



# ASR-Free Pronunciation Assessment

Sitong Cheng<sup>1,2</sup>, Zhixin Liu<sup>1,2</sup>, Lantian Li<sup>1\*</sup>, Zhiyuan Tang<sup>1</sup>, Dong Wang<sup>1\*</sup>, Thomas Fang Zheng<sup>1</sup>

<sup>1</sup>Center for Speech and Language Technologies, Tsinghua University

<sup>2</sup>Beijing University of Posts and Telecommunications

lilt@csl.t.org; wangdong99@mails.tsinghua.edu.cn

## Abstract

Most of the pronunciation assessment methods are based on local features derived from automatic speech recognition (ASR), e.g., the Goodness of Pronunciation (GOP) score. In this paper, we investigate an ASR-free scoring approach that is derived from the marginal distribution of raw speech signals. The hypothesis is that even if we have no knowledge of the language (so cannot recognize the phones/words), we can still tell how good a pronunciation is, by comparatively listening to some speech data from the target language. Our analysis shows that this new scoring approach provides an interesting correction for the phone-competition problem of GOP. Experimental results on the ERJ dataset demonstrated that combining the ASR-free score and GOP can achieve better performance than the GOP baseline.

**Index Terms:** pronunciation assessment, speech processing, normalization flow

## 1. Introduction

Automatic pronunciation assessment plays a key role in Computer Assisted Language Learning (CALL). The research in this field dates back to 90's in the last century [1, 2], and has gained much popularity in second language education, e.g., [3, 4, 5, 6, 7].

Most of existing assessment approaches are based on automatic speech recognition (ASR). A simple approach is to count the words correctly recognized per minute (WCPM). A more careful assessment is often conducted in three steps: (1) Employ an ASR system to segment speech signals into pronunciation units, e.g., phones or words; (2) Compute local or global features based on the segmentation; (3) Derive proficiency scores from the local/global features. The features that have been used can be categorized into two classes: phonetic features and prosodic features. Phonetic features reflect the quality of the pronunciation of single phones (or words). The most successful phonetic features are the phone-level likelihood or posterior [1, 2], others include phone duration, formants, articulation class, phonetic distance, etc [8, 9]. Prosodic features reflect the intonation, stress or fluency. The most popular prosodic features include the pitch and intensity contour, rate of speech (phones per minute), and duration of silence [10, 11]. Deriving pronunciation scores from these features can be accomplished by a simple average (e.g., if the features are phone-level posteriors), or a complex regression model (e.g., if the features are pitch and intensity contour) such as Gaussian process [12]. More technical details for pronunciation assessment can be found in Witt's review paper [13].

The Goodness of Pronunciation (GOP) is perhaps the most popular features for pronunciation assessment [14]. GOP is

based on the posterior probability on the correct phone, given the speech segment of that phone. This is formulated by:

$$GOP = \frac{1}{M} \sum_i^M \ln p(q_i | o_i), \quad (1)$$

where  $q_i$  is the  $i$ -th phone in the speech segment, and  $o_i$  is the corresponding speech segment, and  $M$  is the total number of phones in the speech segment.

Early research computes GOP using Gaussian mixture model-Hidden Markov model (GMM-HMM) [14], and recent study usually uses acoustic models based on deep neural networks (DNNs) [15, 16, 17, 18]. The DNN-based acoustic modeling offers much more robustness against ambient complexity and speaker variation, which in turn leads to better phone segmentation and posterior estimation. Due to the high performance and simple computation, we use the DNN-based GOP as the baseline.

In spite of the prominent success of the existing approaches, in particular those based on GOP, the present studies heavily rely on ASR (for segmentation and posterior estimation). For L2 learners, especially those in the primary stage, the pronunciations tend to be significantly different from native pronunciations, resulting in high alignment/recognition errors and low quality of pronunciation assessment [19]. This performance reduction could be more severe for languages without strong ASR systems.

This inspired us to investigate an ASR-free scoring approach for pronunciation assessment, by which the assessment is not based on ASR and so would not be impacted by the ASR performance. Interestingly, human beings seem to use this way to judge the proficiency of a pronunciation. Our experience is that even if we have no knowledge of the language (so cannot recognize the phones/words), we can still tell how good a pronunciation is, by listening to some native speech samples from the target language. In this paper, we present an ASR-free scoring approach that does not rely on ASR but is based on a generative model  $p(\mathbf{o})$ . Our theoretical study shows that this score offers an interesting correction for the phone-competition problem of GOP, and empirical study demonstrates that better performance can be achieved when combining the ASR-free score and GOP.

The rest of the paper is organized as follows. Section 2 analyzes the potential problem of GOP, and Section 3 presents the ASR-free scoring approach. Section 4 presents the experimental results on the ERJ dataset, and Section 5 concludes the paper.

## 2. GOP is not perfect

Given a phone sequence  $\mathbf{q}$  and the corresponding speech signal  $\mathbf{o}$ , and assuming a perfect alignment, then the averaged conditional probability will be the theoretically sound measurement for testing how well a test speech matches the training speech:

\* Corresponding authors

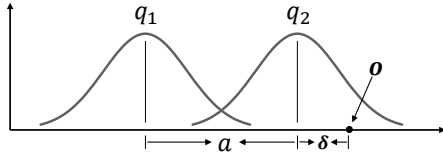


Figure 1: Two conditional distributions for phone  $q_1$  and  $q_2$ . Both are Gaussians with variance 0.5 and the distance of their means is  $a$ . A small change  $\delta$  on the pronunciation in  $q_2$  may lead to either increase or decrease on the posterior, depending on the sign of  $\delta$ .

$$\ln p(\mathbf{o}|\mathbf{q}) = \frac{1}{M} \sum_i \ln p(o_i|q_i). \quad (2)$$

Considering the ideal case where the model is trained on native speech, and the only difference between the test speech and the training speech is pronunciation proficiency, then the conditional above will be perfect for pronunciation assessment. In real situations, however,  $p(\mathbf{o}|\mathbf{q})$  may be varied by numerous factors that are not related to pronunciation proficiency, e.g., noise, channel, speaking rate/volume and speaker trait. This sensitivity to multiple variations means that the conditional is not reliable for pronunciation assessment in real situations.

We can rewrite the conditional into another form:

$$\ln p(\mathbf{o}|\mathbf{q}) = \frac{1}{M} \sum_{i=1}^M \ln \frac{p(q_i|o_i)p(o_i)}{p(q_i)}. \quad (3)$$

If we ignore the prior  $p(q_i)$ , the conditional involves two parts: the posterior  $p(q_i|o_i)$  and the marginal  $p(o_i)$ . The marginal part inherits the property of the conditional and is sensitive to all variations including pronunciation proficiency. The posterior part, which is essentially GOP, however, is more robust. It is purely phone discriminative, and so less sensitive to phone-unrelated variations. This insensitivity offers an important merit for pronunciation assessment and explains the success of GOP.

A potential problem of the GOP score is that there is no guarantee that a worse pronunciation will achieve a smaller posterior. Let's design a simulation experiment to discuss this issue. As shown in Figure 1, we assume two phones  $q_1$  and  $q_2$  are two one-dimensional Gaussians whose variances are both 0.5, and the distance of their means is  $a$ . At the mean of  $q_2$ , which can be regarded as a perfect pronunciation of  $q_2$ , the posterior on  $q_2$  is  $p(q_2|o) = \frac{1}{1+e^{-a^2}}$ . Assume a non-native speaker pronounce  $q_2$  at a position  $\mathbf{o}$ , and the shift from the mean of  $q_2$  to  $\mathbf{o}$  is  $\delta$ . The posterior on  $q_2$  given  $\mathbf{o}$  can be computed as follows:

$$p(q_2|\mathbf{o}) = \frac{e^{-\delta^2}}{e^{-\delta^2} + e^{-(a+\delta)^2}} = \frac{1}{1 + e^{-(a^2+2a\delta)}}. \quad (4)$$

It can be seen that the change of the posterior depends on the sign of  $\delta$ : if  $\delta > 0$ , the posterior essentially increases. This means that a non-native speaker obtains a better GOP than a native speaker. It clearly demonstrates that GOP is not a perfect score, at least in theory.

### 3. Proposed methods

#### 3.1. ASR-free scoring

The undesired behavior of GOP is essentially caused by the *competition between phones*. Recalling that the conditional

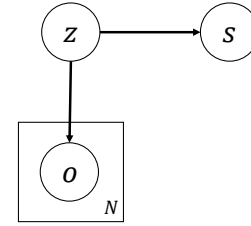


Figure 2: The graphical model for ASR-free score estimation, where  $\mathbf{o}$  is the speech segment that consists of  $N$  frames,  $\mathbf{z}$  is the utterance-level representation of  $\mathbf{o}$ ,  $s$  is the assessment score.

$p(\mathbf{o}|\mathbf{q})$  in Eq. (2) is a perfect assessment and does not suffer from this problem, we conjecture that it is the marginal part  $p(\mathbf{o})$  that solves the phone competition. Therefore, we argue that  $p(\mathbf{o})$  should be involved in the assessment, rather than simply discarded. Since  $p(\mathbf{o})$  concerns neither phones nor words, it is called an **ASR-free score**.

However, simply multiplying  $p(\mathbf{q}|\mathbf{o})$  by  $p(\mathbf{o})$  as in Eq. (2) does not work, as  $p(\mathbf{o})$  is quite noisy. To demonstrate this argument, we trained a GMM model using the WSJ native English dataset and computed the Pearson correlation coefficient (PCC) between the marginal distribution  $p(\mathbf{o})$  and the human-labelled scores on the ERJ (English read by Japanese) dataset, and found that the PCC is nearly zero or even negative. More details will be presented in Section 4.

In order to employ  $p(\mathbf{o})$  but reduce the noise, a discriminative model can be used to discover the factors  $\mathbf{z}_s$  in  $\mathbf{o}$  that are mostly related to pronunciation proficiency, and then build  $p(\mathbf{z}_s)$  rather than  $p(\mathbf{o})$ . This approach, however, requires the native speech being labelled by pronunciation proficiency, which is almost impossible. There is a more direct and parsimonious way: since our goal is to assess the pronunciation, we can build a **conditional model**  $p(s|\mathbf{z})$  that not only selects the factor  $\mathbf{z}_s$ , but also produces the assessment score  $s$  at the same time.

We therefore propose a probabilistic model shown in Figure 2. Firstly, we build a **marginal model**  $p(\mathbf{o})$  that describes all the variations in  $\mathbf{o}$ . This model, however, is not used to compute the marginal probability; instead, it is used to infer the utterance-level representation  $\mathbf{z}$  for the speech  $\mathbf{o}$ . Secondly, we build a **prediction model**  $p(s|\mathbf{z})$  that selects  $\mathbf{z}_s$  from  $\mathbf{z}$  and produces the assessment score  $s$ . This step is a supervised training and a small amount of training data is sufficient. Combining the two models, we can build the conditional model  $p(s|\mathbf{o})$ .

#### 3.2. Marginal model

There are several ways to build the marginal model  $p(\mathbf{o})$ , but not all of them can infer the latent representation  $\mathbf{z}$  in a compact way. We consider three models: i-vector, normalization flow (NF) and discriminative normalization flow (DNF).

##### 3.2.1. i-vector model

The i-vector model is a mixture of linear Gaussians [20]:

$$p(\mathbf{o}) = \sum_k \pi_k N(\mathbf{o}|\mathbf{T}_k \mathbf{z}, \mathbf{\Sigma}_k),$$

$$p(\mathbf{z}) = N(\mathbf{z}; \mathbf{0}, \mathbf{I}),$$

where  $k$  indexes the Gaussian component, and the loading matrices  $\{\mathbf{T}_k\}$  are low-rank. This is a full-generative model, and

the posterior on  $\mathbf{z}$  can be easily computed given a speech signal  $\mathbf{o}$ . The mean of the posterior  $p(\mathbf{z}|\mathbf{o})$  is the i-vector.

The i-vector model possesses several advantages: (1) The model is trained in an unsupervised way and therefore can exploit rich unlabelled data; (2) The Gaussian assumption makes the training and inference simple; (3) The form of Gaussian mixtures absorbs the impact of short-time variations (e.g., speech content), which makes the i-vectors represent the long-term variations. For these reasons, the i-vector model has been widely employed in various speech processing tasks, such as speaker recognition [20] and language recognition [21, 22].

### 3.2.2. Normalization flow

The i-vector model is a shallow model, which may prevent it from representing complex distributions. Normalization flow (NF) is a deep generative model [23] and can describe more complex distributions. The foundation of NF is the principle of distribution transformation for continuous variables [24]. Let a latent variable  $\mathbf{z}$  and an observation variable  $\mathbf{o}$  be linked by an invertible transform  $\mathbf{o} = f(\mathbf{z})$ , their probability density has the following relationship:

$$\ln p(\mathbf{o}) = \ln p(\mathbf{z}) + \ln \left| \det \frac{\partial f^{-1}(\mathbf{o})}{\partial \mathbf{o}} \right|, \quad (5)$$

where  $f^{-1}(\mathbf{o})$  is the inverse function of  $f(\mathbf{z})$ . It has been shown that if  $f$  is flexible enough, a simple distribution (a standard Gaussian) can be transformed to a complex distribution. Usually,  $f$  is implemented as a composition of a sequence of relatively simple invertible transforms [25].

Once the model has been trained, for a speech segment  $\mathbf{o}$ , the latent variable  $\mathbf{z}$  can be inferred by averaging the image of  $\mathbf{o}_i$  in the latent space:

$$\mathbf{z} = \frac{1}{N} \sum_{i=1}^N f^{-1}(\mathbf{o}_i) = \frac{1}{N} \sum_{i=1}^N \mathbf{z}_i. \quad (6)$$

### 3.2.3. Discriminative NF

The vanilla NF model optimizes the distribution of the training data without considering the class labels. This means that data from different classes tend to congest together in the latent space, as shown in the top row of Figure 3. This is not a good property for downstream applications that require discrimination within the latent space, e.g., pronunciation proficiency.

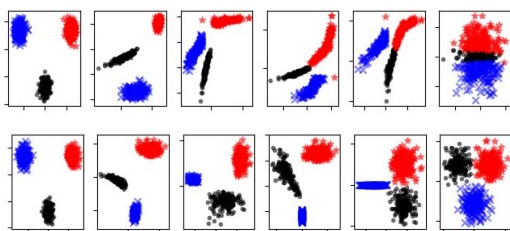


Figure 3: NF (top) pulls all classes together in the latent space. DNF (bottom) keeps data from different classes separated.

Recently, the authors proposed a discriminative NF (DNF) model to deal with this problem [26]. The main advantage of DNF is that it allows each class to have its own Gaussian prior, i.e. all the priors share the same covariance but possess different means, formulated as follows:

$$p_s(\mathbf{z}) = N(\mathbf{z}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}),$$

where  $s$  is the class label. By setting class-specific means, different classes will be separated from each other in the latent space, as shown in the bottom row of Figure 3. If we treat data with different pronunciation proficiency as different classes and train DNF as mentioned above, it will be possible to establish a latent space which is more discriminative for pronunciation assessment. Once the model has been trained, the same averaging approach Eq. (6) as in NF can be used to derive the utterance-level representation  $\mathbf{z}$ .

### 3.3. Prediction model

Modeling the prediction probability  $p(s|\mathbf{z})$  is easy, if  $\mathbf{z}$  has been derived from the marginal model  $p(\mathbf{o})$ . Linear regression is the most popular choice, but we found support vector regression (SVR) often produces better results. Compared to linear regression, SVR chooses a few examples that hold the best predictive power as support vectors. It is argued that this model is more robust against data sparsity and data imbalance. With SVR, we predict the score  $s$  directly, which can be regarded as a special form of the prediction distribution  $p(s|\mathbf{z})$  where all the probability mass concentrates in a single value.

### 3.4. Information fusion

According to Eq. (2), a perfect assessment score should combine the posterior part  $p(\mathbf{q}|\mathbf{o})$  (which is the GOP score) and the marginal part  $p(\mathbf{o})$  (which is the ASR-free score). As we mentioned,  $p(\mathbf{o})$  is very noisy and so we model  $p(s|\mathbf{o})$  instead. Since  $p(\mathbf{o})$  and  $p(s|\mathbf{o})$  model different uncertainties, simply substitute  $p(\mathbf{o})$  for  $p(s|\mathbf{o})$  in Eq. (2) is not theoretically correct. There are two approaches to combine  $p(\mathbf{q}|\mathbf{o})$  and  $p(s|\mathbf{o})$ : score fusion and feature fusion.

#### 3.4.1. Score fusion

In score fusion, we treat the GOP score  $p(\mathbf{q}|\mathbf{o})$  and the ASR-free score  $p(s|\mathbf{o})$  as two independent scores, and interpolate them in the assessment:

$$s^* = \lambda p(\mathbf{q}|\mathbf{o}) + (1 - \lambda) \arg \max_s \{p(s|\mathbf{o})\}. \quad (7)$$

Since SVR outputs the optimal  $s$ , this fusion score reduces to a simple form:

$$s^* = \lambda p(\mathbf{q}|\mathbf{o}) + (1 - \lambda) \gamma(\mathbf{o}), \quad (8)$$

where  $\gamma(\cdot)$  is the prediction function implemented by SVR.

#### 3.4.2. Feature fusion

In the feature fusion, we treat the GOP score  $p(\mathbf{q}|\mathbf{o})$  as a feature and combine it with the latent representation  $\mathbf{z}$ , and then build the SVR model. This feature fusion may discover valuable knowledge in  $\mathbf{z}$  that has not been represented by  $p(\mathbf{q}|\mathbf{o})$ . It produces good performance in our experiments.

## 4. Experiments

### 4.1. Data

Two datasets were used in our experiments: WSJ (Wall Street Journal) and ERJ (English Read by Japanese) [27].

**WSJ:** A native English speech dataset. It contains 37,318 utterances from 282 speakers. The dataset was used to train the DNN-HMM ASR system to build the GOP baseline.

**ERJ:** A standard Japanese-speaking-English dataset. It consists of 1,900 utterances from 190 speakers. Each utterance has 5 human-labelled pronunciation scores. There are 5-scale ratings as the pronunciation proficiency scores. In our experiments, the dataset was separated into two subsets: *ERJ.Train* and *ERJ.Eval*. *ERJ.Train* consists of 1,520 utterances and was used to train the marginal model (i-vector, NF, DNF) and the prediction model (SVR). *ERJ.Eval* consists of 380 utterances and was used for performance tests. The Pearson correlation coefficient (PCC) on *ERJ.Eval* of human-labelled scores is 0.550.

## 4.2. Model Settings

**DNN-HMM:** It was built using the Kaldi toolkit [28], following the WSJ s5 nnet3 recipe. The DNN structure consists of 6 time-delay layers, each followed by a P-norm layer that reduces the dimensionality of the activations from 2,048 to 256. The input features are 23-dimensional Fbanks with the context window of 21 frames, and the output layer contains 3,360 units, corresponding to the number of GMM-HMM senones. This model was used to create the GOP baseline as Eq. (1).

**GMM:** It was created using the Kaldi toolkit, following the SITW recipe. The raw features involve 23-dimensional MFCCs plus the log energy, augmented by first- and second-order derivatives, resulting in a 72-dimensional feature vector. The number of Gaussian components is set to 64. Once trained, the log-likelihood of each frame  $p(o_i)$  can be estimated, and the utterance-level  $p(\mathbf{o})$  can be computed by a simple average.

**i-vector:** The data preparation is the same as GMM. The number of Gaussian components of the UBM is 64, and the dimensionality of the i-vector is 440.

**NF:** It was trained using the PyTorch toolkit. The NF used here is a RealNVP architecture [29]. This model has 10 non-volume preserving (NVP) layers. The input features are 40-dimensional Fbanks with the effective context window of 11 frames, leading to a 440-dimensional feature vector. The Adam optimizer [30] is used to train the model with the learning rate set to 0.001. Similar to GMM, the utterance-level  $p(\mathbf{o})$  can be computed. In addition, the utterance-level representation  $\mathbf{z}$  of the speech  $\mathbf{o}$  can be derived by Eq. (6).

**DNF:** The data preparation and model structure are the same as NF. The number of classes is set to 5 corresponding to the 5-scale pronunciation proficiency scores in the ERJ dataset. The utterance-level representation  $\mathbf{z}$  can also be derived by Eq. (6).

**SVR:** It was implemented using the Scikit-Learn toolkit with the default SVR configuration.

## 4.3. Basic results

Basic results on the *ERJ.Eval* dataset are reported in Table 1. The Pearson correlation coefficient (PCC) is used to measure the correlation between the assessment score and the human-labelled scores. It can be observed that the performance of the GOP score outperforms the human-labelled scores (0.614 vs. 0.550), indicating the robustness of the GOP approach. We also trained GMM and NF models and computed the PCC between the marginal  $p(\mathbf{o})$  and the human-labelled scores. It can be found that the PCCs on both GMM and NF are nearly zero and even negative. This confirms that  $p(\mathbf{o})$  is quite noisy and cannot be directly employed as the assessment score.

Table 1: Performance (PCC) on the baseline models.

	Human	GOP	GMM	NF
PCC	0.550	<b>0.614</b>	-0.065	-0.131

## 4.4. ASR-free scoring

This experiment examines the performance of our proposed ASR-free models. Three marginal models (i-vector, NF and DNF) plus its individual prediction model (SVR) are trained on the *ERJ.Train* dataset. The results are reported in Table 2. Firstly, it can be seen that the performance of these ASR-free models is inferior to GOP, but is still satisfying. This indicates that combining the marginal model  $p(\mathbf{o})$  and the prediction model  $p(s|\mathbf{z})$  is a feasible way to estimate the conditional model  $p(s|\mathbf{o})$  and produce the reasonable assessment score. Secondly, we observed that NF + SVR outperforms i-vector + SVR. The reason is that NF is a deep generative model, compared with the linear-shallow i-vector model, and can better represent the marginal  $p(\mathbf{o})$ . Besides, DNF + SVR obtains the best performance. This demonstrates that DNF can learn more pronunciation-discriminative representations, which are more suitable for the downstream assessment prediction.

Table 2: Performance (PCC) on the ASR-free models.

	i-vector + SVR	NF + SVR	DNF + SVR
PCC	0.434	0.441	<b>0.462</b>

## 4.5. Information fusion

As we mentioned, the posterior  $p(\mathbf{q}|\mathbf{o})$  (the GOP score) and the marginal  $p(\mathbf{o})$  (the ASR-free score) are complementary, and could be combined by a score fusion or feature fusion. The results are reported in Table 3. The optimal hyper-parameter  $\lambda$  was selected based on a small development set which was randomly sampled from the *ERJ.Train* dataset. Experimental results show that the two fusion approaches consistently achieve better performance than the GOP baseline (0.614). This indicates that these ASR-free approaches can discover some valuable knowledge that has not been represented by GOP, and also demonstrates that the marginal  $p(\mathbf{o})$  involved in the assessment score can relieve the phone-competition problem of GOP.

Table 3: Performance (PCC) with information fusion.

	Score-fusion	Feature-fusion
GOP + i-vector	0.640 ( $\lambda = 0.38$ )	0.625
GOP + NF	0.663 ( $\lambda = 0.34$ )	0.656
GOP + DNF	0.676 ( $\lambda = 0.36$ )	0.667

## 5. Conclusions

This paper proposed an ASR-free scoring approach that does not rely on ASR but is based on a generative model. Our theoretical study shows that this scoring approach offers an interesting correction for the phone-competition problem of GOP, and empirical study demonstrated that combining the GOP and this ASR-free approach can achieve better performance than the GOP baseline. Future work will be conducted to understand the behavior of different generative models in the ASR-free assessment, and study more reasonable fusion approaches to combine the posterior and the marginal.

## 6. Acknowledgements

This work was supported by the National Natural Science Foundation of China No. 61633013.

## 7. References

- [1] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1457–1460.
- [2] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *1997 IEEE international conference on acoustics, speech, and signal processing*, vol. 2. IEEE, 1997, pp. 1471–1474.
- [3] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.
- [4] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.
- [5] J. Cheng, "Real-time scoring of an oral reading assessment on mobile devices," in *Proc. Interspeech 2018*, 2018, pp. 1621–1625. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-34>
- [6] A. Miwardelli, I. Gallagher, J. Gibson, N. Katsos, K. M. Knill, and H. Wood, "Splash: Speech and Language Assessment in Schools and Homes," in *Proc. Interspeech 2019*, 2019, pp. 972–973.
- [7] C. Yarra, A. Srinivasan, S. Gottimukkala, and P. K. Ghosh, "SPIRE-fluent: A self-learning app for tutoring oral fluency to second language English learners," in *Proc. Interspeech 2019*, 2019, pp. 968–969.
- [8] K. Kyriakopoulos, K. M. Knill, and M. J. Gales, "A deep learning approach to automatic characterisation of rhythm in non-native English speech," in *Proc. Interspeech 2019*, 2019, pp. 1836–1840. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-3186>
- [9] S. Jenne and N. T. Vu, "Multimodal articulation-based pronunciation error detection with spectrogram and acoustic features," in *Proc. Interspeech 2019*, 2019, pp. 3549–3553. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-1677>
- [10] Q.-T. Truong, T. Kato, and S. Yamamoto, "Automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours," in *Proc. Interspeech 2018*, 2018, pp. 2186–2190. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-1386>
- [11] C. Graham and F. Nolan, "Articulation rate as a metric in spoken language assessment," in *Proc. Interspeech 2019*, 2019, pp. 3564–3568. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-2098>
- [12] K. Knill, M. Gales *et al.*, "Automatically grading learners' English using a Gaussian process." ISCA, 2015.
- [13] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *Proceedings of International Symposium on automatic detection on errors in pronunciation training*, vol. 1, 2012.
- [14] S. M. Witt *et al.*, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, University of Cambridge Cambridge, United Kingdom, 1999.
- [15] 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 *Proc. Interspeech 2019*, 2019, pp. 954–958. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-2363>
- [16] 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.
- [17] 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.
- [18] 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.
- [19] K. Knill, M. Gales, K. Kyriakopoulos, A. Malinin, A. Ragni, Y. Wang, and A. Caines, "Impact of ASR performance on free speaking language assessment," in *Proc. Interspeech 2018*, 2018, pp. 1641–1645. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-1312>
- [20] 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, 2011.
- [21] N. Dehak, P. A. Torrescarrasquillo, D. A. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction." pp. 857–860, 2011.
- [22] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Twelfth annual conference of the international speech communication association*, 2011.
- [23] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *arXiv preprint arXiv:1912.02762*, 2019.
- [24] W. Rudin, *Real and complex analysis*. Tata McGraw-hill education, 2006.
- [25] E. G. Tabak and C. V. Turner, "A family of nonparametric density estimation algorithms," *Communications on Pure and Applied Mathematics*, vol. 66, no. 2, pp. 145–164, 2013.
- [26] Y. Cai, L. Li, D. Wang, and A. Abel, "Deep normalization for speaker vectors," *arXiv preprint arXiv:2004.04095*, 2020.
- [27] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," in *Proc. ICA*, vol. 1, 2004, pp. 557–560.
- [28] 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*, 2011.
- [29] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," 2017.
- [30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.