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

Contrastive Learning-based Syllable-Level Mispronunciation Detection and Diagnosis for Speech Audiometry

Longbin Jin[1,2], *Donghun Min*[1], *Jung Eun Shin*[1], *Eun Yi Kim*[2]

[1]AI Engine Development Team, Voinosis Inc., South Korea
[2]Department of Computer Science and Engineering, Konkuk University, South Korea
lbjin@voinosis.com, dhmin@voinosis.com, eardoc70@voinosis.com, eykim@konkuk.ac.kr

Abstract

Speech audiometry assesses hearing disorders, typically relies on audiologists, making the process subjective and requiring in-person evaluation. In this paper, we introduce *SylPh*, a novel automatic syllable-level mispronunciation detection and diagnosis (MDD) model that generalizes across open-set syllables while also offering phonemic analysis. To capture a wide range of mispronunciation patterns, we construct positive and pseudo-negative bags to extract in-distribution and out-of-distribution features from input audio. Our model aligns audio features with adaptive text embeddings using a contrastive objective, dynamically adjusting decision boundaries for each syllable within a single model. Extensive experiments on a large-scale dataset demonstrate its effectiveness in both closed-set and open-set syllables. Notably, despite training only on syllable-level labels, the Sylph has the capability to localize phoneme-level abnormalities, providing detailed diagnostic insights.
Index Terms: syllable-level mispronunciation detection, speech audiometry, hearing disorder, contrastive learning

1. Introduction

Speech audiometry is a clinical procedure used to assess an individual's ability to understand and distinguish spoken words at various volumes and frequencies [1, 2]. During the test, subjects listen to syllable-level speech stimuli and repeat them, while audiologists calculate the percentage of correct responses to assess auditory capabilities. This manual process often introduces variability across audiologists, making the procedure both time-consuming and inconsistent. Such challenges highlight the need for an objective and efficient approach to assess speech audiometry automatically.

Recently, mispronunciation detection and diagnosis (MDD) systems have been developed to automatically identify pronunciation errors and provide corrective feedback [3, 4]. The most straightforward approach is automatic speech recognition (ASR)-based methods [5, 6, 7], which transcribe speech into text and evaluate according to the correct annotation. ASR models, such as Whisper [8], are trained on sentence-level data that require context information, making them less effective at recognizing syllables in isolation. Such a two-stage process may propagate errors from the ASR stage. Alternatively, end-to-end (E2E) learning models based on connectionist temporal classification (CTC) [9, 10, 11] are trained for direct mispronunciation detection from input audio. Typically, normal speech data are easier to collect and annotate, while listing all possible mispronunciation patterns is challenging, resulting in data imbalance. Consequently, E2E models tend to become biased toward the normal class, leading to output overconfident predictions and poor generalization to the unseen data at test time.

Driven by this analysis, we developed SylPh, a contrastive learning-based MDD model for speech audiometry, focusing on monosyllable and bisyllable stimuli. We treat mispronunciations as anomalies and identify them by detecting deviations from normal speech patterns. Given a large amount of normal data, the phoneme-level normal representation for each stimulated syllable can be easily created using a clustering method, referred to as the positive bag. In contrast, due to the data imbalance, abnormal patterns are sparse and limited [12]. To comprehensively investigate various mispronunciation errors, we construct a pseudo-negative bag containing normal representations of all non-stimulated syllables. Consequently, audio features are derived by comparing their similarity to the normal representations in both the positive bag and pseudo-negative bag, considering both in-distribution (ID) and out-of-distribution (OOD) variations. Then, a projection head maps the audio features to a joint embedding space with the text encoder to optimize contrastive loss. As mispronunciation patterns vary across syllables, we use adaptive text prompts based on the stimulated syllables, enabling the creation of distinct classification boundaries for each stimulus syllable.

Through extensive experiments on a large-scale dataset, our SylPh model outperforms baselines and achieves state-of-the-art performance in syllable-level MDD for both closed-set and open-set syllables. Unlike traditional methods that provide only syllable-level outputs, our approach also offers phoneme-level analysis, despite using only syllable-level labels during training.

2. Speech Audiometry Dataset

We utilize a large-scale hearing disorder voice dataset provided by AI-Hub[1]. The dataset was collected from 12,616 native Korean speakers. For the stimuli, the dataset uses words from the Korean Standard Monosyllabic and Bisyllabic Word Lists for Adults (KS-MWL-A [13] and KS-BWL-A [14]). The KS-MWL-A consists of four word lists, each containing 50 monosyllables, while the KS-BWL-A consists of three word lists, each containing 12 bisyllables. Each participant randomly selects one word list from KS-MWL-A and one from KS-BWL-A. Participants listened to words one at a time, repeated what they heard, and their responses were recorded in audio [15]. Each audio had a length of three seconds. Audiologists then created syllable-level labels, annotating pronunciations as "normal" or "abnormal" based on deviations from the speech stimulus.

Table 1 shows the distribution of normal and abnormal speech data across all word lists. It is observed that normal data is significantly more than abnormal data. Accordingly, we focus on generating the normal patterns for each syllable and then predict anomalies that deviate from the normal distributions.

[1]https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71393

10.21437/Interspeech.2025-1285

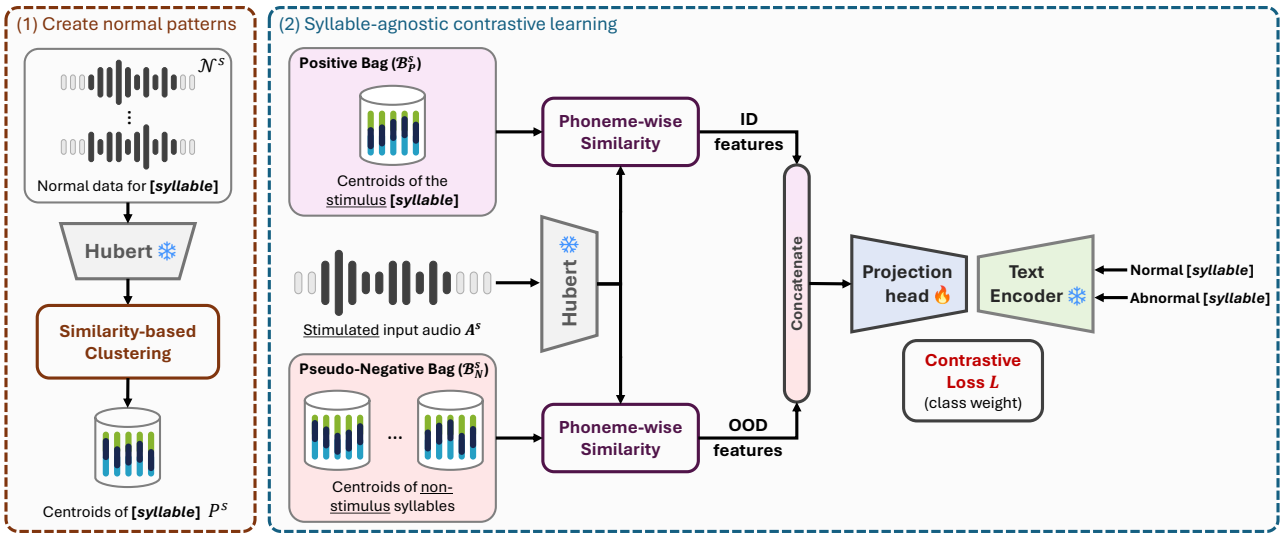


Figure 1: *Framework of the proposed SylPh model: (1) Create phoneme-level normal patterns for each syllable using similarity-based clustering (Section 3.1). (2) Extract audio features by comparing input Hubert features with positive and pseudo-negative bags. The projection head is trained with text supervision using contrastive loss to develop adaptive classifiers for each syllable. (Section 3.2).*

Table 1: *Participants and data distributions of the AI-Hub hearing disorder voice dataset for each word list.*

Dataset	List	# of person		# of data	
		Train	Test	Normal	Abnormal
KS-MWL-A (monosyllable)	M1	3,308	354	154,165	28,935
	M2	2,805	331	133,415	23,385
	M3	2,636	320	124,578	23,222
	M4	2,573	289	118,608	24,492
	Sum	**11,322**	**1,294**	**530,766**	**100,034**
KS-BWL-A (bisyllable)	B1	4,134	451	50,733	4,287
	B2	3,721	426	45,621	4,143
	B3	3,467	417	41,327	5,281
	Sum	**11,322**	**1,294**	**137,681**	**13,711**

3. Methodology

In our study, we aim to detect mispronunciations in syllabic words using contrastive learning. Since mispronunciations often occur at the phoneme level [16, 17], we first extract subject-agnostic phoneme embeddings using the HuBERT model [18], which is pre-trained on a large-scale audio dataset. Then, we identify normal audio patterns for each syllable by clustering phoneme features from normal speech and generating centroids for each syllable. For detecting the diverse patterns of mispronunciation, we define a positive bag including centroids of normal patterns for stimulus syllables, and a pseudo-negative bag containing centroids of all non-stimulus syllables. Acoustic features are extracted by computing phoneme-wise similarity with both bags, capturing ID and OOD characteristics. Finally, a projection head transforms the audio features into the joint subspace of text embeddings. The text embeddings are extracted from a pre-trained encoder, with the input varying based on the stimulus syllable, enabling a dynamic classifier for each syllable within a single model. Figure 1 illustrates the framework of the proposed SylPh model.

3.1. Normal Patterns Generation

Anomalous patterns are difficult to define since they vary largely from person to person. In contrast, it is straightforward to find the normal pattern within the corresponding syllables. Let $\mathcal{N}^s = \{N_i^s\}_{i=1}^m$ be a normal set that contains m correctly pronounced audio recordings stimulated by syllable s in the training set. For each normal set, the syllable-level audio clips are encoded into phoneme-level embedding to represent the smallest units of pronunciation within each recording. Here, we adopt the HuBERT model pre-trained on 4000 hours of Korean spontaneous speech[2]. The features generated from the last layer of the HuBERT model $f_h(\cdot)$, represent the audio segment for each non-overlapping 20ms frame, providing a detailed representation of the phoneme features. The phoneme features of each normal set are formulated as $f_h(N_i^s) = \{n_i^{s,1}, \ldots, n_i^{s,150}\}$, corresponding to a three-second recording.

Since the phoneme-level features of the normal set $\{f_h(N_i^s)\}_{i=1}^m$ contain a large number of normal patterns, potentially featuring duplicates, we apply the k-means clustering method to extract representative features that describe the normal phonemic distributions. We adopt the cosine similarity function as the distance metric of the k-means clustering algorithm. As a results, the centroids of each cluster $P^s = \{p_1^s, ..., p_k^s\}$ are identified as the normal representations of each phoneme.

3.2. Syllable-Agnostic Contrastive Learning

Unlike other methods that can only detect pre-defined anomalies, we aim to detect various mispronunciation-related anomalies by leveraging the centroids of normal speech patterns. To achieve this, we define a positive bag and a pseudo-negative bag. Given the stimulus syllable S, the positive bag $\mathcal{B}_P^s = \{p_i^s \mid \forall i, s = S\}$ contains the normal patterns of the stimulated syllable, while the pseudo-negative bag $\mathcal{B}_N^S = \{p_i^s \mid \forall i, \forall s \neq S\}$

[2]https://huggingface.co/team-lucid/hubert-base-korean

Table 2: *Closed-set performance on monosyllabic and bisyllabic words. Default settings are marked in gray.*

Model	Train Lists	Test Lists	Head	Precision	Recall	Macro F1	AUROC	Accuracy
Whisper [8]	KS-MWL-A	KS-MWL-A	-	0.1583	0.9421	0.2711	0.5227	0.1967
GOPT [19]			linear	0.1508	0.0448	0.0690	0.5322	0.8086
SylPh (monosyllable)	KS-MWL-A	KS-MWL-A	linear	0.7152	**0.7919**	0.7515	0.7819	0.8822
			projection	**0.7236**	0.7876	**0.7542**	**0.7899**	**0.8880**
SylPh (bisyllable)	KS-BWL-A	KS-BWL-A	linear	0.6783	**0.8589**	0.7580	0.8566	0.9126
			projection	**0.6877**	0.8532	**0.7616**	**0.8568**	**0.9255**
SylPh (both)	KS-MWL-A & KS-BWL-A	KS-MWL-A	linear	0.7091	0.7948	0.7495	0.7815	0.8723
			projection	**0.7168**	**0.7988**	**0.7556**	**0.7866**	**0.8821**
		KS-BWL-A	linear	0.7572	0.7629	0.7600	0.7613	0.9438
			projection	**0.7662**	**0.7661**	**0.7661**	**0.7639**	**0.9547**
		Both	linear	0.7137	0.7993	0.7541	0.7866	0.8873
			projection	**0.7209**	**0.8025**	**0.7595**	**0.7878**	**0.8962**

contains all the normal patterns of the non-stimulus syllable.

For an input audio A^s stimulated by syllable S, we first extract phoneme-level embeddings using the pre-trained HuBERT model, as described earlier. The extracted HuBERT features, $f_h(A^s) = \{a^{s,1}, \ldots, a^{s,150} \mid s = S\}$, may contain abnormal phonemic representations. Thus, to detect deviations from normal patterns in the positive bag, the ID features are defined by the distance between each phonemic feature and its nearest centroid neighbor from the positive bag, computed using inverse cosine similarity:

$$f^{id} = \{\min\{1 - sim(a^{s,j}, p_k^s) | \forall p_k^s \in \mathcal{B}_P^S\} | \forall j\} \quad (1)$$

Conversely, to detect various mispronunciation errors, the OOD features are defined by the distance between each phonemic feature and the pseudo-negative patterns, representing the difference from the ID feature:

$$f^{ood} = \{\min\{1 - sim(a^{s,j}, p_k^s) | \forall p_k^s \in \mathcal{B}_N^S\} | \forall j\} \quad (2)$$

Finally, the ID and OOD features are concatenated to form the discriminative audio features.

Since mispronunciation patterns differ across stimulated syllables, using varying decision boundaries can enhance performance. Traditional one-class detectors require training multiple models for each syllable, which limits their application to unseen syllables. To adaptively capture these variations within a single model, we align the audio features with text representations, enabling the development of dynamic classifiers for different stimuli. We use the simple text prompt template of "normal {*syllable*}" and "abnormal {*syllable*}" to differentiate between correct and mispronounced syllables. We adopt the pre-trained text encoder from the CLIP model [20] to extract 512-dimensional text features f^t, which are then aligned with the audio features for mispronunciation detection.

The audio features are mapped to the joint subspace of the text features using a projection head. The projection head consists of two linear layers with an activation function. Thus, the projected audio features can be formulated as follows:

$$f^a = w_2(\text{ReLU}(w_1(\{f^{id}, f^{ood}\}) + b_1)) + b_2 \quad (3)$$

To align the audio and text features, we minimize the contrastive loss, which encourages the audio and text representations to be closer for matching pairs (f_i^a, f_i^t) and farther apart for non-matching pairs in a minibatch. Additionally, to address class imbalance, we incorporate class weights, which are inversely proportional to the number of data in each class. As a result, the class-weighted contrastive loss function is defined as:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} w_{f_i^t} \frac{\exp(sim(f_i^a, f_i^t)/\tau)}{\sum_{c=1}^{C}\exp(sim(f_i^a, f_c^t)/\tau)} \quad (4)$$

4. Experiments

4.1. Performance of MDD on Closed-Set Syllables

We start by comparing the performance of the proposed SylPh with baseline models on closed-set syllables. In this setting, models are trained and tested on the same syllable lists. As shown in Table 2, the ASR-based model Whisper [8] struggles to generate correct syllable-level outputs, resulting in low accuracy. On the other hand, the E2E model GOPT [19] tends to be biased toward the normal class, leading to a low F1 score. Furthermore, both models exhibit low AUROC, indicating poor ability to distinguish between correct speech and mispronunciations. In contrast, the proposed SylPh demonstrates superior performance on monosyllabic and bisyllabic word lists. Moreover, by training the model with both datasets, we achieve higher F1 scores on the KS-MWL-A and KS-BWL-A, respectively. This improvement is attributed to the larger pseudo-negative bag, which enables the detection of a broader range of anomalies and the extraction of more precise OOD features.

To better understand the contribution of the projection head, we also compare our model with an ablation that uses a linear head, which has a similar number of parameters for fair comparison. The projection head maps the audio features to a larger space of text embeddings, while the linear head maps the audio features to the dimensionality of the number of classes. This eliminates the need for a text encoder and contrastive loss. As a result, the absence of a dynamic classification boundary for each syllable leads to performance degradation across all cases.

4.2. Performance of MDD on Open-Set Syllables

To demonstrate the performance on open-set syllables, we conducted a leave-one-list-out evaluation. We ensured that the training and test lists did not overlap in terms of syllables. Importantly, for testing on open-set syllables, SylPh requires correctly pronounced data to construct a positive bag for each syllable, which can be easily collected from the stimulus audio.

Table 3: *Open-set performance with leave-one-list-out evaluation. Closed-set settings are marked in gray.*

Model	Train List	Test List	Precision	Recall	Macro F1	AUROC	Accuracy
SylPh (monosyllable)	M1,M2,M3	M4	0.7595	0.7675	0.7635	0.7564	0.8924
	M1,M2,M4	M3	0.7008	0.7818	0.7391	0.7822	0.8868
	M1,M3,M4	M2	0.7367	0.7932	0.7639	0.7851	0.8940
	M2,M3,M4	M1	0.6819	0.8081	0.7397	0.7955	0.8521
	M1,M2,M3,M4	M1,M2,M3,M4	0.7236	0.7876	0.7542	0.7899	0.8880
SylPh (bisyllable)	B1,B2	B3	0.7337	0.7945	0.7629	0.7876	0.9333
	B1,B3	B2	0.6647	0.8258	0.7365	0.8218	0.9157
	B2,B3	B1	0.6698	0.8549	0.7511	0.8517	0.9283
	B1,B2,B3	B1,B2,B3	0.6877	0.8532	0.7616	0.8568	0.9255

Table 4: *Contribution of ID and OOD features.*

Train/Test List	ID	OOD	Macro F1	Accuracy
KS-MWL-A	✓	-	0.6788	0.8431
	-	✓	0.7283	0.8605
	✓	✓	**0.7542**	**0.8880**
KS-BWL-A	✓	-	0.6910	0.9098
	-	✓	0.6943	0.9154
	✓	✓	**0.7616**	**0.9255**
Both	✓	-	0.6842	0.8662
	-	✓	0.7275	0.8740
	✓	✓	**0.7595**	**0.8962**

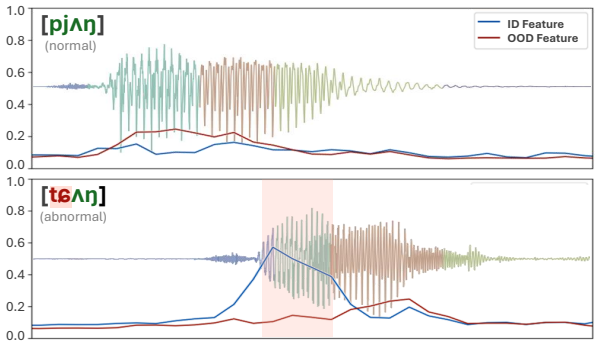


Figure 2: *Visualization of ID and OOD features using audio stimulated by the monosyllable 병[pjʌŋ].*

Meanwhile, the pseudo-negative bag is derived from the training set. As shown in Table 3, SylPh exhibits comparable performance on open-set syllables to that in the closed-set setting, even in a more challenging scenario. This is made possible by the use of adaptive text prompts, which allow the model to dynamically adjust its understanding of various stimuli, enabling generalization to open-set syllables.

4.3. Ablation Study of ID and OOD Features

We further analyze how ID and OOD features contribute to the model's performance. Table 4 shows the various combinations of ID and OOD features in the SylPh. When using each feature independently, OOD features consistently outperform ID features in all cases. Since the OOD features are generated based on the pseudo-negative bag, they capture a broader range of mispronunciation patterns, allowing the model to produce more accurate acoustic representations compared to relying solely on ID features. Due to the larger number of monosyllabic data, the performance gap between ID and OOD features is even more significant in the KS-MWL-A dataset. Notably, by using only the OOD features, it eliminates the need to build a positive bag, thereby enabling zero-shot testing for open-set syllables.

Nevertheless, incorporating ID features allows SylPh to capture stimulus information from multiple modalities, including acoustic and textual cues. Thus, the combination of both feature sets leads to a substantial improvement in the model's robustness, enabling it to better handle variations in pronunciation and adapt to diverse types of speech anomalies. This result underscores the impact of incorporating both ID and OOD features for comprehensive mispronunciation detection.

4.4. Phonemic Analysis

Although the proposed SylPh is trained with syllable-level annotations, it can still perform phoneme-level analysis. Figure 2 visualizes the normal and abnormal audio with ID and OOD features over the time sequence. Both recordings are stimulated by the monosyllable [pjʌŋ] in the International Phonetic Alphabet (IPA) notation. The audio signal is colored to indicate the nearest center in the positive bag. The normal audio consistently shows a low value for the ID feature, while the OOD features are relatively higher because the stimulated phoneme representations are not present in the pseudo-positive bag. In contrast, the abnormal audio of [tɕʌŋ] produces a high ID score at the beginning, indicating an error in the initial consonant (marked in pink). In the normal segment of the abnormal recording, the OOD features remain higher than the ID. During silence, both ID and OOD values are inactive. This demonstrates that SylPh provides both syllable-level outputs and phoneme-level analysis, offering detailed diagnostic insights.

5. Conclusion

In this paper, we introduce SylPh, a novel syllable-level MDD model for automatic speech audiometry. To address data imbalance, we construct a positive bag and a pseudo-negative bag using only normal data. The model is then guided by a simple text prompt with a contrastive objective, enabling generation for open-set syllables. Additionally, SylPh provides phoneme-level analysis, enhancing its diagnostic capability. We believe that the proposed SylPh will be effectively utilized in real-world scenarios to assist audiologists.

6. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2025-RS-2023-00256615) grant funded by the Korea government (MSIT). Following are results of a study on the "Convergence and Open Sharing System" Project, supported by the Ministry of Education and National Research Foundation of Korea.

7. References

[1] A. Epstein, "Speech audiometry," *Otolaryngologic Clinics of North America*, vol. 11, no. 3, pp. 667–676, 1978.

[2] M. Shahin and B. Ahmed, "Phonological-level mispronunciation detection and diagnosis," in *Proc. INTERSPEECH 2024 – 25th Annual Conference of the International Speech Communication Association*, Kos, Greece, Sep. 2024, pp. 307–311.

[3] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.

[4] S. Ondáš, E. Kiktová, M. Pleva, M. Oravcová, L. Hudák, J. Juhár, and J. Zimmermann, "Pediatric speech audiometry web application for hearing detection in the home environment," *Electronics*, vol. 9, no. 6, p. 994, 2020.

[5] W. Hu, Y. Qian, F. 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.

[6] W.-K. Leung, X. Liu, and H. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8132–8136.

[7] H. Jean, N. Wallaert, A. Dreumont, G. Creff, B. Godey, and N. Paraouty, "Automating speech audiometry in quiet and in noise using a deep neural network," *Biology*, vol. 14, no. 2, p. 191, 2025.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, USA, Dec. 2022.

[9] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[10] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[11] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug.-Sep. 2021, pp. 4428–4432.

[12] X. Chen, H. Chen, S. Nan, X. Kong, H. Duan, and H. Zhu, "Dealing with missing, imbalanced, and sparse features during the analysis of electronic medical records: A systematic review," *JMIR Medical Informatics*, vol. 11, p. e42756, 2023.

[13] J.-S. Kim, D. Lim, H.-N. Hong, H.-W. Shin, K.-D. Lee, B.-N. Hong, and J.-H. Lee, "Development of korean standard monosyllabic word lists for adults (ks-mwl-a)," *Audiology and Speech Research*, vol. 4, pp. 126–140, Oct. 2008.

[14] S.-J. Cho, D. Lim, K.-W. Lee, H.-K. Han, and J.-H. Lee, "Development of korean standard bisyllabic word list for adults used in speech recognition threshold test," *Audiology and Speech Research*, vol. 4, pp. 28–36, Jun. 2008.

[15] J. Kim, J. Lee, K. W. Lee, J. Bahng, J. H. Lee, C.-H. Choi, S. J. Cho, E. Y. Shin, and J. Park, "Test-retest reliability of word recognition score using korean standard monosyllabic word lists for adults as a function of the number of test words," *Journal of Audiology and Otology*, vol. 19, no. 2, pp. 68–73, 2015.

[16] Z. Liu, M. Huckvale, and J. McGlashan, "Automated voice pathology discrimination from continuous speech benefits from analysis by phonetic context," in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 2158–2162.

[17] J. Cheng, Y. Yan, Y. Gao, X. Feng, Y. Wang, and J. Zhang, "A study of production error analysis for Mandarin-speaking Children with Hearing Impairment," in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 4840–4844.

[18] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, Oct. 2021.

[19] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7442–7446.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Virtual, Jul. 2021, pp. 8748–8763.