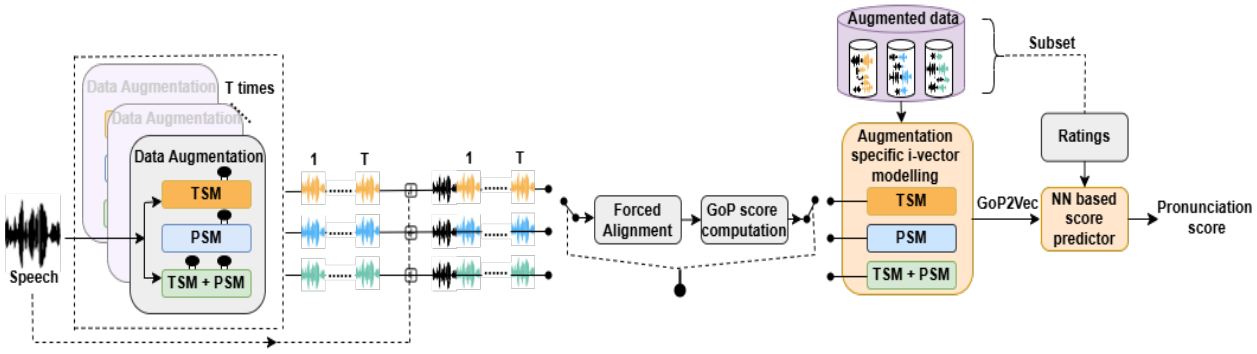

Figure 1: *Block diagram of the proposed GoP2Vec based pronunciation assessment*

state-of-the-art supervised E2E approaches and a significant improvement from the SOTA traditional unsupervised approaches.

2. Dataset

For the experiments conducted in this work, the L2 datasets voisTUTOR [13] and SpeechOcean762 [14] are utilized.

2.1. voisTUTOR

The dataset consists of 1,676 unique stimuli, featuring speech recordings from 16 Indian L2 learners of English. The learners, aged between 19 and 25, represent six native language backgrounds: Malayalam (4), Kannada (5), Telugu (3), Tamil (2), Hindi (1), and Gujarati (1), with an equal distribution of 8 male and 8 female speakers. The stimuli include words with a minimum of 1 and a maximum of 26 in length. The pronunciation quality of each utterance was evaluated by an expert on a scale of 0 to 4, reflecting the overall quality of the speech. The dataset spans a total duration of 14 hours.

2.2. SpeechOcean762

It consists of 5,000 English utterances from 250 L2 learners, with train and test splits of 2,500 utterances each. The learners are evenly split between children and adults, all native Mandarin speakers with a 1:1 gender ratio. Each utterance was manually rated by five experts at the phoneme, word, and sentence levels. The phoneme-level score reflects the pronunciation accuracy of each phone, while the word-level scores include accuracy and stress. At the sentence level, scores capture accuracy, completeness, fluency, and prosody. The dataset spans 6 hours, with each learner reading 20 sentences ranging from 1 to 20 words in length. In this work, we consider the median of the five expert scores provided for each utterance, utilizing the sentence-level total scores that indicate overall pronunciation quality on a scale of 0 to 10.

3. Background

An i-vector (identity vector) [15] is a compact, fixed-dimensional representation of variable-length sequences. It builds on the concept of factor analysis to capture underlying variability in a concise form. Introduced as an improvement over Joint Factor Analysis (JFA) [16], the key advantage of the i-vector approach is that it consolidates all variabilities into a single total variability space. The process generally involves two main steps: training a Universal Background Model (UBM) consisting GMM and extracting i-vectors using the Total Variability matrix. Since speech signals vary in length, the i-vector's ability to provide a compact and fixed-dimensional representation makes it highly effective for various speech applications. It is widely used in tasks such as speaker recognition [17, 18], language recognition [19], accent recognition [20], acoustic event detection [21], emotion recognition [22], and speaker diarization [23]. Using i-vectors for these applications has proven to be highly effective [24]. A key observation noted by Kenny et al. [25] is that i-vectors extracted from short utterances tend to be less reliable, whereas their reliability improves with longer utterances. Motivated by this insight, we propose augmentation strategies to extend speech sample length in this work.

4. Methodology

The block diagram in Figure 1 shows the steps involved in extracting the proposed GoP2Vec and its use in pronunciation score prediction. In the first step, each speech sample length is increased by T times by concatenating the augmented samples obtained from the respective speech. The augmented speech sample is obtained by changing its length and pitch using time-scale modification (TSM) and pitch-scale modification (PSM), respectively. Further, the speech sample also increased by considering the combination of augmented samples from both modifications. In the second step, goodness of pronunciation scores are obtained at the phoneme level using force-aligned phoneme segment boundaries. In the third step, phoneme level GoP scores for the entire increased utterance are converted to a vector (GoP2Vec) using i-vector-based modelling [15]. For each augmentation method, a background GMM and total variability matrix are learned separately and considered for the respective augmented speech sample. In the fourth step, the GoP2Vec is passed to the NN-based score predictor for obtaining a pronunciation quality score.

4.1. Data Augmentation

Generally, the vector obtained from i-vector computation is effective when the length of the speech sample is greater than 5s, i.e., a sequence of more than 500 frames [15]. However, the phoneme level score sequences are much shorter than this requirement. To achieve a longer length sequence, augmentation strategies that do not affect the pronunciation quality are considered for increasing the speech sample length with the concatenation process. The learners speaking with a normal rate range may not affect the pronunciation quality due to minimal changes in the phoneme length. The TSM technique changes

the duration of the sample without affecting the message content and speaker identity. In this work, the augmented samples are generated with TSM, considering the scale ranging from 0.8 to 1.2. The phoneme level scores are independent of the pitch in the speech sample. The PSM technique changes the pitch from one value to another without changing the duration and message content of the speech sample. The augmented speech samples are generated considering a pitch scale ranging from -20Hz to 20Hz. Further, the same time and pitch scale ranges are applied when combined augmentation from both techniques is considered. The augmented samples of these ranges were randomly selected and used in the concatenation process in all three types of length-increasing strategies.

4.2. GoP Score Computation

Typically, the GoP computation methods provide phoneme level scores indicating its pronunciation quality in reference to the respective native speaker's phoneme pronunciation. These methods were developed using the ASR framework, considering phoneme segment boundaries, which were obtained by force-aligning second language learner's speech with native speaker phoneme sequence. For this, the ASR is trained on the native speaker's spoken data and its respective pronunciation lexicon. Further, in the force-aligning, the native phoneme sequence is automatically identified by the ASR when the native pronunciation lexicon is used. We compute GoP scores for second language learners' English speech following the work by Sweekar et al. [5]. The choice is due to its rating-independent state-of-the-art nature in GoP computation and its effectiveness in correlating with ground-truth ratings. Following their work, for the computation, we have considered Kaldi ASR tool kit [26] trained on English speech data from Libri-speech [27] for force-alignment and GoP computation.

4.3. Gop2Vec Computation

Gop2Vec p_l embeds pronunciation quality for l^{th} concatenated speech sample (A_l) is computed as follows:

1) The GoP scores sequence (G_l) for A_l is obtained for each phoneme using the GoP computation indicated as $G_l = \{g_i; 1 \leq i \leq n_l\}$, where n_l is the number of phonemes in A_l .

2) The p_l is computed as, $(\mathbf{I} + \mathbf{V}^T\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{V})^{-1}\mathbf{V}^T\mathbf{\Sigma}^{-1}\mathbf{F}$, where, I is the identity matrix; V is the total variability matrix; $\mathbf{\Sigma}$ is the background GMM covariance matrix; N is the diagonal matrix with diagonal entries as N_k and F is vector with elements as F_k. The matrices V and $\mathbf{\Sigma}$ are obtained from the training data sample, whereas N_k and F_k are specific to the A_l and is computed as, $N_k = \sum_{i=1}^{n_l} \gamma_{k,i}, \quad F_k = \sum_{i=1}^{n_l} \gamma_{k,i} g_i$, where N_k, F_k represents the zeroth and first-order statistics of k^{th} mixture for a given GoP score sequence G_l , and $\gamma_{k,i}$ is the posterior probability of the k^{th} mixture of background GMM given g_i.

3) **Training:** The GMM is trained on the GoP score sequences of all the utterances present in the train set and obtained its parameters: $\mu_k; \sigma_k; \lambda_k \; \forall k \in 1 \leq k \leq K$, where K is total number of mixture components. We obtain diagonal matrix Σ whose diagonal elements are σ_k. Using the super-vector $\mu = [\mu_1, \ldots, \mu_k]^T$ from the background GMM, we obtain the total variability matrix V by solving the following equation iteratively; $\mu_l = \mu + Vq_l + \epsilon_l$, where, the covariance matrix of ϵ_l is approximated by Σ.

Table 1: *Comparison of the proposed approach using three data augmentation strategies with the baselines.*

Dataset	Proposed approach				Baselines	
	# Train samples	TSM	PSM	TSM+PSM	SV	USV
voisTUTOR	150	0.66	**0.69**	0.67	-	0.61
SpeechOcean762	241	0.68	**0.71**	0.69	**0.74**	0.62

4.4. NN-based score Computation

A simple multi-layer perception (MLP) is considered for predicting a score in a supervised manner. The considered MLP layer has the following architecture: The MLP architecture consists of an input layer, followed by two hidden layers with ReLU activation and an output layer. The output layer has r units, where 'r' represents the number of rating classes. During training, the logits are passed through a softmax function to compute class probabilities, and the model is optimized using CrossEntropyLoss. The architectural choices, including the number of hidden layers and hidden units, are determined based on performance on validation set. It is to be noted that the architecture is simple and needs a few samples with labels to predict a score for the pronunciation quality.

Table 2: *Cross-corpus analysis along with varying word lengths*

Train		Test (Correlation)	
	Word length	voisTUTOR	SpeechOcean762
voisTUTOR	1	0.64	0.65
	2-7	0.73	0.74
	>7	0.76	0.77
	All	0.69	0.70
SpeechOcean762	1	0.63	0.66
	2-7	0.72	0.75
	>7	0.75	0.78
	All	0.68	0.71

5. Experimental setup

We conduct the experiments on two datasets: voisTUTOR and SpeechOcean762. From voisTUTOR data, a random 150 out of 12,535 samples are used for training, while the remaining samples serve as the test set. From the SpeechOcean762 data, 241 out of 2,500 training samples are utilized for training while for the testing all 2500 samples in the test set are used. The training samples were chosen to ensure balanced representation of all rating levels. We use Kaldi toolkit for implementing i-vector framework. We consider two baselines: 1. A supervised approach proposed in [6], which integrates GoP features with a Transformer-based architecture. 2. An unsupervised approach proposed in [5], which relies solely on GoP features. We evaluate model performance using the Pearson correlation coefficient [28], which measures the correlation between predicted and actual ratings.

6. Results

In this section, we present the results as follows: 1) Comparison of the baseline and proposed approach; 2) Effect of sentence length under cross and matched conditions and 3) Analysis on GMM components & GoP2Vec dimensions.

6.1. Comparison with baselines:

Table 1 presents the correlations achieved by the proposed approach, which utilizes three data augmentation strategies—TSM, PSM, and TSM+PSM—against two baselines:

Table 3: *Correlation with varying no of Gaussian components and GoP2Vec dimensions for both voisTUTOR & SpeechOcean762 (in brackets) test datasets*

GMM (k)	Correlation					
	GoP2Vec dimension (d)					
	2	**5**	**8**	**20**	**50**	**100**
2	0.59 (0.66)	0.6 (0.65)	0.61 (0.65)	0.62 (0.67)	0.63 (0.67)	0.64 (0.68)
4	0.6 (0.66)	0.62 (0.66)	0.63 (0.67)	0.64 (0.67)	0.65 (0.68)	0.66 (0.68)
8	0.61 (0.67)	0.63 (0.67)	0.65 (0.67)	0.66 (0.68)	0.67 (0.68)	0.67 (0.68)
16	0.63 (0.67)	**0.69** (0.67)	0.68 **(0.71)**	0.67 (0.68)	0.66 (0.68)	0.65 (0.69)
32	0.62 (0.67)	0.68 (0.68)	0.67 (0.68)	0.66 (0.68)	0.65 (0.69)	0.64 (0.69)
64	0.61 (0.68)	0.67 (0.68)	0.66 (0.68)	0.64 (0.69)	0.63 (0.70)	0.62 (0.70)
128	0.60 (0.68)	0.65 (0.68)	0.64 (0.69)	0.62 (0.69)	0.61 (0.70)	0.60 (0.70)
256	0.59 (0.69)	0.63 (0.69)	0.62 (0.69)	0.61 (0.70)	0.60 (0.70)	0.59 (0.71)

the Supervised Baseline (SV) and the Unsupervised Baseline (USV). The results are summarized for two datasets, voisTUTOR and SpeechOcean762.

voisTUTOR: The proposed approach achieves the highest correlation of 0.69 using the PSM strategy. This result outperforms the TSM (0.66) and TSM+PSM (0.67) variations, as well as the USV baseline (0.61). The supervised baseline (SV) is unavailable for the voisTUTOR dataset because it lacks detailed utterance and word level annotations.

SpeechOcean762: A similar trend is observed. The proposed approach with PSM obtains the best correlation of **0.71**, followed by TSM (0.68) and TSM+PSM (0.69). While the proposed method surpasses the USV baseline (0.62), it falls slightly short of the supervised SV baseline, which achieves a correlation of 0.74.

Additionally the supervised baseline was trained on approximately 2500 samples, whereas the proposed approach achieves competitive performance using only the minimal sample sizes indicated in the table. PSM outperforms TSM as it involves modifying the pitch of speech while maintaining the original duration and message content, preserving the overall structure of pronunciation. Since pitch variations are speaker-dependent and do not inherently lead to incorrect pronunciation, PSM allows for more natural speech variations, which might explain its superior performance. On the other hand, TSM alters the duration of the speech while keeping the pitch constant. When the duration of speech is modified, it can disrupt the natural flow and rhythm, potentially leading to degraded speech quality and affecting the perceived accuracy of the pronunciation. This disruption can be particularly detrimental when evaluating the correlation between predicted and actual ratings, which is why TSM may result in lower correlation scores compared to PSM.

6.2. Effect of sentence length on performance:

Table 2 presents the correlation results for sentences with different lengths across both the voisTUTOR and SpeechOcean762 datasets, including both within-dataset (matching) and cross-dataset (training on one dataset and testing on the other) evaluations. The analysis primarily focuses on the performance of the proposed approach, utilizing the PSM-based augmentation strategy, and highlights the impact of sentence length on pronunciation assessment. For the voisTUTOR dataset, the correlation improves as the sentence length increases. The correlation for single-word utterances is 0.63, but for sentences with 2–7 length, the correlation increases to 0.73, showing a 15.9% relative improvement. For sentences with more than 7 length, the correlation further increases to 0.76, with a 20.6% relative improvement compared to single-word utterances. Similarly, for the SpeechOcean762 dataset, the correlation for single-word utterances starts at 0.66, increases to 0.72 for sentences with 2–7 length (a relative improvement of 9.1%), and further increases to 0.75 for words with more than 7 length (a relative improvement of 13.6%).

These results suggest that longer sentences provide richer pronunciation information, leading to more reliable and accurate pronunciation assessments. The cross-dataset analysis also reveals similar trends. For instance, when the model is trained on voisTUTOR and tested on SpeechOcean762, the correlation increases from 0.64 for single-word utterances to 0.74 for sentences with 2-7 length (15.6% relative improvement), and reaches 0.77 for sentences with more than length 7 (20.3% relative improvement). Conversely, when trained on SpeechOcean762 and tested on voisTUTOR, the model shows similar improvements, with correlations rising as word length increases.

Therefore, the results indicate that the proposed approach is not only effective within a single dataset but also exhibits strong cross-dataset generalization, performing well across different word lengths and datasets.

6.3. GMM components & GoP2Vec dimensions analysis:

Table 3 presents the correlation values for pronunciation assessment performance across both the test datasets. The rows represent different values of k (ranging from 2 to 256), while the columns correspond to different values of d (ranging from 2 to 100). Each cell contains two correlation values: the first for the voisTUTOR dataset and the value in parentheses for SpeechOcean762. The results indicate a general trend of increasing correlation as both k and d grow. However, the best performance is observed at k = 16 with d = 5 for voisTUTOR and d = 8 for SpeechOcean762. While higher values of k and d provide similar performance, they are computationally less efficient. Interestingly, the optimal performance occurs when the GoP2Vec dimension matches the number of classes, suggesting that i-vector training may be learning class-specific information in each dimension. These findings emphasize the importance of balancing model complexity and computational efficiency for optimal pronunciation assessment.

7. Conclusion

In this work, we addressed the challenge of automatic pronunciation assessment focusing on reducing the reliance on large amounts of annotated data. We proposed a novel approach that leverages a few data samples by augmenting speech through TSM, PSM and combination of both. We develop a method to obtain a vector (GoP2Vec) representation leveraging the i-vector framework. Our experiments on two datasets demonstrated that the proposed GoP2Vec approach outperforms unsupervised methods and performs on par with supervised approaches when trained with a limited number of samples. Future works include to deduce strategies for better augmentation strategies in GoP2Vec computation for better performance.

8. References

[1] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, University of Cambridge, Department of Engineering, 2000.

[2] J. J. H. C. van Doremalen, C. Cucchiarini, and H. Strik, "Using non-native error patterns to improve pronunciation verification," in *Proceedings of Interspeech – 11th Annual Conference of the International Speech Communication Association (ISCA)*, 2010.

[3] Y. Song, W. Liang, and R. Liu, "Lattice-based gop in automatic pronunciation evaluation," in *The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, 2010.

[4] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and utilization of mllr speaker adaptation technique for learners' pronunciation evaluation," in *Proceedings of Interspeech – 10th Annual Conference of the International Speech Communication Association (ISCA)*, 2009.

[5] 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 *Proceedings of Interspeech – 20th Annual Conference of the International Speech Communication Association (ISCA)*, 2019.

[6] 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 *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7262–7266.

[7] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "A hierarchical context-aware modeling approach for multi-aspect and multi-granular pronunciation assessment," *arXiv preprint arXiv:2305.18146*, 2023.

[8] J. Yang, A. Wumaier, Z. Kadeer, L. Wang, S. Guo, and J. Li, "Attention-cnn combined with multi-layer feature fusion for english l2 multi-granularity pronunciation assessment," in *2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*, pp. 449–457.

[9] Y. Liang, K. Song, S. Mao, H. Jiang, L. Qiu, Y. Yang, D. Li, L. Xu, and L. Qiu, "End-to-end word-level pronunciation assessment with mask pre-training," *arXiv preprint arXiv:2306.02682*, 2023.

[10] H. Ryu, S. Kim, and M. Chung, "A joint model for pronunciation assessment and mispronunciation detection and diagnosis with multi-task learning," in *Proceedings of Interspeech – 24rd Annual Conference of the International Speech Communication Association (ISCA)*, 2023.

[11] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," *arXiv preprint arXiv:2204.03863*, 2022.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[13] C. Yarra, A. Srinivasan, C. Srinivasa, R. Aggarwal, and P. K. Ghosh, "voistutor corpus: A speech corpus of indian l2 english learners for pronunciation assessment," in *22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2019, pp. 1–6.

[14] 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.

[15] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[16] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, vol. 14, 2005.

[17] S. Biswas, J. Rohdin, and K. Shinoda, "i-vector selection for effective plda modeling in speaker recognition," *ODYSSEY, The Speaker and Language Recognition Workshop*, pp. 100–105, 2014.

[18] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *Proceedings of Interspeech – 12th Annual Conference of the International Speech Communication Association (ISCA)*, 2011.

[19] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Proceedings of Interspeech – 12th Annual Conference of the International Speech Communication Association (ISCA)*, 2011.

[20] H. Behravan, V. Hautamäki, and T. Kinnunen, "Factors affecting i-vector based foreign accent recognition: A case study in spoken finnish," *Speech Communication*, vol. 66, pp. 118–129, 2015.

[21] Z. Huang, Y.-C. Cheng, K. Li, V. Hautamäki, and C.-H. Lee, "A blind segmentation approach to acoustic event detection based on i-vector." in *Proceedings of Interspeech – 14th Annual Conference of the International Speech Communication Association (ISCA)*, 2013.

[22] R. Xia and Y. Liu, "Using i-vector space model for emotion recognition." in *Proceedings of Interspeech – 13th Annual Conference of the International Speech Communication Association (ISCA)*, 2012.

[23] J. Silovsky and J. Prazak, "Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4193–4196.

[24] P. Verma and P. K. Das, "i-vectors in speech processing applications: a survey," *International Journal of Speech Technology*, vol. 18, pp. 529–546, 2015.

[25] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7649–7653.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

[27] 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)*, pp. 5206–5210.

[28] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.