10th Workshop on Speech and Language Technology in Education (SLaTE)
22-24 August 2025, Nijmegen, Netherlands

Articulatory-Enhanced Mispronunciation Detection and Diagnosis Models: A Multi-dimensional Error Analysis

Xing Wei[1], *Catia Cucchiarini*[2], *Roeland van Hout*[2], *Helmer Strik*[1,2,3]

[1]Centre for Language and Speech Technology (CLST), Radboud University, The Netherlands
[2]Centre for Language Studies (CLS), Radboud University, The Netherlands
[3]Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands
{xing.wei, catia.cucchiarini, roeland.vanhout, helmer.strik}@ru.nl

Abstract

This study presents a multi-dimensional error analysis of integrating articulatory features (AFs) into End-to-End (E2E) models for mispronunciation detection and diagnosis (MDD). We examine two output representation frameworks: phoneme-based (PHN) and articulatory-based (ART), employing both customized Conformer-based models and fine-tuned Wav2Vec 2.0 models. Experimental results reveal that AF-integrated ART models demonstrate enhanced capability in identifying common mispronunciations, thereby lowering diagnosis error rates. In contrast, PHN models maintain a slight advantage in overall detection accuracy but provide less articulatory insight. Additional analysis reveals key challenges, such as degraded detection accuracy on medium-length utterances and inter-speaker variability. These findings emphasize critical trade-offs between detection and diagnosis, while proposing practical considerations for building more accurate and learner-aware L2 pronunciation assessment systems.

Index Terms: mispronunciation detection and diagnosis, error analysis, articulatory features

1. Introduction

Mispronunciation detection and diagnosis (MDD) plays a pivotal role in computer-assisted language learning (CALL), providing second language (L2) learners not only accurate detection outcomes but also precise feedback to improve their pronunciation proficiency [1-3]. Conventional MDD systems generally rely on complex, multi-stage pipelines, such as forced alignment and pre-defined acoustic models [4-5]. Recent advances in deep learning have shifted the focus toward end-to-end (E2E) models, which directly learn mappings from acoustic inputs to pronunciation labels, simplifying system framework and improving performance [6-8].

A growing line of research investigates how to further enhance these models, particularly in challenging L2 learning contexts marked by high phonetic variability and learner-specific deviations [9]. One promising avenue is the integration of articulatory features (AFs), which represent speech based on the physical configurations of the vocal tract (e.g., voicing, place, and manner of articulation) [10-12]. Unlike raw acoustic features, AFs are more interpretable and linguistically grounded, offering robustness against speaker variability, a particularly valuable attribute in L2 scenarios characterized by divergent first language (L1) backgrounds and varying proficiency levels [13-14]. Moreover, AFs have the potential to support fine-grained feedback, such as subsegmental articulatory guidance (e.g., lip rounding or tongue position), which is essential for effective pronunciation training [15-16].

Prior studies have demonstrated the general benefits of AFs in enhancing MDD accuracy, yet a systematic understanding of their impact across different E2E frameworks remains limited. In particular, the comparative effectiveness of phoneme-based (PHN) models versus articulatory-based (ART) models, which operate at distinct representational levels, has yet to be thoroughly examined. Of greater concern, despite notable gains in overall accuracy, critical factors such as utterance length [17-18] or inter-speaker variability [13] remain underexplored, even though they may significantly impact model behavior. These nuanced patterns are often hidden by coarse-grained evaluation methods, underscoring the need for more detailed, fine-grained analysis.

This study addresses these critical gaps by conducting a comprehensive and multi-dimensional error analysis of E2E MDD models enhanced with AFs. Departing from prior works that predominantly target accuracy improvements or proposing novel architectures, we place our emphasis on understanding model behavior across a broad design space, including different modeling paradigms (customized Conformer-based E2E model vs. fine-tuned Wav2Vec 2.0), feature configurations (with vs. without AFs), and output representations (PHN vs. ART). The key novelty of this study lies in its systematic and comparative error analysis spanning multiple architecture choices, feature sets, and representation levels, an area still underexplored in existing MDD research. To this end, we carry out a structured investigation into how factors such as utterance length, speaker variability, and mispronunciation types impact both detection performance and diagnostic precision. Through these analyses, we seek to uncover latent patterns and model limitations that are routinely overlooked, thereby offering practical insights for designing more robust and learner-aware MDD systems. Based on this objective, we pose the following overarching research question: Can in-depth error analysis of AF-enhanced E2E MDD models yield meaningful insights into their performance and behavior patterns? To explore this question, we investigate two sub-questions:

RQ1): Which performance bottlenecks and behavioral trends emerge from in-depth error analysis of AF-enhanced E2E MDD models?

RQ2): How do different output representations (PHN vs. ART) influence the relationship between detection accuracy and diagnostic precision?

10.21437/SLaTE.2025-7

2. Experimental setup for error analysis

2.1. Speech material

Two speech corpora were employed in this study: Librispeech [19] and L2-ARCTIC [20]. Librispeech involves approximately 1,000 hours of read English speech, originally sourced from audiobooks publicly released via the LibriVox project [21]. A 100-hour subset of clean speech was selected to train the AF classifiers, enabling robust acoustic-to-articulatory mapping. The L2-ARCTIC corpus contains read-aloud English speech produced by L2 speakers with diverse language backgrounds. The 3.5 hours of manually annotated data out of 27 hours of speech were selected to train the MDD models. The selected L2 data was randomly partitioned into training (12 speakers), validation (6 speakers), and test sets (6 speakers), with speaker and utterance independence across all splits. The six speakers in the test set include three native Vietnamese speakers (PNV, THV, and TLV), two native Hindi speakers (RRBI and SVBI), and one native Arabic speaker (SKA). Two are male (RRBI and TLV), and the other four are female.

2.2. Models

Following the concept outlined in [22-23], three AF categories were defined for vowels: *Backness* (e.g., *Front, Central, Back, Back2front*), *Height* (e.g., *High, Middle, Low, Low2high*), and *Roundness* (e.g., *Rounded, Unrounded, Rounded2unrounded*), and three for consonants: *Manner* (e.g., *Affricate, Fricative, Nasal, Stop, Approximant*), *Place* (e.g., *Alveolar, Bilabial, Dental, Glottal, Labiodental, Palatal, Post-Alveolar, Velar*), and *Voicing* (e.g., *Voiced/Unvoiced*). An independent DNN-HMM classifier was trained on 39-dimensional MFCC features extracted from the selected subset of Librispeech for each of the six categories. Each classifier consisted of six hidden layers, each consisting of 2048 sigmoid units, and produced frame-level posteriors through a softmax layer. These posteriors represent probability distributions over the possible values within each respective AF category (e.g., the *Voicing* classifier outputs posterior probabilities for *Voiced* and *Unvoiced*). Finally, the posteriors from all six classifiers were concatenated to construct a composite AF vector for each frame.

This study investigates two distinct E2E modeling methods for MDD. The first method constructs three custom E2E models that share a unified architecture, comprising a Conformer-based encoder, a Transformer-based decoder, and a Connectionist Temporal Classification (CTC) module. The key difference among these models lies in the input features: raw speech (RS) and 83-dimensional FBank pitch (FP) features serve as baselines, while the proposed M1 model incorporates a frame-by-frame fusion of FP and AFs. To be specific, given two input feature matrices of shape $T \times D_1$ and $T \times D_2$ (where T is the number of frames of the features, and D_1, D_2 are their respective feature dimensions), fusion yield a matrix of shape $T \times (D_1 + D_2)$. The second method involves fine-tuning XLSR [24], a multilingual extension of the pre-trained Wav2Vec 2.0 model[1]. In this setup, the baseline model (FT) utilizes only speech embeddings from the encoder, whereas the proposed M2 model integrates AFs with these embeddings before feeding them into a Transformer decoder and CTC for joint inference. These two methods yield five core model configurations: RS, FP, M1, FT, and M2. While the primary objective is segment-level prediction (i.e., phoneme transcription), this study additionally investigated subsegmental modeling by mapping phonemes to their corresponding articulatory labels, as defined by the AF categories. Accordingly, each model configuration is evaluated under two output representation frameworks: the phoneme-based (PHN) framework, which directly predicts phonemes, and the articulatory-based (ART) framework, which leverages articulatory labels for subsegmental analysis. In total, ten models are evaluated, covering all five configurations across both PHN and ART frameworks.

2.3. Evaluation metrics

To perform a multi-dimensional error analysis, a set of metrics targeting both detection and diagnostic precision is considered. For correctly pronounced speech, predictions were categorized as either Correct Acceptance (CA) or False Rejection (FR). For mispronounced speech, outcomes were classified as Correct Rejection (CR) or False Acceptance (*FA*). *CR* cases were further divided based on diagnostic precision: Correct Diagnosis (CD) refers to accurately identifying the specific mispronunciation, while Diagnosis Error (DE) denotes a misidentified error. Based on these classifications, the following metrics were computed: Detection Accuracy (DA), False Acceptance Rate (FAR), False Rejection Rate (FRR), Diagnosis Error Rate (DER), and Matthews Correlation Coefficient (MCC), as defined below:

$$DA = (CA + CR)/(CA + CR + FA + FR) \quad (1)$$

$$FAR = FA/(CR + FA) \quad (2)$$

$$FRR = FR/(CA + FR) \quad (3)$$

$$DER = DE/(CD + DE) \quad (4)$$

$$MCC = \frac{CA * CR - FA * FR}{\sqrt{(CA + FR) * (CA + FA) * (CR + FA) * (CR + FR)}} \quad (5)$$

3. Results

3.1. Overall model performance evaluation

Table 1 demonstrates the mean and standard deviation (SD) of all models concerning key evaluation metrics, calculated by averaging results across the six speakers. AF-enhanced models (M1 and M2) generally outperform their respective feature-based baselines (FP and FT) across PHN and ART frameworks, supporting the effectiveness of AFs in enhancing DA and MCC.

Table 1: *Mean (SD) performance across speakers for all evaluated models.*

Model	DA	MCC	FAR	FRR
PHN-RS	81.75(1.33)	0.33(0.08)	57.65(6.60)	9.76(2.71)
PHN-FP	81.03(1.36)	0.32(0.08)	57.67(6.02)	10.62(3.15)
PHN-M1	82.52(1.41)	0.38(0.09)	52.03(5.42)	10.14(2.84)
PHN-FT	86.29(1.22)	0.47(0.10)	47.31(8.89)	7.14(1.25)
PHN-M2	86.90(1.24)	0.48(0.11)	48.23(9.42)	6.30(0.81)
ART-RS	82.41(1.59)	0.35(0.10)	57.53(6.10)	9.16(2.87)
ART-FP	81.83(1.23)	0.34(0.09)	55.93(6.27)	10.16(3.06)
ART-M1	82.48(1.39)	0.38(0.10)	50.98(6.84)	10.49(3.07)
ART-FT	85.99(1.87)	0.44(0.10)	54.74(8.48)	5.76(1.15)

[1] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

ART-M2	86.76(2.30)	0.46(0.09)	55.85(7.37)	4.60(0.77)

Table 2 presents the paired t-test results, including *p*-values and Cohen's *d* (effect size), computed across six speakers per model. The results suggest that the improvements from AF integration were often statistically significant and practically meaningful for various metrics in both frameworks.

Table 2: *Statistical significance (p) and effect size (d) for DA and MCC across key model comparisons.*

Comparison	DA	MCC
	p/d	*p/d*
PHN-M1 vs. PHN-FP	0.0045/1.9985	0.0026/2.2637
PHN-M2 vs. PHN-FT	0.0183/1.4062	0.1412/0.7128
ART-M1 vs. ART-FP	0.1158/0.7759	0.0059/1.8708
ART-M2 vs. ART-FT	0.0188/1.3975	0.0180/1.4142

3.2. Speaker-specific analysis

Figure 1 illustrates the DA heatmaps for six speakers (the six rows) across all models. Across both frameworks (PHN and ART), all speakers generally achieve higher DA scores in the proposed models M1 and M2 (compare columns 3 to 2, 5 to 4, 8 to 7, and 10 to 9). Despite these general trends, however, noticeable inter-speaker variability in DA remains evident. For instance, although ART-M1 generally outperforms ART-FP, speaker *RRBI* shows a marginally higher DA in ART-FP (82.74%) compared to ART-M1 (82.64%).

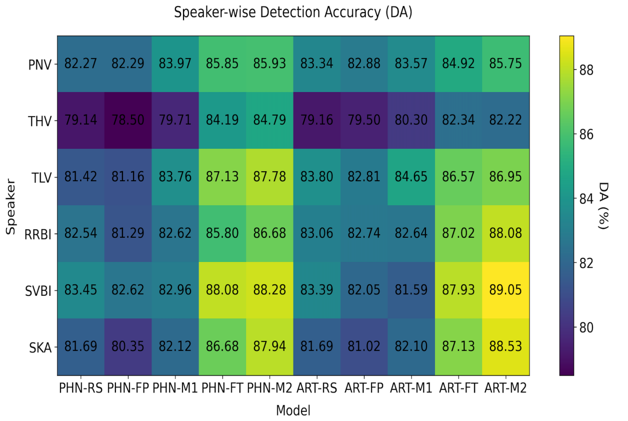

Figure 1: *Speaker-wise detection accuracy (%) across all evaluated models. (Darker indicates lower DA).*

The speaker-wise DA heatmap, along with the relatively higher mispronunciation rates of speakers *THV* (27.00%) and *TLV* (24.73%), as reported in Table 3, demonstrates a notable discrepancy between speakers. Specifically, *THV* attained the lowest DA scores across all models, whereas *TLV* achieved substantially higher values. To further examine this unexpected divergence, we first computed FAR, FRR, and MCC separately for each speaker within each model, and then averaged the results across the ten models to obtain speaker-wise metrics. The results show that *TLV* recorded the lowest average FAR at 41.81%, followed by *THV* with the second lowest average FAR of 52.56%. Both speakers also demonstrated low average FRR (*THV*: 6.61%, *TLV*: 6.72%) and high average MCC scores, ranking first (0.56) and second (0.47) for *TLV* and *THV*, respectively.

Table 3: *Distribution (number/percentage) of mispronunciation by speakers.*

Speaker	Sub.	Ins.	Del.	Total
PNV	622/12.32	16/0.32	246/4.87	884/17.51
THV	903/17.72	67/1.31	406/7.97	1376/27.00
TLV	776/15.28	43/0.85	437/8.60	1256/24.73
RRBI	451/9.03	47/0.94	65/1.30	563/11.27
SVBI	426/8.61	28/0.57	89/1.80	543/10.98
SKA	464/9.14	116/2.29	59/1.16	639/12.59

3.3. Effect of utterance length

Figure 2 illustrates the relationship between utterance-wise DA and utterance length, measured by the number of phonemes or articulatory labels, across the ten models. A noteworthy overall trend is the superiority of the proposed models (M1 and M2) over their respective baselines (FP and FT). For example, PHN-M1 (dark purple curve) surpasses PHN-FP (dark yellow curve), highlighting the benefits of integrating AFs. Furthermore, a pattern consistent with the results described in Table 1 emerges when comparing models across the PHN and ART frameworks. Specifically, RS and FP achieve better results under the ART framework, whereas M1, FT, and M2 yield marginally better performance within the PHN framework.

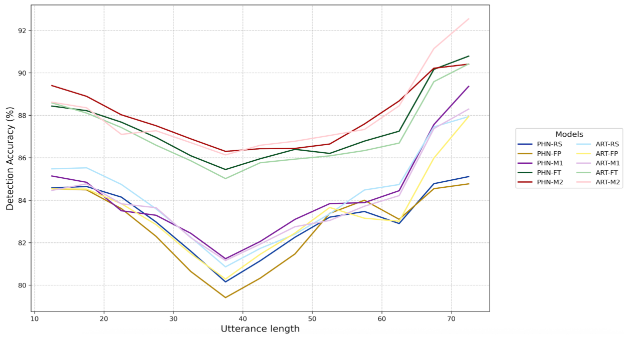

Figure 2: *Utterance-wise DA (%) vs. utterance length. Color of models: RS (blue), FP (yellow), M1 (purple), FT (green), and M2 (red). PHN (dark colors) and ART (light colors).*

As illustrated in Figure 2, model performance noticeably declines when processing utterances shorter than 40 labels, particularly in the 20-40 range. To better understand the factors contributing to this pattern, we first partitioned the test set into three utterance length categories: short (<21 labels), medium (21-40 labels), and long (>40 labels). For each length category, FAR and FRR were computed separately for each of the ten models, resulting in three FAR and FRR values per model. These values were then averaged across the ten models within each length category to identify general trends. The test set is primarily composed of medium-length utterances, with fewer short and long cases. Our analysis reveals that the average FRR increases from 6.13% for short utterances to 8.55% for medium-length utterances, the highest among all categories, before slightly decreasing to 8.23% for long utterances. Similarly, the average FAR increases from 47.07% (short) to 53.32%

(medium), with long utterances reaching 56.20%. These concurrent increases in FAR and FRR account for the decline in DA from short to medium-length utterances, suggesting that models encounter greater challenges in handling intermediate-length utterances.

3.4. Diagnostic precision on frequent mispronunciations

A critical distinction between the PHN and ART frameworks, particularly when integrated with AFs, was observed in their diagnostic capabilities, as measured by DER. Figure 3 presents a heatmap illustrating the DER values associated with the ten most frequent mispronunciations made by the speakers in the test set, selected by ranking error types based on their frequency. The corresponding distributions of these ten mispronunciations are summarized in Table 4, where the DH/D substitution error emerges as the most frequent in the test set, followed by Z/S, indicating their prevalence among the L2 speakers examined. Notably, these two error types also exhibit higher frequency in the entire dataset, which may partially explain their relatively lower DER values in Figure 3.

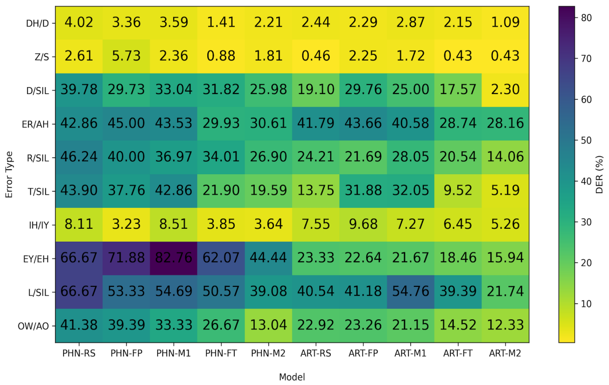

Figure 3: *DER for the most frequent mispronunciations in the test set. (Darker indicates higher DER).*

In Figure 3, the DH/D pair on the first row indicates that the phoneme /DH/ was mispronounced as /D/ by the speaker, representing a substitution error. As observed, the integration of AFs generally leads to reduced DER across both frameworks, especially for M2 vs. FT. More strikingly, ART models integrating AFs (ART-M1 and ART-M2) consistently obtained substantially lower DER values for most of these most frequent mispronunciations compared to the corresponding PHN models (PHN-M1 and PHN-M2) that integrated AFs. These findings highlight the superior capacity of the ART framework to leverage AFs for fine-grained, precise diagnosis of frequent mispronunciation types. However, despite these improvements, certain error types like *L/SIL* remain challenging, as evidenced by persistently high DER values even in the best-performing ART models, as detailed in Figure 3.

Table 4: *Number of the most frequent mispronunciations in the test set and their occurrences in the entire datasets in this study.*

Error type	Test set	Entire dataset
DH/D	348	1676
Z/S	296	1701
D/SIL	238	706
ER/AH	222	400
R/SIL	207	534
T/SIL	193	565
IH/IY	134	855
EY/EH	131	243
L/SIL	126	298
OW/AO	125	389

4. Discussion and conclusions

Through a systematic, multi-dimensional error analysis of AF-enhanced E2E MDD models, several key insights were gained that directly address our research questions. This in-depth examination of model behavior helped identify the core factors affecting both detection accuracy and diagnostic precision.

While the integration of AFs generally improved model performance across PHN and ART frameworks, our analysis for RQ1 also uncovered certain limitations and performance bottlenecks in current models. As shown in Figure 2, utterance-wise DA declines for utterances shorter than 40 labels. Note that medium-length utterances make up approximately 65% of the test set. As demonstrated in Section 3.3, this length group also exhibits the highest FAR and FRR, indicating that models struggle the most in this region. Interestingly, the proportion of mispronunciations in the short, medium, and long utterances is relatively similar (16.76%, 16.95%, and 18.31%, respectively), suggesting that the performance drop is not simply due to error frequency. We hypothesize that medium-length utterances fall into a contextual ‘gray zone’: they are too long for simple modeling yet lack the redundancy for longer utterances needed for disambiguation. This limitation may be attributed to the insufficient respective fields of current architectures, such as Conformer and XLSR. Similar observations have been reported in [17], underscoring the need for improved context modeling, especially for mid-length speech segments.

Another salient finding related to RQ1 was the considerable inter-speaker variability noted, although the generalizability of this finding is constrained by the six-speaker sample. The contrast between *THV* and *TLV* serves as an example: *THV* showed the lowest DA despite its second-best average FAR and MCC, as detailed in Section 3.2, likely a consequence of its substantially higher error rate (27.00%), illustrated in Table 3, amplified false acceptances. Conversely, *TLV*'s similar error rate resulted in higher DA, potentially due to more detectable errors (lowest average FAR: 41.81%). This implies that for L2 learners with high mispronunciation rates, detectability may be impacted by both frequency and distinctiveness. This remains a persistent challenge, even with AFs, and calls for validation on larger and more diverse L2 cohorts with varied L1 backgrounds.

Regarding the impact of output representations on the trade-off between DA and diagnostic precision (RQ2), the choice of model framework emerged as a key differentiator. While AFs generally enhanced performance over baselines, the nature of these gains differed across frameworks. ART models exhibited higher diagnostic precision, reflected in lower DER for common errors, whereas PHN models achieved comparable or slightly better DA. This suggests potential performance biases: ART outputs may better support error type identification, whereas PHN outputs tend to preserve phonetic detail, enhancing DA. These trends were further shaped by input configurations. RS and FP models performed better with ART outputs, likely due to their simplified classification categories. In contrast, M1 and M2 yield higher accuracy in the PHN setting, where the fine-grained AF inputs align more closely with phoneme outputs, maximizing the advantage of AF information. This benefit may diminish under ART, where coarser output granularity creates a mismatch with detailed inputs. Phoneme distribution may also contribute to these differences, with consonants accounting for 60.8% of L2-ARCTIC [20], their disproportionate influence may skew overall DA. Finally, PHN-FT outperformed ART-FT, possibly due to greater compatibility with the phoneme-centric pretraining of XLSR, facilitating more efficient fine-tuning.

In summary, our multi-dimensional error analysis identifies three principal findings for AF-enhanced MDD models: 1) The ART framework provides better diagnostic precision for most frequent errors compared to PHN models, despite marginally lower DA and MCC scores; 2) Considerable inter-speaker variability in DA is observed within the limited speakers, even with comparable proportions of mispronunciations, suggests differences in error detectability, potentially linked to acoustic-phonetic distinctiveness, may better explain model performance variations than error quantity alone; 3) While current AF-enhanced models show consistent improvements, they still struggle with medium-length utterances. These insights suggest that AF integration yields distinct but context-dependent L2 assessment advantages, highlighting intricate links among representation granularity, feature compatibility, and training strategies, motivating further studies on larger, diverse datasets.

This study has limitations, most notably the small speaker sample ($N = 6$), reducing statistical power and generalizability. While our multi-metric analysis revealed meaningful individual performance differences, a thorough investigation of acoustic-phonetic factors was beyond the scope of this work and needs to be addressed in future exploration. Future studies should: 1) validate the findings on larger and more diverse L2 datasets; 2) improve contextual modeling to better address utterance length; and 3) examine mispronunciation patterns at the acoustic-phonetic level.

5. References

[1] C. Cucchiarini and H. Strik, “Automatic speech recognition for second language pronunciation training,” *The Routledge handbook of contemporary English pronunciation*, pp. 556-569, 2017.

[2] A. Neri, C. Cucchiarini, and H. Strik, “The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch,” *ReCALL*, vol. 20, no. 2, pp. 225–243, Mar. 2008.

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

[4] W. Hu, Y. Qian, and F.K. Soong, “An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners’ speech,” *Proc. Speech and Language Technology in Education (SLaTE 2015)*, pp. 71-76, 2015.

[5] F. Nazir, M. N. Majeed, M. A. Ghazanfar, and M. Maqsood, "Mispronunciation Detection Using Deep Convolutional Neural Network Features and Transfer Learning-Based Model for Arabic Phonemes," in *IEEE Access*, vol. 7, pp. 52589-52608, 2019.

[6] B.-C. Yan, M.-C. Wu, H.-T. Hung, and B. Chen, “An End-to-End Mispronunciation Detection System for L2 English Speech Leveraging Novel Anti-Phone Modeling,” *Proc. Interspeech 2020*, pp. 3032-3036, 2020.

[7] M. Wu, K. Li, W.-K. Leung, and H. Meng, “Transformer Based End-to-End Mispronunciation Detection and Diagnosis,” *Proc. Interspeech 2021*, pp. 3954-3958, 2021.

[8] H. -W. Wang, B. -C. Yan, H. -S. Chiu, Y. -C. Hsu, and B. Chen, "Exploring Non-Autoregressive End-to-End Neural Modeling for English Mispronunciation Detection and Diagnosis," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 6817-6821,2022.

[9] M. Tu, A. Grabek, J. Liss, and V. Berisha, “Investigating the role of L1 in automatic pronunciation evaluation of L2 speech,” *Proc. Interspeech 2018*, pp. 1636-1640, 2018.

[10] M. Shahin, J. Epps, and B. Ahmed, “Phonological-Level Mispronunciation Detection and Diagnosis,” *Proc. Interspeech 2024,* pp. 307-311, 2024.

[11] Q. Chen, B. Lin, and Y. Xie, “An Alignment Method Leveraging Articulatory Features for Mispronunciation Detection and Diagnosis in L2 English,” *Proc. Interspeech 2022*, pp. 4342–4346, Sep. 2022.

[12] S. Mao, Z. Wu, X. Li, R. Li, X. Wu, and H. Meng, "Integrating Articulatory Features into Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech," *2018 IEEE International Conference on Multimedia and Expo (ICME)*, San Diego, CA, USA, pp. 1-6, 2018.

[13] Y. E. Kheir, S. A. Chowdhury, and A. Ali, "L1-Aware Multilingual Mispronunciation Detection Framework," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, pp. 12752-12756, 2024.

[14] S. Khanal, M. T. Johnson and N. Bozorg, "Articulatory Comparison of L1 and L2 Speech for Mispronunciation Diagnosis," *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, pp. 693-697, 2021.

[15] B. -C. Yan, H. -W. Wang, Y. -C. Wang, and B. Chen, "Effective Graph-Based Modeling of Articulation Traits for Mispronunciation Detection and Diagnosis," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, pp. 1-5, 2023.

[16] W. Li, S. M. Siniscalchi, N. F. Chen and C. -H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, pp. 6135-6139, 2016.

[17] R. Cumbal, B. Moell, J. Lopes, and O. Engwall, “You don’t understand me!: comparing ASR results for L1 and L2 speakers of Swedish,” *Proc. Interspeech 2021*, pp. 4463-4467, 2021.

[18] Y.Y. Lin, T. Han, H. Xu, V.T. Pham, Y. Khassanov, T.Y. Chong, L. Lu, and Z. Ma, “Random utterance concatenation based data augmentation for improving short-video speech recognition,” *Proc. Interspeech 2023*, pp. 904-908, 2023.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015.

[20] G. Zhao et al., “L2-ARCTIC: A Non-native English Speech Corpus,” *www.isca-archive.org*, 2018.

[21] J. Kearns, “Librivox: Free public domain audiobooks*,” Reference Reviews*, vol. 28, no. 1, pp. 7–8, 2014.

[22] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, Ö. Çetin, “Articulatory feature classifiers trained on 2000 hours of telephone speech,” *Proc. Interspeech 2007*, pp. 2485-2488, 2007.

[23] M. Morshed, M. Hasegawa-Johnson, “Cross-lingual articulatory feature information transfer for speech recognition using recurrent progressive neural networks,” *Proc. Interspeech 2022*, pp. 2298-2302, 2022.

[24] A. Babu et al., “XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale,” *arXiv.org*, Dec. 16, 2021.