Assessment of Non-Native Speech Intelligibility using Wav2vec2-based Mispronunciation Detection and Multi-level Goodness of Pronunciation Transformer

Ram C.M.C Shekar[1], *Mu Yang*[1], *Kevin Hirschi*[2], *Stephen Looney*[3], *Okim Kang*[2], *John Hansen*[1]

[1]Center for Robust Speech Systems (CRSS), University of Texas at Dallas, TX, USA
[2]Northern Arizona University, AZ, USA, [3]Pennsylvania State University, PA, USA

[1]{ramcharan.chandrashekar, mu.yang, john.hansen}@utdallas.edu,
[2]{kevin.hirschi, okim.kang}@nau.edu, [3]sdl16@psu.edu

Abstract

Automatic pronunciation assessment (APA) plays an important role in providing feedback for self-directed language learners in computer-assisted pronunciation training (CAPT). Several mispronunciation detection and diagnosis (MDD) systems have achieved promising performance based on end-to-end phoneme recognition. However, assessing the intelligibility of second language (L2) remains a challenging problem. One issue is the lack of large-scale labeled speech data from non-native speakers. Additionally, relying only on one aspect (e.g., accuracy) at a phonetic level may not provide a sufficient assessment of pronunciation quality and L2 intelligibility. It is possible to leverage segmental/phonetic-level features such as goodness of pronunciation (GOP), however, feature granularity may cause a discrepancy in prosodic-level (suprasegmental) pronunciation assessment. In this study, Wav2vec 2.0-based MDD and Goodness Of Pronunciation feature-based Transformer are employed to characterize L2 intelligibility. Here, an L2 speech dataset, with human-annotated prosodic (suprasegmental) labels, is used for multi-granular and multi-aspect pronunciation assessment and identification of factors important for intelligibility in L2 English speech. The study provides a transformative comparative assessment of automated pronunciation scores versus the relationship between suprasegmental features and listener perceptions, which taken collectively can help support the development of instantaneous assessment tools and solutions for L2 learners.

Index Terms: Wav2vec 2.0, Transformer, goodness of pronunciation, phoneme, prosody, suprasegmental.

1. Introduction

Recent advancements in acoustic modeling and automatic speech recognition (ASR) techniques have allowed for the development of CAPT tools. CAPT is aimed at self-directed language learning and automatic mispronunciation detection [1, 2, 3]. This facilitates non-native (L2) speakers to learn foreign-spoken (L1) languages. Most advancement efforts focus on scoring phoneme-level pronunciation quality [4, 5, 6, 7, 8, 9, 10]. The major focus is on providing diagnosis on phonetic-level errors (phoneme substitution, deletion, insertion) [11, 12, 13, 14, 15]. Recently, the importance of assessing prosodic-level features (e.g. lexical stress, intonation) has increased substantially [16]. L2 speech is influenced by suprasegmental and temporal differences from their first language, lexical stress and speech rate, hypothesized to account for large variations in L2 speech are believed to play a key role in influencing L2 speech intelligibility [17], which is different from L1 speech intelligibility, typically assessed in the presence of noise [18].

International Teaching Assistants (ITAs) at North American universities have often faced communication difficulties due to differences in their speech [19, 20]. Two important prosodic components, speech rate [21, 22, 23, 24] and pause units [24, 25, 26] are hypothesized to be related to intelligibility and perceived accentedness [21]. However, [24] suggests that the relationship between speech rate and perception of L2 speech is curvilinear, implying an optimal rate may neither be too high or too low and it may lie in between. Pause patterns, different from L1 speakers, negatively impact perceptions of fluency[25] and ITA effectiveness [26]. Lexical stress deviations also impact several dimensions of listener judgments. L2 speakers' speech rate is typically considered a strong predictor of perceived fluency [22, 23]. However, results from [25] indicate that perception of L2 speech and speech rate is curvilinear in that L2 speech that is spoken too quickly or too slowly inhibits comprehension. Nevertheless, the complex relationship between speech rate, lexical stress, other important prosodic factors, and intelligibility still remains an ongoing research question. In our study, an ITA dataset is prepared by collecting data from international students who serve as teaching assistants (TAs) in North American universities. The dataset is subjectively rated for intelligibility and accentedness. Most existing approaches in L2 speech analysis focus on assessing pronunciation. Pronunciation quality can be modeled at multiple levels: phonetic, word and utterance, and multiple factors such as prosody, stress, lexical stress, etc., These factors are typically modeled separately [27, 28, 29, 30, 31, 32]. However, many multi-level scores on phonetic, word, and utterance-level features can be correlated. Recent advancements in machine learning have allowed us to learn a more comprehensive representation. In our task, a transformer-based model that is trained on multiple aspects of pronunciation simultaneously is leveraged to study the relationship between L2 intelligibility and multi-level aspects of pronunciation. Furthermore, in this study, we also consider both human transcripts and transcripts generated by ASR, in order to (i) simulate conditions that exist in several instantaneous assessment tools, and (ii) assess the robustness of our L2 speech intelligibility assessment framework. Additionally, the phonetic-level MDD solution is also leveraged to study the relationship between L2 speech intelligibility and the lowest-level phonetic error characteristics.

2. Related work

2.1. GOPT Overview

Conventional methods like GOP [4, 5, 7, 8, 10] have been extensively studied. Recent advancements in transformer-based models have enabled advanced GOP-based approaches that can effectively capture the relationship between phonemes and

10.21437/Interspeech.2023-2371

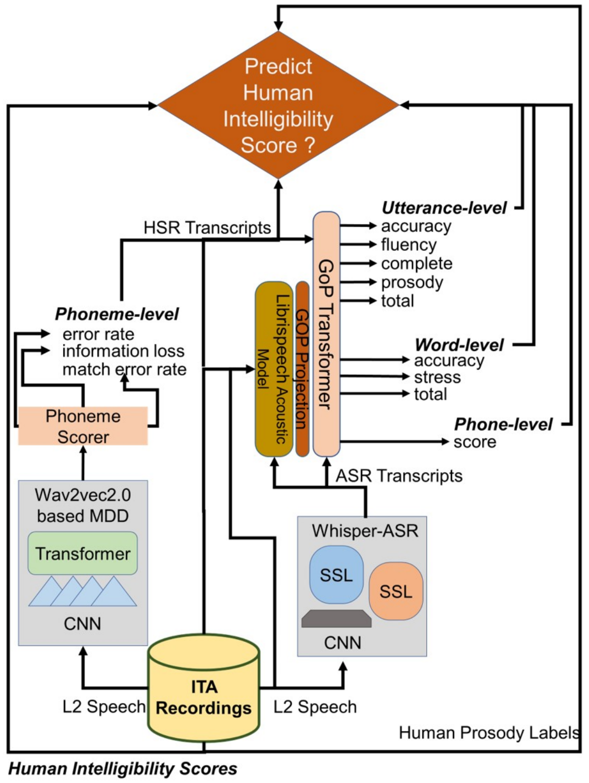

Figure 1: *Assessment of L2 Speech Intelligibility using GOPT and Wav2Vec2.0 models*

words within an utterance. Development of Self-attention-based models [32, 33] resulted in GOPT (Goodness of Pronunciation Transformer), which is a recently developed pronunciation assessment model that is based on Goodness of Pronunciation (GOP) features and a Transformer self-attention architecture [34]. Multi-level labels (one phoneme-level, three word-level, and five utterance-level accuracy, prosody, and fluency) are used to achieve multi-aspect and multi-grained supervision for GOPT training. GOPT is a state-of-the-art model that jointly predicts multi-aspect scores of pronunciation assessment with different granularities. This contrasts with several conventional approaches which also independently incorporate different aspects of pronunciation markers. GOPT learns the correlation between utterance-level tokens and phoneme-level tokens through the attention mechanism

2.2. Wav2vec2.0 Overview

Wav2vec 2.0 [35] is a pre-trained method that uses a feature encoder, a context network, and a quantization module, to learn a contrastive learning objective. Wav2vec 2.0 has achieved state-of-the-art results on phone recognition for CAPT. Wav2vec 2.0's latent representation also renders rich phonetic information [36].

3. Experimental Setup

In this work, we make use of the international teaching assistance (ITA) dataset and leverage state-of-the-art GOP-based Transformer and Wav2vec 2.0 for characterizing utterance, word and phoneme level pronunciation. The proposed framework for the assessment of L2 intelligibility using human/machine annotations and multi-level pronunciation properties is shown in Fig. 1.

Table 1: *Statistics of ITA Dataset and Human/Machine Annotations for Intelligibility Assessment.*

Intelligibility	Annotation/Scoring		# Speakers	# Utterances
	Human	Machine		
L2	HSR	-	15	887
	-	ASR	15	5012
	HSR+ Prosody	-	57	3000
	Prosody	ASR	57	4839
L2-High	HSR	-	19	438

3.1. ITA Dataset

For this study, 54 adult learners, with diverse L1 backgrounds and English-speaking abilities, produced a total of 76 conversational in-class speech recordings, with an average length of 6.06 minutes. The ITA recordings can be broadly categorized based on the level of non-native speech proficiency and intelligibility: (i) 'L2' (Low intelligibility) and (ii) 'L2-High' (High intelligibility). Along with the speech utterances, some parts of ITA recordings are supplemented by both human and machine annotations. Human annotations include transcripts referred to as human speech recognition (HSR) and prosodic annotations, whereas at the machine level ASR transcripts are generated. 'L2-High' has HSR transcripts available whereas some portions of the 'L2' speech has both HSR transcripts and prosody annotations. Only a few utterances have only HSR transcripts. Whisper-ASR [37] is used to obtain ASR transcription for utterances that have missing HSR transcripts. 'L2' level has several combinations of supplementary human and/or machine annotations for every portion of the ITA recordings. The number of speakers and utterances for every portion of the ITA recordings as described in Table 1.

3.1.1. Human labeled Transcripts and Prosody Ratings

Fifteen trained linguists rated the speech based on accentedness and intelligibility. Speech rate, measured in syllables per second, silent pause and filled pause durations were recorded and omitted from syllable count. Additionally, lexical stress scores were calibrated using polysyllabic words extracted from speech event transcripts scored against the CMU Pronouncing Dictionary [38]. A lexical stress score indicates the duration, intensity, or pitch emphasis of the stressed syllable. Furthermore, the accentedness and intelligibility of the speakers were also rated by the linguistic listener pool.

3.1.2. Whisper ASR Transcripts Generation

Although the ITA dataset is nearly labeled and carefully compiled, transcribers have generally reported difficulty in reliably deciphering the content. Therefore, additionally Whisper ASR network [37] is considered to provide auxiliary transcription, and assess the robustness of intelligibility scoring. Whisper [37] performs end-to-end ASR on 30-second audio chunks. The Whisper architecture is based on an encoder-decoder Transformer.

3.2. Acoustic Model and GOP Feature Extraction

Goodness-of-pronunciation (GOP), an early DNN-based method proposed to characterize MDD, focuses on evaluating phonetic errors. In our study, a Kaldi-based ASR acoustic model, which is based on the factorized time-delay neural network (TDNN-F), is trained with Librispeech [39]. GOP features for the ITA dataset are extracted using Kaldi Librispeech S5 recipe. GOP extraction involves processing the audio sample along with its corresponding canonical transcription within the acoustic module to obtain a sequence of frame-level phonetic posterior probabilities, these are then force-aligned at the phoneme level and converted to 84-dimensional goodness of pronunciation (GOP) features.

3.3. GOPT Inference and Scoring

The GOPT is trained using the 84-dim GOP feature as input. GOPT is trained on Speechocean762 [40], a free open-source dataset designed for pronunciation assessment. Speechocean762 is multi-labeled and provides rich label information at every level. During GOPT training, the mean squared error (MSE) as the loss is computed at the utterance, word and phoneme levels and averaged.

3.4. MDD using Wav2vec 2.0

Along with the GOPT, the MDD evaluation is considered an auxiliary machine annotation/score, used for the characterization of the reliability of the pronounced phonemes and aspects of L2 speech. More recently, MDD has been achieved via end-to-end phoneme recognition [41]. The implemented model applies the momentum pseudo labeling technique to Wav2vec 2.0 fine-tuning, to leverage the unlabeled L2 speech for improved phoneme recognition performance [41]. MDD information is computed using the phoneme scorer (JiWER) that interprets the phoneme predictions from Wav2Vec2.0 [42].

4. Results

All the utterances are transcripted using either HSR or ASR techniques. All the utterances from the ITA-Recordings, including both ’L2’ and ’L2-High’ portions, are assessed using the multi-granular and multi-level GOPT. Now, some portions may have additional prosodic annotations, and MDD information. In summary, for every utterance of the ITA dataset, GOPT produces five utterance-level scores (accuracy, fluency, completeness, prosody, and total score), three world-level scores (accuracy, stress, and total score) and a phoneme level score. Additionally, annotations/scores are available based on the portions in ITA-Recording. The prosodic annotations are: (i) articulation rate, (ii) lexical stress score, and (iii) silent pause duration and MDD information is: (i) PER: phoneme error rate (ii) MER: match error rate and (iii) IL: information loss. The intelligibility scores are characterized by (i) intelligibility at phrase level, (ii) intelligibility as a total number of words understood, and (iii) average intelligibility for the entire duration.

4.1. Multicollinearity Test and Regression Analysis

The ITA dataset was tested for multicollinearity and it showed that there was no significant correlation between the predictors. For all variations of the ITA dataset, a simple linear regression, random forest regressor, and XGBoost regressor were applied and the results are as shown in Table 2. Random forest regressor was found to be the best-performing model based on mean square error and R2 metrics as highlighted in Table 2.

Table 2: *Comparative analysis of regression types on different combinations of annotations/scoring on ITA dataset*

ITA	Annotation/Scoring		Regression	MSE	R2
	Human	Machine			
L2	HSR	-	Linear	0.0761	0.0213
			Random Forest	5.24E-09	1.0000
			XGBoost	6.46E-08	1.0000
	-	ASR	Linear	0.0547	0.0352
			Random Forest	1.88E-07	1.0000
			XGBoost	6.78E-05	0.9987
	HSR+ Prosody	-	Linear	0.0472	0.1384
			Random Forest	8.28E-10	1.0000
			XGBoost	1.02E-06	1.0000
	Prosody	ASR	Linear	0.03137	0.4680
			Random Forest	9.30E-29	1.0000
			XGBoost	2.77E-09	1.0000
	-	ASR + MDD	Linear	0.046829	0.1708
			Random Forest	1.15E-08	1.0000
			XGBoost	2.18E-07	1.0000
	Prosody	ASR + MDD	Linear	0.0249	0.5772
			Random Forest	9.22E-29	1.0000
			XGBoost	5.50E-10	1.0000
L2-High	HSR	-	Linear	0.0761	0.0210
			Random Forest	3.18E-10	1.0000
			XGBoost	6.07E-08	0.9999

4.2. Characterization of L2 Intelligibility using Feature Importance

Random forest regressor was found to be the best-performing model and they have been used for assessing feature importance. The feature importance results for all combinations of annotations/scoring for the ITA dataset are described in Fig. 2. In Fig. 2 A), a comparative analysis of L2 intelligibility predictors for ‘L2-High:HSR’, ‘L2:HSR’, and ‘L2:ASR’ configurations reveal that only utterance level scores of GOPT are significant and other word level and phonetic predictors are not dominant predictors. Fluency is the top predictor for ‘L2-High:HSR’, prosody is the top predictor for ‘L2:HSR’ and interestingly, accuracy is the top predictor for ‘L2:ASR’ configuration. In Fig. 2 B), four configurations for L2 speech, namely: ‘HSR + Prosody’, ‘ASR + Prosody’, ‘ASR + MDD’, ‘ASR + Prosody + MDD’ are considered for comparative analysis of non-native speech intelligibility predictors. As earlier, among GOPT scores, only utterance level scores are significant and comparing ‘HSR + Prosody’ based model with ‘ASR + Prosody’ suggests that human-labeled prosodic predictors are significant for L2 intelligibility. Among GOPT utterance scores, only prosody and total scores contribute towards predicting L2 intelligibility for ‘HSR + Prosody’ configuration and Whisper-ASR generated transcripts seems to degrade the reliability of GOPT-generated utterance-level scores. Also, the important prosodic features such as ‘articulation rate’, ‘lexical stress’ and ‘silent pause duration’ remain the most important features which influence human labeled intelligibility scores. Among ‘ASR + MDD’ and ‘ASR + MDD + Prosody’ configurations, human prosodic information seems to play a very dominant role in predicting L2 intelligibility. In the absence of any human labels and relying on Whisper-ASR generated transcripts and the MDD information, namely: information loss, phoneme error rate and match error rates were significantly influential for predicting L2 intelligibility characteristics. Additionally, all utterance-level GOPT predictors also contributed towards predicting L2 intelligibility. However, we observe that between GOPT and MDD scores, the phoneme-level MDD scores played a dominant role in influencing L2 intelligibility.

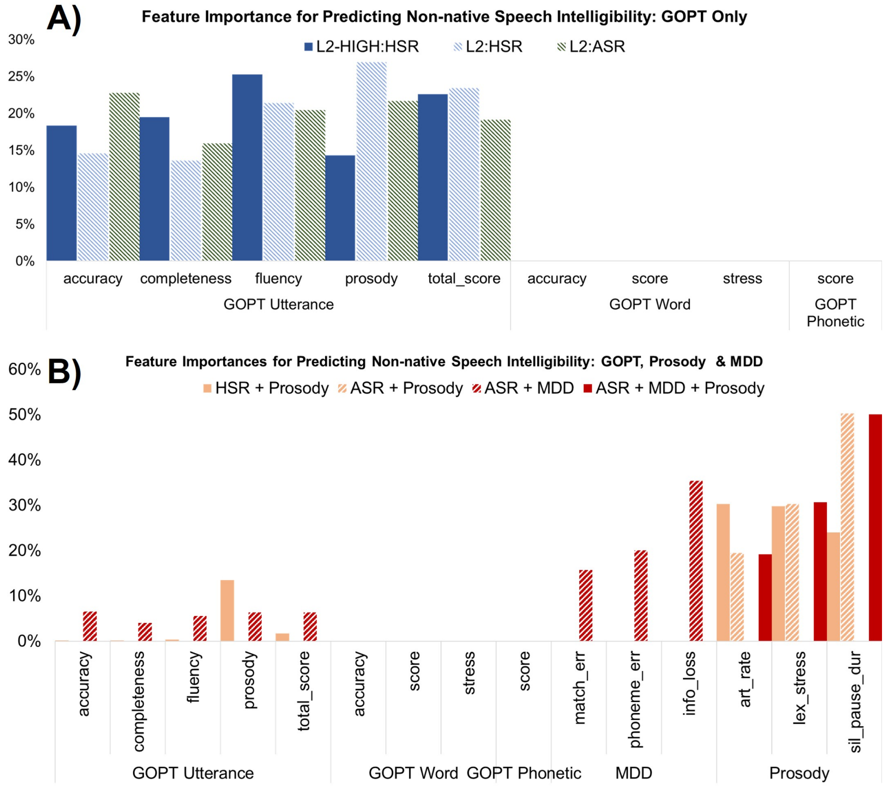

Figure 2: *Evaluating most significant predictors of L2 speech intelligibility using feature importance analysis of random forest regressor: A) GOPT features importance analysis among L2:ASR, L2:HSR and L2-High:HSR; B) GOPT, prosody and MDD features importance analysis among L2:HSR + Prosody, L2:ASR + Prosody, L2:ASR + MDD and L2:ASR + MDD + Prosody.*

5. Discussion

Previously it has been found that lexical stress and speech rate independently relate to L2 speech intelligibility. In our assessment, prosodic features were found to outweigh the multi-level granular features of pronunciation significantly, especially when human-labeled prosodic features were available. Among all scores, only the utterance level scores, specifically prosody, completeness, total score, and others were dominant in the absence of any prosodic features. Furthermore, a comparative analysis of phoneme-level MDD diagnostic features was found to be more influential than GOPT scores, despite the availability of multi-level scoring. Analysis of features showed interesting aspects of L2 intelligibility: (i) Human or ASR transcription does not significantly influence the factors that impact L2 speech intelligibility; (ii) Apart from transcription, human-rated prosodic factors are dominant in influencing L2 intelligibility; (iii) MDD is still a major factor for assessing L2 intelligibility; (iv) Among automatic/machine learning based predictors, lower phonetic level features are more important compared to multi-level GOPT features, however, human transcribed prosodic features are more relevant compared to lower level phonetic transcription.

6. Conclusion

This study has considered a framework to assess L2 speech intelligibility by considering several aspects of pronunciation: (i) lower phonetic level pronunciation-based MDD solution; (ii) Multi-level granular pronunciation assessment tool; (iii) Transcription robustness: Human vs ASR; (iv) Human rated prosodic labels. Most existing approaches use MDD solutions for characterizing L2 pronunciation. However, our study shows that an automatic assessment of L2 speech intelligibility can be carried out reliably, irrespective of human or automatic speech transcription. Human-rated prosodic predictors are the dominant factors in the assessment of L2 speech intelligibility. Also, there is no significant difference in the dominant predictors for L2 speech intelligibility when HSR transcripts are compared to ASR generate versions. However, MDD predictors do become highly dominant in characterizing L2 speech intelligibility. Further, MDD features are still dominant in the presence of multi-level GOPT features. The findings from this analysis provide direction for the development of robust systems for characterizing L2 speech intelligibility. Furthermore, with the increasing demand for instantaneous L2 speech assessment tools, the results of this experimental framework underscore the need to develop more accurate tools to estimate human-rated prosodic features and MDD solutions.

7. Acknowledgements

This study is supported by NSF EAGER CISE Project 2140415, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

8. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, 2009.

[2] K. Zechner, D. Higgins, X. Xi *et al.*, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, 2009.

[3] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *International Symposium on automatic detection on errors in pronunciation training*, 2012.

[4] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, 2000.

[5] F. Zhang, C. Huang, F. K. Soong *et al.*, "Automatic mispronunciation detection for mandarin," in *Proc. ICASSP*, 2008.

[6] D. Luo, Y. Qiao, N. Minematsu *et al.*, "Analysis and utilization of mllr speaker adaptation technique for learners' pronunciation evaluation," in *Proc. Interspeech*, 2009.

[7] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. ICASSP*, 2012.

[8] W. Hu, Y. Qian, F. K. Soong *et al.*, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, 2015.

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

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

[11] S. Sudhakara, M. K. Ramanathi, C. Yarra *et al.*, "An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities." in *Proc. Interspeech*, 2019.

[12] W. Li, S. M. Siniscalchi, N. F. Chen *et al.*, "Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling," in *Proc. ICASSP*, 2016.

[13] M. Wu, K. Li, W.-K. Leung *et al.*, "Transformer based end-to-end mispronunciation detection and diagnosis." in *Proc. Interspeech*, 2021.

[14] L. Peng, K. Fu, B. Lin *et al.*, "A study on fine-tuning wav2vec2. 0 model for the task of mispronunciation detection and diagnosis." in *Proc. Interspeech*, 2021.

[15] Y. Feng, G. Fu, Q. Chen *et al.*, "Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis," in *Proc. ICASSP*, 2020.

[16] D. Korzekwa, R. Barra-Chicote, S. Zaporowski *et al.*, "Detection of Lexical Stress Errors in Non-Native (L2) English with Data Augmentation and Attention," in *Proc. Interspeech*, 2021.

[17] O. Kang, "Linguistic analysis of speaking features distinguishing general english exams at cefr levels," *Research notes*, 2013.

[18] N. Mamun, M. S. Zilany, J. H. L. Hansen *et al.*, "An intrusive method for estimating speech intelligibility from noisy and distorted signals," *The Journal of the Acoustical Society of America*, 2021.

[19] M. Inglis, "The communicator style measure applied to nonnative speaking teaching assistants," *International Journal of Intercultural Relations*, 1993.

[20] J. M. Levis, "Changing contexts and shifting paradigms in pronunciation teaching," *TESOL quarterly*, 2005.

[21] O. Kang, R. I. Thomson, and M. Moran, "Empirical approaches to measuring the intelligibility of different varieties of english in predicting listener comprehension," *Language Learning*, 2018.

[22] 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*, 2006.

[23] M. J. Munro and T. M. Derwing, "The effects of speaking rate on the comprehensibility of native and foreign-accented speech," *Canadian Acoustics*, 1996.

[24] P. Tavakoli, "Pausing patterns: Differences between l2 learners and native speakers," *ELT journal*, 2011.

[25] J. Kahng, "The effect of pause location on perceived fluency," *Applied Psycholinguistics*, 2018.

[26] A. E. Tyler, A. A. Jefferies, and C. E. Davies, "The effect of discourse structuring devices on listener perceptions of coherence in non-native university teacher's spoken discourse," *World Englishes*, 1988.

[27] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, 2000.

[28] P. C. Bagshaw, "Automatic prosodic analysis for computer aided pronunciation teaching," Ph.D. dissertation, University of Edinburgh PhD thesis, 1994.

[29] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proc. ICASSP*, 2005.

[30] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech communication*, 2010.

[31] K. Li, X. Wu, and H. Meng, "Intonation classification for l2 english speech using multi-distribution deep neural networks," *Computer Speech & Language*, 2017.

[32] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *Proc. NeurIPS*, 2017.

[33] B. Lin, L. Wang, X. Feng *et al.*, "Automatic scoring at multi-granularity for l2 pronunciation." in *Proc. Interspeech*, 2020.

[34] Y. Gong, Z. Chen, I.-H. Chu *et al.*, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *Proc. ICASSP*, 2022.

[35] A. Baevski, Y. Zhou, A. Mohamed *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, 2020.

[36] T.-A. Hsieh, C. Yu, S.-W. Fu *et al.*, "Improving Perceptual Quality by Phone-Fortified Perceptual Loss Using Wasserstein Distance for Speech Enhancement," in *Proc. Interspeech*, 2021.

[37] A. Radford, J. W. Kim, T. Xu *et al.*, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[38] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[39] V. Panayotov, G. Chen, D. Povey *et al.*, in *Proc. ICASSP*, 2015.

[40] J. Zhang, Z. Zhang, Y. Wang *et al.*, "speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment," in *Proc. Interspeech*, 2021.

[41] M. Yang, K. Hirschi, S. D. Looney *et al.*, "Improving Mispronunciation Detection with Wav2vec2-based Momentum Pseudo-Labeling for Accentedness and Intelligibility Assessment," in *Proc. Interspeech*, 2022.

[42] A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Proc. Interspeech*, 2004.