



Phonetic Posteriorgram-Based Phoneme Selection for Vocal Cord Disorder Classification in Continuous Mandarin Speech

Chih-Ning Chen¹, Yu-Lan Chuang¹, Ming-Jhang Yang¹, Wei-Cheng Hsu², Yung-An Tsou², Yi-Wen Liu¹

¹Department of Electrical Engineering, National Tsing Hua University, Taiwan

²Dept. of Otorhinolaryngology, Head and Neck Surgery, China Medical University Hospital, Taiwan

andrewman71@gapp.nthu.edu.tw, ywliu@ee.nthu.edu.tw

Abstract

Automatic classification of vocal cord disorders (VCDs) in dysphonia benefits society by enabling home screening when immediate clinical consultation is unavailable. Previous studies focused on VCD classification using single vowels or isolated words. This research advances the field by classifying VCDs from continuous speech, using data from diagnosed patients. By parsing continuous speech into phonemes based on phonetic posteriorgrams (PPGs), we investigated VCD classification using the Mel frequency cepstral coefficients (MFCCs) corresponding to each Mandarin phoneme as the features. Results show a 15% accuracy improvement over a baseline model that ignores phonetic context and a prior study on the same patients using single-word utterances and sustained vowels. Our findings enhance the understanding of phonetic characteristics in VCDs and underscore the significance of continuous speech in automatic classification.

Index Terms: vocal cord disorders classification, phonetic posteriorgrams, continuous speech, machine learning

1. Introduction

Vocal cord disorders (VCDs) are prevalent in cases of dysphonia, significantly impacting speech production and quality. Common VCDs include vocal cord polyps, palsy, atrophy, paralysis, and laryngitis. These disorders often present with similar symptoms, such as hoarseness, sore throat, cough, and a sensation of tightness in the throat. However, the underlying causes of different VCDs vary considerably. While traditional diagnosis relies on imaging and clinical expertise, automatic detection offers a promising, efficient alternative for early intervention.

Previous studies have primarily focused on VCD classification using isolated speech samples and their acoustic features such as spectrum and Mel-frequency cepstral coefficients (MFCCs) [1][2]. However, isolated speech samples do not fully capture the natural flow of speech, while acoustic features may not sufficiently differentiate between VCD types. Despite the application of deep learning models, performance remains limited, suggesting the potential for exploring alternative features more closely related to VCD characteristics. Moreover, interpreting results solely from deep learning models is highly challenging, as it is difficult to pinpoint the features most relevant to VCDs[3][4]. This limitation significantly hampers detailed analysis and understanding of the underlying characteristics of VCDs.

To address these challenges, we collected continuous speech data from 380 hospital patients diagnosed with various VCDs. Unlike previous studies that used isolated vowels and words, our dataset captures the natural flow of speech. This

comprehensive data collection allows for a more accurate representation of how VCDs manifest in longer speech utterances. The longer, word-rich utterances contain more acoustic features, thus providing a richer source for analysis and classification.

To capture the rich articulation information during continuous speech, we need to determine which types of features contain the information that might be crucial for VCD classification. Phonetic posteriorgrams (PPGs) are an excellent way to achieve this goal. PPGs, originating from the work of Hazen et al. [5], are extracted from the spectrum of speech signals. A PPG chart illustrates the probability of each phonetic class at each moment in time, which provides a compact representation of the phonetic characteristics in speech. By using phonetic features like PPGs, we can identify which phonemes are articulated most differently across various VCDs.

Upon capturing PPG as a phonetic feature, we select specific phonemes and extract their corresponding acoustic features, specifically using MFCCs, which provide a more concise and targeted analysis than full-spectrum approaches. We employed machine learning methods such as neural networks (NN) and support vector machines (SVM) to classify VCDs. Instead of relying on the currently popular pretrained deep learning models, this approach allowed us to achieve high classification accuracy while providing more comprehensible explanation of speech-related disorders. Our contributions can be summarized as follows:

1. We incorporate PPGs to select specific phonemes and extract corresponding MFCCs, significantly enhancing VCD classification accuracy beyond previous works.
2. We collected continuous speech data from real-world hospital scenarios and utilized it to classify VCDs.
3. We analyze critical phonemes in VCD classification, contributing to a deeper understanding of how VCDs explicitly affect the articulation of different phonemes.

2. Related Work

VCD classification aims to identify vocal cord diseases by analyzing features associated with pathological voices. Kadiri et al. investigated glottal source features to differentiate between pathological and normal voices [6]. Recent studies have demonstrated the effectiveness of acoustic features, such as spectral analysis and MFCCs, in detecting pathological voices [7][8]. These techniques have also shown promise in identifying other diseases and disorders from voice recordings [9][10]. However, their approach, which focused on single vowels and isolated voice recordings, may not be ideal for VCD classification because it potentially omits crucial continuous acoustic information needed for comprehensive analysis. Additionally, it is

important to recognize that these methods might inadvertently capture information unrelated to the specific causes of the disorders. Wang et al. [11] and Chowdary et al. [12] have implemented deep learning models, including bi-directional long-short term memory networks, gated recurrent units, and transformer models, on datasets composed of word recordings and continuous speech. While these approaches provide a richer analytical context than those based solely on single vowels, the lack of detailed feature extraction has hindered a deeper understanding of the outcomes.

Liu et al. developed a method to detect VCDs by analyzing continuous speech within phonetic contexts, specifically to distinguish spasmodic dysphonia from vocal fold palsy [13]. They used forced alignment in a speech recognition system to identify phoneme-specific time segments, and employed a low-level descriptors configuration for feature extraction in training their classification model. In the present research, we argue that PPG can more accurately capture the true articulatory movements by directly extracting phonetic features from speech recordings, offering a precision advantage over acoustic analyses through forced alignment.

PPG was originally designed to enable speech analysis based on phonetic similarity without relying on large-scale data-trained speech recognition models. Empirically, PPG works well for tasks with limited corpora, such as spoken short queries [5], and it has been extensively utilized across various applications, including voice conversion [14][15], text-to-speech synthesis [16], language learning [17][18], and health examinations [19][20]. It serves as a pivotal bridge between spoken language and written text, enabling the extraction of detailed articulation information from utterances.

In our study, we utilized a dataset of continuous speech that we collected, aiming to capture a comprehensive range of articulatory information. We extracted phonetic features from these datasets, employing both PPG and MFCC to incorporate phonetic and acoustic features, respectively. This novel approach solves the challenges of capturing real utterance features for VCD classification and thus has the potential to enhance the performance of VCD classification and to identify which phonetic features are most influential in distinguishing VCDs.

3. Materials and Methods

3.1. Dataset Description

We established a specialized dataset for our study, sourced from the Voice and Swallowing Center of China Medical University Hospital in Taichung, Taiwan. This research is dedicated to the classification of three primary types of VCD: vocal cord atrophy, vocal cord paralysis, and benign organic lesions (BOL). Each category encompasses a range of subtypes—atrophy includes conditions such as sulcus and general thinning; paralysis is primarily associated with palsy; and BOL covers a variety of issues, including polyps, nodules, cysts, and edema.

Table 1: Number of cases, gender distribution (Male (M), Female (F)), average speech duration, and average age for each VCD type.

Disorder	M	F	Total	Duration(s)	Age
Atrophy	62	57	119	27.4±8.2	59.3±11.4
Palsy	55	39	94	31.6±10.4	59.5±13.4
BOL	32	100	132	27.7±8.0	54.1±11.7

Table 1 presents the number of cases, the average duration, and the average age with standard deviation for each audio recording in the VCD datasets. It shows the VCD types and gender distributions, including a total of 345 cases with an average duration of 28.63±8.95 seconds. Each case involves recording an audio of a script in Mandarin, detailed in Table 2, alongside its phonetic transcription used in our experiments.

The audio recordings in this dataset consist of continuous speech. This approach more closely mimics the natural speaking patterns observed in everyday communication. The paragraph, lasting approximately 30 seconds, incorporates a variety of vocabulary, which enriches the phonetic features captured by the PPG. It is important to note that although each individual reads the same script, actual pronunciation styles vary from person to person, resulting in PPGs that may appear similar but are not identical. This variability is crucial for extracting rich, individualized phonetic features that reflect more realistic speech dynamics.

3.2. Feature Extraction

In our research, audio files were processed to extract MFCCs as speech features, with 40 coefficients selected for each sample. The extraction utilized short-time Fourier transform settings of a 25ms window length and a 10ms hop length, at a sampling rate of 16 kHz. Figure 1 shows an example of a PPG. The PPG was generated using an acoustic model developed with the open-source toolkit Kaldi [21], which was trained on the AISHELL-1 corpus [22]. This model is specifically designed for Mandarin, as the AISHELL-1 corpus is composed of Mandarin recordings.

Strictly speaking, the PPG identifies the initials (consonants) and finals (vowels) in Mandarin, rather than actual phonemes. However, in Mandarin, these initials and finals form a set of pronunciation units and they function similarly to phonemes in their linguistic role. Therefore, throughout this article, we will refer to these units extracted via PPG as “phonemes” for simplicity and clarity, although they technically consist of initials and finals. PPG uses a sampling rate of 16 kHz and a window size of 25ms. Note that the hop size used in Kaldi is 30ms, resulting in three times fewer PPG frames than speech feature frames for the same audio file.

3.3. Alignment and Phoneme Selection for MFCC Extraction

After extracting both MFCCs and PPGs, we begin by aligning these features. Due to the differing hop sizes, the number of MFCC frames is three times that of PPG frames, with each PPG frame corresponding to three MFCC frames. Subsequently, we select specific target phonemes and extract the corresponding MFCC frames for each occurrence of these phonemes within the speech signal. These extracted frames are averaged and used as input for training machine learning models for VCD classification. This approach refines the focus on relevant phonetic characteristics and improves accuracy by filtering out irrelevant information, thereby enhancing the interpretability of the classification results.

3.4. Machine Learning Model Settings

3.4.1. Neural Network

The architecture of the neural network consists of sequential layers with dimensions (40, 128), (128, 32), and (32, 3), utilizing the ReLU activation function. The network was trained over 50 epochs with a learning rate of 0.001 and the Adam optimizer.

Table 2: Script, phonetic transcription, and Pinyin transcription of Mandarin sentences utilized in VCD classification. This table presents each sentence from the test paragraph in traditional Mandarin characters alongside its corresponding phonetic transcription and Romanized pronunciation.

Mandarin Script	Phonetic Transcription	Romanized Pronunciation
我聽見有人敲門 緩緩的說了一聲：請進來 門開了，我看見一個年輕人 瘦長的身體，明亮的眼睛 還有一張誠懇的臉。 看他臉上的表情 以及誠懇的態度 好像有什麼事情要我幫助。	uu uo3 t ing1 j ian4 ii iu3 r en2 q iao1 m en2 h uan3 h uan3 d i5 sh uo1 l e5 ii i4 sh eng1 q ing3 j in4 l ai2 m en2 k ai1 l e5 uu uo3 k an4 j ian4 ii i2 g e4 n ian2 q ing1 r en2 sh ou4 ch ang2 d e5 sh en1 t i3 m ing2 l iang4 d e5 ii ian3 j ing1 h ai2 ii iu3 ii i1 zh ang1 ch eng2 k en3 d e5 l ian3 k an4 t a1 l ian3 sh ang4 d e5 b iao3 q ing2 ii i3 j i2 ch eng2 zh ix4 d e5 t ai4 d u4 h ao3 x iang4 ii iu3 sh en2 m e5 sh ix4 q ing5 ii iao4 uu uo3 b ang1 zh u4	wǒ tīngjiàn yǒurén qiāomén huǎn huǎn de shuōle yī shēng: "Qǐng jìn lái" mén kāile, wǒ kànjiàn yīgè niánqīng rén shòucháng de shēntǐ, míngliàng de yǎnjīng hái yǒu yī zhāng chéngkěn de liǎn. kàn tā liǎn shàng de biǎoqíng yǐjī chéngzhì de tàidù hǎoxiàng yǒu shénme shìqíng yào wǒ bāngzhù.

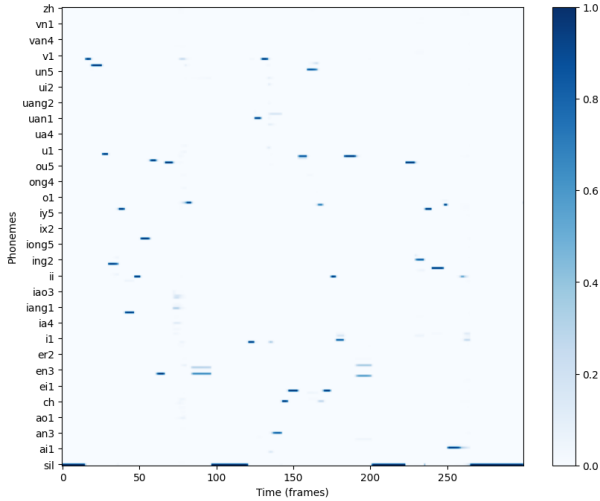


Figure 1: Example of a PPG from the present dataset. The horizontal axis shows time in frames, while the vertical axis lists Mandarin-based phonemes extracted using the Kaldi toolkit, which was trained on the AISHELL-1 corpus. Color intensity indicates the probability of each phoneme’s occurrence, with higher probabilities shown in deeper blue and lower probabilities in white. For display clarity, only 30 of the total 217 phonemes are shown.

3.4.2. Support Vector Machine

To select the optimal SVM model, we performed a grid search over a range of parameters, including the penalty parameter C , kernel types, and kernel coefficient γ . The parameter grid is detailed in Table 3.

Table 3: SVM Parameter Grid

Parameter	Values
C	0.1, 0.5, 1, 5
γ	0.1, 0.01, 'scale' 1
Kernel	'linear', 'rbf'

'scale' adjusts γ based on the number of features.

$$\gamma = 1/n_{\text{features}}.$$

Table 4: The accuracy (%) of NN and SVM models from average of MFCCs as an ablation study.

Model	Male	Female	Male & Female
NN	55.1±8.3	52.9±7.6	50.0±6.6
SVM	54.4±7.3	53.6±5.6	49.9±4.6

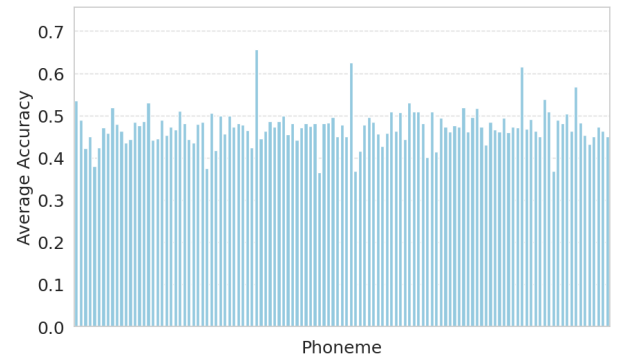


Figure 2: Accuracy distribution by phoneme for the SVM model predictions in the combined gender task. The horizontal axis represents the phonemes analyzed, totaling 119 out of the complete set of 217, after excluding those that appeared less than 40 times. Due to the large number of phonemes, individual names are not displayed on the axis to maintain clarity.

4. Results

In our experiments, we used 5-fold cross-validation combined with 10 different random seeds to assess the model’s performance. The results are reported as means and standard deviations of accuracy for 3-class (atrophy, paralysis, BOL) VCD classification. Only phonemes that appeared more than 40 times in the dataset were included in the analysis.

4.1. Baseline Results

In our study, we investigate the influence of phonemes on VCD classification. We present baseline results derived from machine learning models trained on features generated through the average pooling of MFCCs without selecting frame segments of the target phoneme. Table 4 displays these baseline results. Notably, separating genders has a discernible impact on performance. The results for combined gender classification are slightly lower compared to male or female. Specifically, in the combined gender scenario, the mean baseline accuracy is 0.500. Our prior work on the same task—3-class VCD classifi-

Table 5: Comparison of the accuracy (%) of NN and SVM on VCD classification, presented separately for male, female, and combined male and female groups. Each group lists the top 10 phonemes by accuracy.

Male				Female				Male & Female			
Neural Network		SVM		Neural Network		SVM		Neural Network		SVM	
Phoneme	Accuracy	Phoneme	Accuracy	Phoneme	Accuracy	Phoneme	Accuracy	Phoneme	Accuracy	Phoneme	Accuracy
vv	56.2±13.0	vv	66.2±11.4	ang3	55.8±13.3	ang3	65.5±13.4	ie1	53.2±14.2	eng4	65.6±12.8
s	56.1±9.1	u3	62.7±13.0	iao1	54.9±13.6	ai3	59.8±11.5	eng4	52.4±14.5	ie1	62.5±12.7
q	55.1±9.3	ian1	62.3±15.5	en3	54.3±7.6	ei3	59.6±14.5	iao1	50.7±10.2	ua4	61.5±17.1
c	54.8±10.9	iao1	61.6±12.9	ang2	54.2±13.0	ing5	58.6±12.5	en3	50.4±7.5	v2	56.8±12.9
ian1	54.1±13.9	ai3	59.8±11.4	ing5	54.0±16.7	ie4	58.6±10.8	uan4	49.0±9.8	uang2	53.9±13.1
an4	53.9±8.2	ing4	59.7±11.2	m	53.9±6.9	an2	58.2±12.6	ix4	48.7±6.5	ing5	53.1±11.1
ao4	53.7±14.8	aa	59.5±11.7	ing3	53.0±10.8	uo4	57.8±16.0	e5	48.6±5.9	ang3	53.0±7.7
l	53.4±7.0	e2	59.0±13.5	e5	52.9±6.9	k	56.8±6.7	ing5	48.5±11.3	m	51.9±5.9
z	52.9±8.3	an4	57.7±7.3	en1	52.8±9.3	iao1	56.8±11.8	an4	48.5±5.6	ai3	51.9±7.8
ai3	52.8±15.8	iu3	57.4±7.5	d	52.5±6.7	uan2	56.7±9.2	uo3	48.4±5.7	ou3	51.8±13.9

cation—shows similar performance, with the best result achieving 0.507, which closely aligns with the baseline result in this research [2].

4.2. Impact of Phoneme Selection

Figure 2 reveals a substantial variation in VCD classification accuracy across different phonemes, underscoring the pivotal influence of target phonemes on overall classification performance. Table 5 presents the experimental results from both the neural network and SVM models across different genders.

In the combined gender results, the extraction of the /eng4/ phoneme and subsequent utilization of its corresponding frame segments yielded a 15.7% increase in accuracy when employing the SVM model. Similarly, accuracy improvements of 11.8% and 11.9% were observed in the male and female tasks, with the male improvement linked to the /vv/ phoneme and the female improvement to the /ang3/ phoneme, respectively. The models achieved higher classification accuracy than the baseline when utilizing a more compact feature set, where only phoneme-specific MFCC segments were retained instead of the full MFCC representation. This underscores the importance of selecting phonemes that capture critical articulation features of VCDs. However, this improvement was not universal across all phonemes. Figure 2 reveals that 101 phonemes performed below baseline, suggesting that a large portion of articulation data may be redundant for classification. Systematic evaluation of Mandarin phonemes identified a subset that improves classification accuracy, confirming the effectiveness of targeted phoneme selection for VCD classification.

4.3. Gender-Based Differences in Phoneme Performance

From Table 5, we observe that the top 10 phonemes vary by gender. For instance, /vv/ ranks first in the male results but drops to 24th in the female results. Additionally, there is minimal overlap among the top-performing phonemes across male, female, and combined gender results. This variation arises from differences in articulation frequency composition, leading to distinct MFCC patterns even when the same phonemes are produced. These findings underscore the necessity of considering gender when classifying VCDs, as pronunciation differences significantly impact classification accuracy.

Further analysis of the best-performing phonemes from the SVM results reveals that male speakers exhibit a strong presence of the /i/ series phonemes, such as /ian1/, /iao1/, and /iang4/. These phonemes involve a high front unrounded vowel /i/, which requires precise tongue positioning and dynamic movement during articulation. In Mandarin, /i/-based

diphthongs and nasals often demand greater control of tongue height and advancement, both of which are susceptible to disruptions caused by VCDs. Medical studies indicate that vocal cord impairments can alter vowel articulation by affecting the stability of high front vowels, making them more acoustically distinct in pathological speech. Additionally, clinical research highlights that high vowels like /i/ are more prone to phonation instability due to their reliance on controlled airflow and vocal fold tension [23]. Given that male speakers generally have lower fundamental frequencies, the resulting differences in vocal tract resonance amplify the articulatory instabilities of high front vowels. In contrast, female speakers show higher classification accuracy for phonemes such as /ang3/, /an2/, and /uan2/, which involve complex tongue and airflow coordination. Their naturally higher fundamental frequency leads to greater formant dispersion, amplifying VCD-induced distortions [24]. This effect is particularly evident in nasalized vowels like /ang3/ and /an2/, where vocal tract resonance changes due to VCDs further enhance their distinctiveness.

Notably, /ai3/ is a unique phoneme, ranking fifth in the male group and second in the female group classification. This can be attributed to its composition, as it combines characteristics of both /i/, which is well-classified in males, and /a/, which is more distinctive in females. As a falling diphthong transitioning from a low open front vowel /a/ to a high front vowel /i/, /ai3/ requires precise tongue movement and airflow control, making it highly susceptible to VCD-induced impairments. Vocal cord paralysis can disrupt glottal closure, while vocal cord atrophy weakens phonation control, reducing diphthong stability. Additionally, BOL alters vowel resonance, further distinguishing /ai3/ in pathological speech. These effects make /ai3/ a strong acoustic marker for VCD classification.

5. Conclusion

In this paper, we incorporate PPGs to harness phonetic features for classifying VCDs in continuous Mandarin speech. By integrating these features with MFCCs and conducting a comprehensive evaluation of all potential phonemes, we pinpointed specific target phonemes that significantly boosted classification accuracy from 50% to 65.6% in combined gender scenarios for 3-class VCD classification. Our findings reveal that phonetic articulation varies significantly across different VCDs and genders, enriching our understanding of VCD classification and underscoring the critical role of phonetic features. This research not only enhances VCD classification using speech audio but also raises awareness of the importance of considering phonetic features in future related studies.

6. Acknowledgements

This research was supported by the National Science and Technology Council of Taiwan under grant No. 112-2410-H-007-048 and 113-2221-E-007-100-MY2.

7. References

- [1] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, vol. 33, no. 5, pp. 634–641, 2019.
- [2] C.-C. Chen, W.-C. Hsu, T.-H. Lin, K.-D. Chen, Y.-A. Tsou, and Y.-W. Liu, "Classification of vocal cord disorders: Comparison across voice datasets, speech tasks, and machine learning methods," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 1868–1873.
- [3] M. T. Ribeiro, S. Singh, and C. Guestrin, "“why should I trust you?” explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [4] S. Luo, H. Ivison, S. C. Han, and J. Poon, "Local interpretations for explainable natural language processing: A survey," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–36, 2024.
- [5] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 421–426.
- [6] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 367–379, 2019.
- [7] H.-C. Kuo, Y.-P. Hsieh, H.-H. Tseng, C.-T. Wang, S.-H. Fang, and Y. Tsao, "Toward real-world voice disorder classification," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 10, pp. 2922–2932, 2023.
- [8] E. C. Compton, T. Cruz, M. Andreassen, S. Beveridge, D. Bosch, D. R. Randall, and D. Livingstone, "Developing an artificial intelligence tool to predict vocal cord pathology in primary care settings," *The Laryngoscope*, vol. 133, no. 8, pp. 1952–1960, 2023.
- [9] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic covid-19 diagnosis via voice and symptoms from crowd-sourced data," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8328–8332.
- [10] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for health: the use of vocal biomarkers from research to clinical practice," *Digital Biomarkers*, vol. 5, no. 1, pp. 78–88, 2021.
- [11] S.-S. Wang, C.-T. Wang, C.-C. Lai, Y. Tsao, and S.-H. Fang, "Continuous speech for improved learning pathological voice disorders," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 3, pp. 25–33, 2022.
- [12] P. N. Chowdary, M. S. Akshay, V. S. Aravind, M. S. Aashish, G. V. V. Vardhan, and G. J. Lal, "A few-shot approach to dysarthric speech intelligibility level classification using transformers," in *2023 14th International Conference on Computing Communication and Networking Technologies*. IEEE, 2023, pp. 1–6.
- [13] Z. Liu, M. Huckvale, and J. McGlashan, "Automated voice pathology discrimination from continuous speech benefits from analysis by phonetic context," in *Proceedings of Interspeech*. ISCA, 2022, pp. 2158–2162.
- [14] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.
- [15] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5314–5318.
- [16] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," in *Proceedings of Interspeech*, 2016, pp. 322–326.
- [17] Y.-B. Wang and L.-S. Lee, "Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8232–8236.
- [18] A. Sini, A. Perquin, D. Lolive, and A. Delhay, "Phone-level pronunciation scoring for I1 using weighted-dynamic time warping," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1081–1087.
- [19] G. Gosztołya, V. Svindt, J. Bóna, and I. Hoffmann, "Extracting phonetic posterior-based features for detecting multiple sclerosis from speech," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3234–3244, 2023.
- [20] A. K. Suresh, S. R. KM, and P. K. Ghosh, "Phoneme state posteriorgram features for speech based automatic classification of speakers in cold and healthy condition," in *Proceedings of Interspeech*, 2017, pp. 3462–3466.
- [21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldı speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.
- [22] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.
- [23] E. J. Park, J. H. Kim, Y. H. Choi, J. E. Son, S. A. Lee, and S. D. Yoo, "Association between phonation and the vowel quadrilateral in patients with stroke: A retrospective observational study," *Medicine*, vol. 99, no. 39, p. e22236, 2020.
- [24] E. J. Hunter, K. Tanner, and M. E. Smith, "Gender differences affecting vocal health of women in vocally demanding careers," *Logopedics Phoniatrics Vocology*, vol. 36, no. 3, pp. 128–136, 2011.