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

Automatic assessment of pronunciation and its dependent factors by exploring their interdependencies using DNN and LSTM

Aparna Srinivasan, Chiranjeevi Yarra, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India

{aparnasrinivasa,chiranjeeviy,prasantg}@iisc.ac.in

Abstract

In the applications of computer assisted language learning, it is important to assess the pronunciation quality of second language learners in an automatic manner. Typically, this assessment is posed as a classification problem wherein the overall pronunciation quality is estimated at discrete levels. For classification, features are heuristically computed for an entire utterance considering factors influencing the pronunciation quality. However, the heuristic computation at the utterance level could not help in exploring the interdependencies between the factors and their effect at the sub-segment level. In this work, we learn the interdependencies between the factors by jointly modeling the labels representing the qualities of factors as well as pronunciation. Further, we also consider sub-segment level features for modeling. Experiments are conducted on data collected from Indian learners, considering the accuracy between the estimated qualities and the human expert ratings as performance measure. The highest improvements are found to be 19.13% and 14.93% (relative) when the proposed joint model is used with sub-segment and utterance level features respectively, and are compared to that of the baseline scheme without using a joint model.

Index Terms: Pronunciation assessment, Joint modeling, Sequential modelling, Sub-segment level features

1. Introduction

In second language (L2) learning, it is important to acquire the ability to pronounce words and sentences in the L2 language correctly. This is because, incorrect pronunciation causes miscommunication. The L2 learners often acquire good pronunciation skills with effective training methods. In the applications of computer assisted language learning (CALL) where computer aided pronunciation training (CAPT) is an important module, it has been shown that the training could be provided in an automatic manner [1]. For this the pronunciation quality of L2 learners is assessed automatically. Generally, in almost all the applications related to CALL, the task of automatic assessment is a necessary component.

Usually, pronunciation is assessed in terms of either a score or a set of ratings representing its quality. For example, Deshmukh et al. assessed the pronunciation using a set containing ratings from one to four in steps of one, where one and four denote poor and good pronunciation quality [2]. Hacker et al. used a set of one to five integer values for the assessment task [3]. Kim et al. [4] used a set containing one and two, to indicate that the pronunciation is not intelligible and intelligible respectively. Similarly, Xiao et al. and Dutta et al. used a set containing one and two for representing high and low nativity influence in the pronunciation [5, 6].

In order to assess the pronunciation at discrete levels, typically, classifier based approaches are considered. Franco et al. used a Deep Neural Network (DNN) based classifier trained with features based on phoneme duration and the phoneme posterior probabilities [7]. Deshmukh et al. used a logistic regression based classifier with features consisting of the averaged log posterior probability of phonemes [2]. Suzuki et al. trained a DNN classifier with heuristically computed distance matrix between the phoneme posterior probabilities of the learner and expert that are aligned with dynamic time warping [8]. Xiao et al. trained a DNN classifier with features obtained by augmenting the averaged frame level log posterior probabilities from learner's and expert's utterance [5]. Nikolav et al. performed pronunciation assessment at the phoneme level using a regression tree followed by a simple threshold using phoneme posterior probabilities [9].

In addition to the features based on posterior probability, features have also been computed based on the factors influencing the pronunciation quality like intelligibility, syllable stress, intonation and fluency [10]. Hacker et al. considered additional features which were computed based on rate of speech and duration of pauses [3]. Dutta et al. used mel frequency cepstral coefficients (MFCCs) as the additional feature and support vector machines as the classifier [6]. Cincarek et al. considered additional features based on measure of phoneme errors (insertion, deletions and substitutions) and mean phone duration [11]. Tepperman et al. used a binary decision tree classifier trained with features including word durations [12]. In almost all these works, the features for an utterance are computed heuristically by applying statistics on the word or phoneme level features.

Apart from these works on pronunciation assessment, there also exist other works that estimate factors like intonation, nativity, intelligibility [13, 14, 5, 6, 4] etc. and hypothesize that the estimated factors could be useful in the assessment task. However, these factors have been estimated independently from the assessment task. Though the influences from these factors are represented as features for assessing an utterance, these features have been combined together in a heuristic manner to obtain an utterance level feature. Such heuristic approaches used in feature computation might not adequately represent the factors in the task of pronunciation assessment. Further, there exists no work that models the interdependencies between the factors and the pronunciation quality in a data driven manner while assessing the pronunciation quality.

In this work, we proposed to learn the interdependencies of the factors and the pronunciation quality using a joint DNN model trained with the ratings representing factors' and overall pronunciation quality of each utterance. The factors considered include intelligibility, phoneme quality, phoneme mispronunciation, syllable stress, intonation, pause placement and mother tongue influence (MTI) which have been shown to affect the overall pronunciation quality [15, 16, 17, 18, 19, 20]. In order to overcome the heuristic computations involved in the utterance level feature, we propose to model sub-segment level features directly using long-short term memory (LSTM) networks. For modeling, we propose a pair of features augmented with baseline features that have been computed following the

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

10.21437/SLaTE.2019-7

work by Xiao et al. [5]. Experiments are conducted on the data collected from Indian learners and considering the work proposed by Xiao et al. [5] as the baseline scheme, where baseline features alone were modelled using DNNs. The accuracies in predicting the rating for overall pronunciation quality with the proposed LSTM and DNN based joint models and the proposed augmented features are found to be higher than the baseline with relative improvements of 19.13% and 14.93% respectively. Further, the improvement in the accuracies with the proposed LSTM and DNN based joint models are compared with LSTM and DNN based individual models to shows the effectiveness of modeling the interdependencies.

2. Database

In this work, we consider a read English corpus collected from 16 Indian learners who were in spoken English training at the time of the recording. Due to the language diversity in India, we consider the learners from six different native languages – Malayalam, Kannada, Telugu, Tamil, Hindi and Gujarati. There are a total of 4 (3+1), 5 (1+4), 3 (2+1), 2 (2+1), 1 (0+1) and 1 (0+1) speakers (male + female) from each of these languages respectively. All the learners were either undergraduate or postgraduate students whose age ranged from 19 to 25. There is a total of 12375 utterances present in the corpus and approximately 800 utterances per subject. A spoken English expert with 25 years of training experience, manually rated each utterance on a scale of 5 to 1 for the overall pronunciation quality, where the rating 5, 4, 3, 2 and 1 indicate excellent, good, average, poor and incorrect pronunciation respectively. From the ratings for all 12375 utterances, 2513, 2595, 2916, 2143 and 2208 utterances are assigned with rating 1, 2, 3, 4, and 5 respectively. In this process, the expert also indicated a binary rating on the quality of the seven factors influencing the overall pronunciation as follows – 1) intelligible (1) or not (0), 2) phoneme quality is good (1) or not (0), 3) phoneme mispronunciation exists (1) or not (0), 4) syllable stress is proper (1) or not (0), 5) intonation is proper (1) or not (0), 6) pause locations are proper (1) or not (0) and 7) MTI is present (1) or not (0). For these factors, the percentage of utterances with rating one are 88.50, 68.71, 49.25, 37.42, 62.26, 81.21 and 57.60 respectively. Further, 1200 utterances were randomly repeated in order to know the consistency of the expert. The expert was found to have more than 70% consistency in the ratings of the repeated stimuli which is closer to the value observed in other databases [21, 22]. Further, the 800 unique utterances were also recorded from the expert.

3. Proposed Approach

The block diagram in Figure 1 shows the two major steps involved in the proposed approach. The first step computes either an utterance level feature (f_{utt}) or a sub-segment level feature sequence (f_{seg}) for a given utterance. The second step predicts the overall and factor-specific quality ratings considering a classifier that jointly models both the qualities separately using – 1) f_{utt} with single layer neural networks (SLNNs) and 2) f_{seg} with LSTMs and SLNNs. The joint model has four stages, namely, shared layer, factor-specific layer, overall quality layer and soft-max layer. The shared layer takes either f_{utt} or f_{seg} as input and estimates a common representation ϕ^s. For f_{utt}, it uses SLNNs and for f_{seg} it uses LSTMs. From ϕ_s, the factor-specific layer predicts representations $\phi_i^{fs}, 1 \leq i \leq n$, where n is the total number of factors. The overall quality layer obtains representations ϕ^{oq} for overall quality considering both ϕ^s and $\{\phi_i^{fs}, 1 \leq i \leq n\}$. The soft-max layer predicts rating specific posterior probabilities $P(R_i^{fs}|\phi_i^{fs}) \; \forall \; i$ and $P(R^{oq}|\phi^{oq})$ for each factor and overall quality from the respective ϕ_i^{fs} and ϕ^{oq}. Finally, the rating with the highest probability in each factor and overall quality is considered as the predicted rating for those qualities.

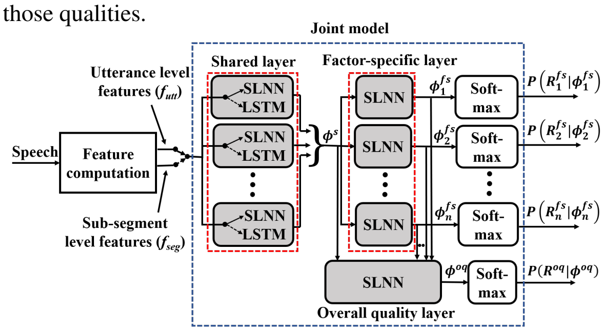

Figure 1: *Block diagram showing the steps involved in the proposed approach.*

3.1. Feature computation

The f_{utt} and f_{seg} are computed based on the frame level logarithm of posterior probability values from all phoneme models, referred to as log posteriors. Figure 2 shows the log posteriors and its average for two utterances of an exemplary sentence "draughts man", whose overall ratings are one and five respectively. The average is performed across the frames in an utterance and referred to as utterance level averaging, which is considered to obtain the feature for an utterance in most of the works related to pronunciation assessment [2]. For the exemplary sentence, the phonemes in the canonical pronunciation are "d r aa f t s m ae n". In general, the log posteriors within the aligned boundaries of a canonical phoneme are consistently high, when they are obtained from the respective phoneme model. For example, for the canonical phonemes "s" and "n" in the figure, it is observed that the log posteriors have high values (highlighted in green boxes) with respect to their corresponding phoneme models. However, the strength of these values is reduced, when the utterance is influenced by the learner's nativity. This reduction spreads across other phoneme models based log posteriors and it could be discriminative for the assessment of nativity influences. Xiao et al. explored this for the classification of native and non-native speech [5] considering utterance level averaged log posterior as the feature. However, it is observed that the feature could be insufficient for obtaining good performance in pronunciation assessment task due to utterance level averaging.

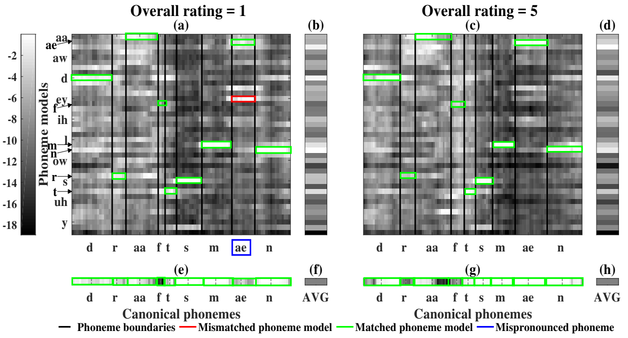

Figure 2: *An illustrative example showing (a), (c) log posteriors from all phoneme models and (b), (d) their utterance level average (AVG), and (e), (g) log posteriors from matched phoneme models and (f),(h) their average for overall rating of 1 and 5 respectively. Though the log posteriors are computed for 39 phoneme models, we indicate only 15 phonemes in the y-axis to highlight selected locations.*

3.1.1. Motivation

From Figure 2a it is observed that, the log posteriors have consistently high values from the phoneme model “ey” (highlighted in red box) within the boundaries of the canonical phoneme “ae”. This mismatch between the canonical phoneme and the phoneme model could be indicative of incorrect phoneme pronunciation, which is believed to be the cause for obtaining overall rating of one. On the other hand, the matched condition is an indicator of correct phoneme pronunciation. From Figure 2b, which has a rating of five, it is observed that, the log posteriors have consistently high values within the boundaries of the canonical phoneme “ae” from the same phoneme model (highlighted in green box). However, after utterance level averaging, the difference between log posteriors of utterances with low and high ratings may diminish. For example, in Figure 2a and b, it is observed that, for both utterances, the utterance level averaged log posteriors look similar. This could be because the differences as observed in the log posteriors, are masked by the log posteriors at other frame locations after the utterance level averaging. In order to overcome this, we derive two sets of features.

3.1.2. Utterance level features

When there is an incorrect phoneme pronunciation, the log posteriors from a phoneme model matched with the canonical phoneme are typically lower than those from the model corresponding to the incorrect phoneme. This is found to be consistent with Figure 2a, where it is observed that, the log posteriors corresponding to the phoneme model “ae” (highlighted in green box) is relatively lower than those corresponding to the phoneme model “ey” (highlighted in red box), within the boundaries of the canonical phoneme “ae”. Therefore, considering the log posteriors from the matched phoneme model separately (indicated by all green boxes in Figure 2a), could be indicative of pronunciation quality. Based on this hypothesis, we construct a one-dimensional vector that consists of the log posteriors from the matched phoneme models of the canonical phonemes in an utterance as shown in Figure 2e and g. When the average of the elements of this vector is appended to the utterance level averaged log posteriors considered by Xiao et al. [5], it could improve the performance of pronunciation assessment task. This resultant feature vector is considered as f_{utt}.

3.1.3. Sub-segment level features

Generally, pronunciation mismatch occurs only within a few sub-segments in an utterance and these sub-segments could be phonemes, syllables or words. For example, in Figure 2a it is observed that pronunciation mismatch has occurred only in the phoneme sub-segment “ae” (highlighted with blue box). As the utterance level averaged log posteriors for rating one and five are similar, sub-segment level average of log posteriors is proposed. Since the average is performed over fewer frames in the sub-segments, we believe that it could indicate pronunciation mismatch between the two ratings better than the utterance level averaged log posteriors. Therefore, similar to f_{utt}, sub-segment level average is performed for each sub-segment, and to it the average of the log posteriors from the matched phoneme models of the canonical phonemes in the sub-segment is appended. The resulting sequence of sub-segmental vectors for an utterance is considered as f_{seg}. Further, the sub-segment level features are modelled in a data driven manner using LSTMs, in contrast to the heuristic utterance level averaging typically used in pronunciation assessment task. Considering this, the dependencies between the sub-segment level features are learnt for assessing pronunciation quality of an utterance.

3.2. Joint model

It is known that the pronunciation quality depends on several factors like syllable stress, intonation, MTI etc. Thus, learning the dependencies from these factors could be helpful in pronunciation assessment. Further, these factors are not completely independent of one another. For example, intelligibility depends on the phoneme mispronunciation including other factors like the type of mispronounced phoneme. In order to explore the interdependencies between the factors as well as the interdependencies between the overall quality and the factors, we propose to use a joint model that consists of shared, factor-specific and overall quality layer. We believe that the shared layer explores these interdependencies by learning common representations in conjunction with factor-specific and over quality layer. The factor-specific and overall quality layer learn representations specific to each factor and overall quality separately.

Shared layer: It consists of six SLNNs or LSTMs depending on f_{utt} or f_{seg}. The number of SLNNs or LSTMs was determined empirically. SLNNs are used when f_{utt} is considered, since f_{utt} is of the same dimension for all utterances. However, when f_{seg} is considered, LSTMs are used since they have been shown to model features with different sequence lengths.

Factor-specific layer: It consists of a SLNN with 32 units for each factor.

The overall quality layer is a SLNN with 32 units. Finally, the representations learnt by the factor-specific and overall quality layers are fed to softmax layers. The number of softmax layers is equal to the number of factors considered plus overall quality. The number of hidden units is equal to the total number of ratings in the factors and overall quality.

4. Experimental results

4.1. Experimental setup

Classification accuracy is used as the objective measure to compare the performance of the proposed approach with the baseline scheme proposed by Xiao et al. [5]. The baseline scheme uses the 78-dimensional paired log posteriors as the feature which is obtained by considering 39 phoneme models, and concatenating the utterance level averaged log posteriors of learner and expert. Further, the baseline scheme also uses a DNN with two hidden layers and 16 units each as the model which we refer to as the baseline model (BM). Following this work, f_{utt} and f_{seg} are also constructed by concatenating the feature from the learner and expert. The dimensions of f_{utt} and f_{seg} are 80 and $n \times 80$ respectively, where n is the number of sub-segments in the utterance. For f_{seg}, the sub-segment is considered as word. Since the number of sub-segments is required to be the same in the learner’s and expert’s pronunciation for concatenation, it is not guaranteed when smaller sub-segments like phonemes and syllables are considered.

In order to know the effectiveness of the proposed joint model that consists of only SLNNs (JDM) or has LSTMs (JLM), we consider independent DNN and LSTM models (IDM and ILM) for each factor as well as overall quality respectively. Each IDM has two hidden SLNN, with each layer having 32 units, which is identical to that considered in the shared and factor-specific layer of JDM. Each ILM is a network having one LSTM and SLNN with 128 and 32 units respectively, which are identical to those considered in the shared and factor-specific layers of JLM.

All seven factors and the overall quality are considered for the joint models. Experiments are carried out in a 10-fold cross-validation setup, where 8 folds are used for train, 1 for validation (val) and 1 for test. The data is divided such that the distribution of the overall pronunciation quality ratings is the same in all 10 folds. Mean and standard deviation (STD) normalization is performed on the features in the train, val and test sets using the mean and STD values computed from the train set. Sigmoid and hyperbolic tangent are used as the activations for SLNNs and LSTMs respectively. The models are compiled using Adam optimizer and categorical cross-entropy as the cost function. They are trained for a maximum of 25 epochs. IDM and JDM are trained with mini-batch size of 32 whereas, ILM and JLM are trained with mini-batch size of 1 since the sequence length varies from one training sample to another. To prevent overfitting, early stopping is invoked if the validation loss does not improve for 3 consecutive epochs.

4.2. Results and discussion

Figure 3 shows the average accuracy on test set across 10 folds obtained for overall quality and all the seven factors from five models (BM, IDM, ILM, JDM and JLM) considering both baseline and proposed features. In the figure, we do not report the accuracies on val set since those are close to that on the test set. From the figure, it is observed that the accuracies obtained with the proposed features are higher than those with the baseline feature across all five models. The highest relative improvements in overall quality with JLM and JDM are found to be 19.13% and 14.93% respectively. This indicates the effectiveness of the proposed features. Further, with the proposed features, we analyse the model specific variations in accuracies for overall quality and the factors.

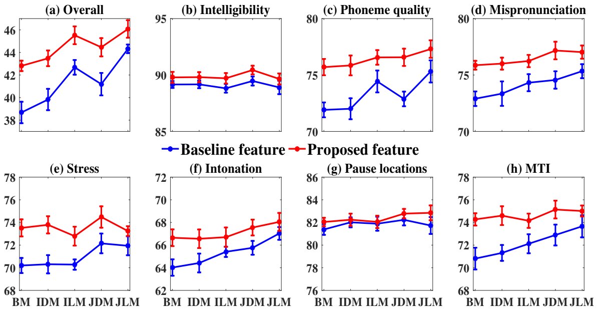

Figure 3: *Average accuracies (standard deviations with error bars) on test set from the five models considering baseline and proposed features.*

From Figure 3a, it is observed that the accuracies obtained from JDM and JLM are found to be 2.25% and 1.23% (relative) higher than that from IDM and ILM respectively. Further, from Figure 3b to 3h, similar observations appear consistently across all the factors. This indicates that the joint models perform better than the independent models. From Figure 3a, it is also observed that ILM and JLM provide an improvement of 4.69% and 3.64% (relative) when compared to IDM and JDM respectively. Further, from Figure 3b to 3h it is found that on an average for all the factors except intelligibility, stress and MTI, LSTM based sequential models provide comparable or better performance than their respective DNN models. This indicates the benefit of f_{seg} with LSTM modelling. The lower performance in those factors could be because, phonemes or syllables might have been the better choice of sub-segment, but it remains a challenge to model them as the number of these sub-segments vary between the learner's and expert's utterance.

In order to know the effect of both representations $\{\phi^s$, $\phi^{fs}\}$ on the overall quality, we obtain accuracies from JDM and JLM considering either only ϕ^s or ϕ^{fs} as the input to the overall quality layer. Table 1 shows the difference between the average accuracies obtained with $\{\phi^s$, $\phi^{fs}\}$ and that obtained with either ϕ^s or ϕ^{fs} separately for JDM and JLM. From the table it is observed that the differences are positive in all cases for overall quality. This indicates the benefit of both the representations $\{\phi^s$, $\phi^{fs}\}$ for assessing overall pronunciation quality. Further, the differences are found to be positive in most of the cases for the factors. This benefit of joint training could be due to the interdependencies between the factors and overall quality.

Table 1: *Difference between the average accuracies obtained with $\{\phi^s$,$\phi^{fs}\}$ and those obtained with either ϕ^s or ϕ^{fs}. The negative entries are indicated in red.*

	JDM		JLM	
	Only ϕ^{fs}	Only ϕ^s	Only ϕ^{fs}	Only ϕ^s
Intelligibility	0.3	0.3	0.09	0.21
Phoneme quality	0.27	0.33	0.03	-0.11
Mispronunciation	0.52	0.44	0.29	0.22
Stress	-0.01	0.32	-0.04	-0.11
Intonation	0.79	1.18	0.13	0.31
Pause locations	0.08	0.17	0.17	-0.01
MTI	-0.1	-0.04	0.19	-0.39
Overall quality	0.81	0.95	0.5	0.78

Table 2: *Confusions among the ratings in overall quality computed from a) BM with baseline feature (BM with baseline), b) JDM with f_{utt} and c) JLM with f_{seg}.*

	(a) BM with baseline					(b) JDM with f_{utt}					(c) JLM with f_{seg}				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	38.0	30.6	16.3	4.3	10.8	54.0	28.7	11.7	2.3	3.3	57.9	25.4	9.2	3.2	4.3
2	24.0	39.8	26.2	3.5	6.5	22.7	46.4	23.6	3.7	3.6	20.4	48.3	21.4	5.2	4.7
3	12.6	22.5	38.2	9.4	17.3	9.8	24.8	39.9	11.4	14.1	9.5	22.5	35.9	16.3	15.8
4	8.3	8.5	29.2	14.7	39.3	4.2	8.1	31.6	20.5	35.6	5.2	8.2	23.2	25.7	37.7
5	4.7	2.4	19.0	11.8	62.1	2.1	2.1	17.7	17.4	60.7	2.9	2.9	11.1	19.7	63.4

Further, in order to know the effectiveness of the proposed approach (JDM with f_{utt} and JLM with f_{seg}) in predicting each of the five overall ratings, the confusions are computed among the ratings considering only overall quality. Table 2 shows the confusions in percentage averaged across 10 folds from BM with baseline feature as well as JDM and JLM with the proposed features. The true and predicted ratings are given along the rows and columns respectively. The red colored entries indicate where JDM and JLM have values lower in the diagonal and higher in the off-diagonal than the respective values from BM with baseline feature. The fewer red colored entries in JDM and JLM indicate that they perform better at correctly predicting most of the ratings instead of biasing to only one of the ratings.

5. Conclusions

We predict the ratings for overall pronunciation quality and its influencing factors by exploring interdependencies among them. For this, we considered sequence containing sub-segment level features from expert's & learner's utterances and jointly modelled their interdependencies using LSTMs, in contrast to heuristically computed utterance level averaged features. Experiments on the data collected from Indian learners reveal that the proposed joint approach performs better than the baseline scheme with utterance level averaged features and without using a joint model. Further investigations are required to identify better sub-segment level features for improving quality of all factors and overall quality. Future works also include better modeling strategies when the length of sub-segment level features from expert and learner are not identical and finding another rater for the data.

6. References

[1] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Computer assisted language learning*, vol. 15, no. 5, pp. 441–467, 2002.

[2] O. D. Deshmukh, S. Joshi, and A. Verma, "Automatic pronunciation evaluation and classification," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[3] C. Hacker, A. Batliner, S. Steidl, E. Nöth, H. Niemann, and T. Cincarek, "Assessment of non-native children's pronunciation: Human marking and automatic scoring," *Proc. SPEECOM*, vol. 1, pp. 123–126, 2005.

[4] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. Narayanan, "Intelligibility classification of pathological speech using fusion of multiple high level descriptors," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

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

[6] P. Dutta and A. Haubold, "Audio-based classification of speaker characteristics," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2009, pp. 422–425.

[7] 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.

[8] M. Suzuki, Y. Qiao, N. Minematsu, and K. Hirose, "Integration of multilayer regression analysis with structure-based pronunciation assessment," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[9] M. Nicolao, A. V. Beeston, and T. Hain, "Automatic assessment of English learner pronunciation using discriminative classifiers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5351–5355.

[10] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *Proc. IS ADEPT*, vol. 6, 2012.

[11] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-natives first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, 2009.

[12] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation verification of children's speech for automatic literacy assessment," in *Ninth International Conference on Spoken Language Processing*, 2006.

[13] M. A. Shahin, J. Epps, and B. Ahmed, "Automatic Classification of Lexical Stress in English and Arabic Languages Using Deep Learning." in *INTERSPEECH*, 2016, pp. 175–179.

[14] K. Li, X. Wu, and H. Meng, "Intonation classification for L2 English speech using multi-distribution deep neural networks," *Computer Speech & Language*, vol. 43, pp. 18–33, 2017.

[15] A. Raux and T. Kawahara, "Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning," in *Seventh International Conference on Spoken Language Processing*, 2002, pp. 737–740.

[16] M. J. Munro and T. M. Derwing, "The foundations of accent and intelligibility in pronunciation research," *Language Teaching*, vol. 44, no. 3, pp. 316–327, 2011.

[17] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication*, vol. 69, pp. 31–45, 2015.

[18] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010.

[19] J. Geertzen, T. Alexopoulou, B. Post, and A. Korhonen, "Native language effects on pronunciation accuracy in L2 English."

[20] P. Trofimovich and W. Baker, "Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech," *Studies in second language acquisition*, vol. 28, no. 1, pp. 1–30, 2006.

[21] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech communication*, vol. 30, no. 2-3, pp. 83–93, 2000.

[22] F. de Wet, C. Van der Walt, and T. Niesler, "Automatic assessment of oral language proficiency and listening comprehension," *Speech Communication*, vol. 51, no. 10, pp. 864–874, 2009.