Towards Robust Automated Perceptual Voice Quality Assessment with Deep Learning

Whenty Ariyanti, *Student Member, IEEE*, Kuan-Yu Chen, *Member, IEEE*, Sabato Marco Siniscalchi, *Member, IEEE*, Hsin-Min Wang, *Senior Member, IEEE*, and Yu Tsao, *Senior Member, IEEE*

Abstract*— *Objective:* Perceptual voice quality assessment plays a critical role in diagnosing and monitoring voice disorders by providing standardized evaluation of vocal function. Traditionally, this process relies on expert raters utilizing standard scales, such as the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) and Grade, Roughness, Breathiness, Asthenia, and Strain (GRBAS). However, these metrics are inherently subjective and susceptible to inter-rater variability, which motivates the need for automated and objective assessment methods. *Methods:* In this study, we propose Voice Quality Assessment Network (VOQANet), a deep learning-based framework with an attention mechanism that leverages a Speech Foundation Model (SFM) to capture high-level acoustic and prosodic information from raw speech for automated voice quality assessment. To further enhance robustness and interpretability, we present VOQANet+, which integrates handcrafted acoustic features such as jitter, shimmer, and harmonics-to-noise ratio (HNR) with SFM embeddings into a hybrid representation. Unlike previous studies that focus only on vowel-based phonation (the PVQD-A subset) on the Perceptual Voice Quality Dataset (PVQD), we evaluate our models on both vowel-based phonation and sentence-level speech (the PVQD-S subset) to improve generalizability in real-world applications. *Results:* Experimental results show that sentence-based input yields stronger performance than vowel-based input, especially at the patient level, highlighting the value of longer utterances in capturing perceptual voice attributes. VOQANet consistently outperforms the baseline methods in terms of root mean squared error (RMSE) and Pearson correlation coefficient (PCC) across CAPE-V and GRBAS dimensions, while VOQANet+ performs even better. Additional experiments under noisy conditions show that VOQANet+ maintains higher prediction accuracy and robustness. *Conclusion:* These findings highlight the effectiveness of combining SFM embeddings with domain-informed acoustic features for interpretable and resilient automated voice quality assessment. *Significance:* VOQANet+ shows strong potential for deployment in real-world and telehealth settings, addressing the limitations of subjective perceptual assessments with an interpretable and noise-resilient solution.

***Index Terms*— Perceptual voice quality assessment, VOQANet, CAPE-V, GRBAS, speech foundation models, voice disorder assessment**

Manuscript received xxxx; revised xxxx, 2025; accepted xxxxx. This work was supported by xxxx. The review of this paper was arranged by Editor Paolo xxxxx. (Corresponding author: Yu Tsao.)

Whenty Ariyanti is with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan, and also with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan (e-mail: d11115805@mail.ntust.edu.tw).

Kuan-Yu Chen is with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: kychen@mail.ntust.edu.tw).

Sabato Marco Siniscalchi is the University of Palermo, Palermo, Italy (e-mail: sabatomarco.siniscalchi@unipa.it).

Hsin-Min Wang is with the Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan (e-mail: whm@iis.sinica.edu.tw).

Yu Tsao is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan, and also with the Department of Electrical Engineering, Chung Yuan Christian University, 11529, Taiwan (e-mail: yu.tsao@citi.sinica.edu.tw).

I. Introduction

VOICE disorders are common in modern society, and impaired voice quality can seriously affect an individual's communication ability and social well-being [1], [2]. Voice disorders can be caused by a variety of diseases, including vocal fold nodules, polyps, paralysis, neurological diseases, such as Parkinson's disease, and head and neck cancers [3]. These diseases often affect vocal characteristics, such as hoarseness, breathiness, roughness, or strain, which require perceptual examination by trained clinicians. Therefore, vocal signal analysis has become a non-invasive and accessible screening tool and is increasingly being adopted by otolaryngology and neurology clinics to help early detection and monitoring of voice-related disorders [4]. Auditory-perceptual voice quality assessment (VQA) aims to improve the diagnosis, monitoring, and treatment of voice disorders by providing an objective and standardized assessment of vocal function. It serves as a critical tool for identifying pathological voice conditions, tracking disease progression, and evaluating the effectiveness of therapeutic interventions. Traditionally, VQA relies on perceptual assessments by experienced clinicians using standardized scales, such as the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) and the Grade, Roughness, Breathiness, Asthenia, Strain (GRBAS) [5], [6], [7]. CAPE-V has been adapted to multiple languages, including French, Turkish, European Portuguese, and Japanese, further demonstrating its clinical relevance and international applicability [8], [9], [10], [11]. It provides continuous ratings (0–100) for perceptual attributes, whereas GRBAS uses a discrete 4-point ordinal scale (0–3). For both, the larger the value, the more severe the condition. Although these perceptual ratings are widely used and provide valuable qualitative insights into voice disorders [12], they are inherently subjective and prone to inter- and intra-examiner variability. Recent studies have explored other strategies, such as crowdsourcing perceptual ratings, particularly in neurological disorders like Parkinson's disease, to improve scalability and maintain rating validity [13]. The reliance on expert raters makes standardization difficult and increases the need for automated VQA.

To address the above challenges, several machine learning (ML) and deep learning methods have been explored to automate perceptual voice assessment [14]. Traditional ML models, such as Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN), have been used to predict CAPE-V values based on handcrafted acoustic features [15], [16]. Dispite providing interpretability, they often struggle with generalizability and fail to capture the complex characteristics of speech that influence perceptual voice quality ratings. To overcome these limitations, recent studies have employed deep learning and ensemble frameworks for pathological voice classification [2], demonstrating the benefits of combining multiple modalities and learned representations for more robust voice assessment. Recent advances in Speech Foundation Models (SFM), especially large-scale pre-trained models such as WavLM [17], HuBERT [18], and Whisper [19], have performed well in extracting high-level speech representations. WavLM and HuBERT are trained using self-supervised learning (SSL), while Whisper adopts a semi-supervised paradigm, leveraging large-scale audio-text pairing data collected through weak supervision. These models provide robust and transferable features, which are well-suited for downstream tasks, including VQA. Notably, WavLM incorporates a denoising pre-training objective, which may improve the model's robustness in noisy clinical recordings; while Whisper's multilingual training provides a wider range of phonetic coverage, although there may be a domain mismatch with disordered speech.

In this study, we propose VOQANet (Voice Quality Assessment Network), a deep learning-based framework with an attention mechanism that leverages SFM embeddings of WavLM for perceptual VQA.To further enhance clinical relevance and model robustness, we further propose VOQANet+, an extended version that combines SFM embeddings with handcrafted acoustic features such as jitter, shimmer, and harmonics-to-noise ratio (HNR), which have been widely used in vocal pathology analysis for their ability to capture voice irregularities and instabilities [20]. This combination enables VOQANet+ to benefit from both high-level learned representations and complementary low-level signal-based features. We evaluate our models on the Perceptual Voice Quality Dataset (PVQD), reporting both utterance-level and patient-level results, where the latter predictions are averaged across each speaker's utterances. Experimental results show that VOQANet provides a strong baseline, while VOQANet+ consistently improves prediction accuracy and generalization, especially under noisy conditions.

The main contributions of this study are summarized as follows: First, we propose VOQANet, a deep learning framework with an attention mechanism that systematically evaluates the effectiveness of SFM embeddings of pre-trained speech models for perceptual VQA. Second, we propose VOQANet+, an extended version of VOQANet that combines handcrafted acoustic features with SFM embeddings to improve model interpretability and performance by integrating domain-specific knowledge. Third, we conduct a comprehensive evaluation on the PVQD dataset, including both utterance-level and patient-level evaluations, aligning with clinical assessment practice. Through comprehensive evaluations on the PVQD dataset, our results demonstrate the potential of SFM-driven methods for robust, interpretable, and clinically relevant automated VQA.

II. Related Work

A. Automated Perceptual Voice Quality Assessment

Ensuring robustness and generalizability is critical for real-world applications of perceptual VQA. Traditional methods rely on machine learning models, such as RF, SVM, and KNN, which utilize handcrafted acoustic features like jitter, shimmer, zero crossing rate, and HNR [16]. A lightweight feature extraction method has been proposed to leverage these classical ML models for CAPE-V prediction. However, while these models provide interpretability and domain relevance, they often struggle to generalize across datasets due to the variability of speech patterns and the sensitivity of handcrafted acoustic features to recording conditions [16]. While CAPE-V provides continuous ratings and GRBAS uses a discrete scale, both are susceptible to inter-rater differences [21], [22], which further motivates the development of objective, automated assessment methods.

To overcome these limitations of classical methods, recent studies have adopted deep learning-based methods, especially leveraging SFMs such as Whisper [19] and WavLM [17], which learn rich acoustic representations from large-scale raw waveforms. Among these models, WavLM includes a denoising pre-training objective, which enables the model to learn robust representations even in the presence of background noise and acoustic variations. This feature is particularly beneficial for disordered speech, which often deviates from typical acoustic patterns. On the other hand, Whisper is trained on a large-scale multilingual and multitask corpus, which makes it highly effective in a variety of automatic speech recognition (ASR) tasks. Whisper has been explored for other applications, such as improving the performance of speaker verification (SV) tasks [23]. These findings highlight the potential of SFMs in clinical applications, while also underscoring the need to tailor the representations to domain-specific characteristics. However, SFM embeddings are typically learned from general speech corpora and may not fully capture task-specific or clinically salient characteristics relevant to voice pathology. To overcome this limitation, recent studies have explored hybrid modeling strategies that combine the generalization capabilities of SFM embeddings with handcrafted features to improve the prediction performance of VQA systems. This has shown that such hybrid audio representations can simultaneously enhance robustness and interpretability in cognitively and physically demanding speech tasks [24]. Based on these insights, we explore deep learning-based methods that combine SFM representations with clinically relevant feature representations.

B. Speech Foundation Models

Recent advances in self-supervised learning have introduced SFMs, which provide more robust and generalizable representations for downstream speech tasks. These models have attracted much attention in the speech processing community due to their ability to learn meaningful representations directly from raw audio. Models such as WavLM and HuBERT are trained using SSL, where contextual representations are learned from unlabeled data, while Whisper takes a semi-supervised approach, leveraging large-scale audio-text pairing data collected with weak supervision. Compared to traditional supervised methods, SFMs trained using SSL or semi-supervised methods can capture low-level acoustic features and high-level linguistic patterns without the need for extensive annotations.

Models pre-trained on large-scale corpora such as Librispeech [25] have achieved excellent performance in various speech tasks such as ASR, SV, speech synthesis, and speech emotion recognition [26], [27]. In the context of VQA, SFMs have been explored for their ability to extract rich and contextualized acoustic representations that can be related to perceptual voice attributes such as breathiness, strain, and roughness. These deep representations are able to capture complex speech patterns that are usually difficult to model using only handcrafted acoustic features.

C. Hybrid Models Combining SFM and Handcrafted Features

Combining SFM embeddings with handcrafted acoustic features has been explored to capture data-driven and clinically interpretable speech characteristics. A hybrid approach combining data-driven features from a BYOL-derived model with handcrafted features extracted using openSMILE has demonstrated strong performance on speech analysis tasks [28]. Similarly, self-supervised speech models jave been shown to outperform traditional acoustic features in predicting parameters such as heart activity, highlighting the efficacy of combining SSL embeddings with handcrafted features [24]. While prior studies have explored hybrid designs in affective or biomedical speech tasks, few have systematically examined such approaches in the context of clinical voice assessment, particularly with respect to GRBAS and CAPE-V prediction. Although clinical findings suggest that CAPE-V ratings may correlate more strongly with objective acoustic and aerodynamic measures than GRBAS in certain populations [29], both scales remain widely used in perceptual VQA and offer complementary insights. Therefore, we include both in this study.

D. Evaluation Strategies in Voice Assessment Models

Assessment strategies play a critical role in evaluating the reliability and clinical applicability of voice assessment models. Most previous studies on automated VQA focus on predicting perceptual scores at the utterance level, where each audio segment is treated as an independent sample [16], [28]. While this approach enables fine-grained analysis, it may not capture broader patterns of the patient's overall vocal characteristics, especially when there are phonetic differences between different utterances. To improve reliability in clinical settings, some studies have begun to explore patient-level assessments by aggregating predictions from multiple utterances from the same individual [24]. This strategy better reflects the real-world diagnostic process, where clinicians assess voice quality based on a complete set of speech samples rather than isolated fragments.

III. Proposed Method

The overall architectures of VOQANet and VOQANet+ are shown in Fig. 1. VOQANet only uses SFM embeddings as input, while VOQANet+ combines SFM and handcrafted features as input.

A. VOQANet: SFM-Based Feature Learning

As shown in Fig. 1(a), VOQANet leverages SFM embeddings extracted by a pre-trained model to capture rich acoustic and prosodic characteristics of the input waveform. Given a waveform $\mathbf{w}$, the SFM embedding is calculated as a weighted sum of the hidden representations of all transformer layers:

$$X_S = \sum_{\ell=0}^{L} \alpha_\ell \cdot h^{(\ell)}(\mathbf{w}), \tag{1}$$

where $h^{(\ell)}$ represents the hidden state at layer ℓ of the pre-trained SFM, and α_ℓ is a learnable scalar weight normalized by softmax. This layer-wise aggregation strategy follows prior work that showed that weighted combinations of intermediate layers outperform single-layer embeddings in speech assessment tasks [30]. This mechanism allows the model to adaptively emphasize the layers most relevant to perceptual voice quality.

B. VOQANet+: Joint Representation Learning

As shown in Fig. 1(b), to further enhance the model's ability to capture clinically relevant speech characteristics, VOQANet+ combines handcrafted acoustic features with SFM embeddings. These handcrafted features include Jitter, Shimmer, and HNR, which are extracted from the same waveform $\mathbf{w}$ using signal processing techniques:

$$X_A = f_A(\mathbf{w}), \tag{2}$$

where $f_A(\cdot)$ denotes the handcrafted feature extraction function. The final feature representation is obtained by concatenating the two feature types:

$$X = [X_S, X_A]. \tag{3}$$

As a result, a 1027-dimensional hybrid feature vector sequence is obtained. This combination enables the model to jointly learn from high-level SFM embeddings and low-level handcrafted acoustic features, thereby improving the robustness and interpretability of perceptual VQA.

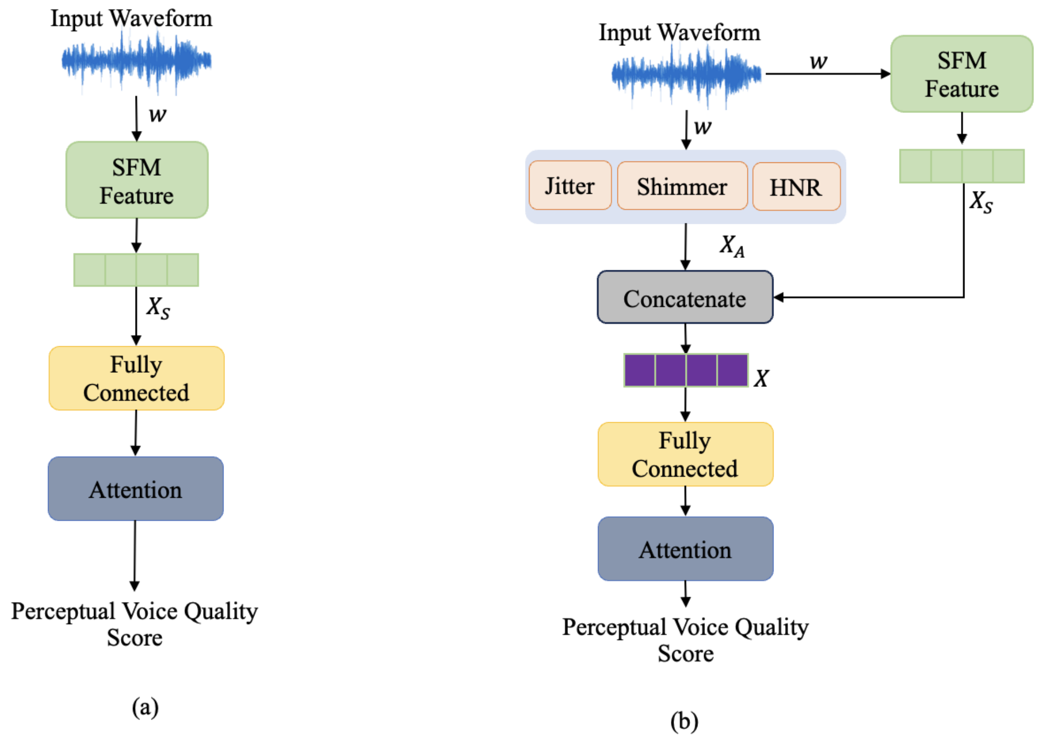


Fig. 1. Architectures of VOQANet (a) and VOQANet+ (b).

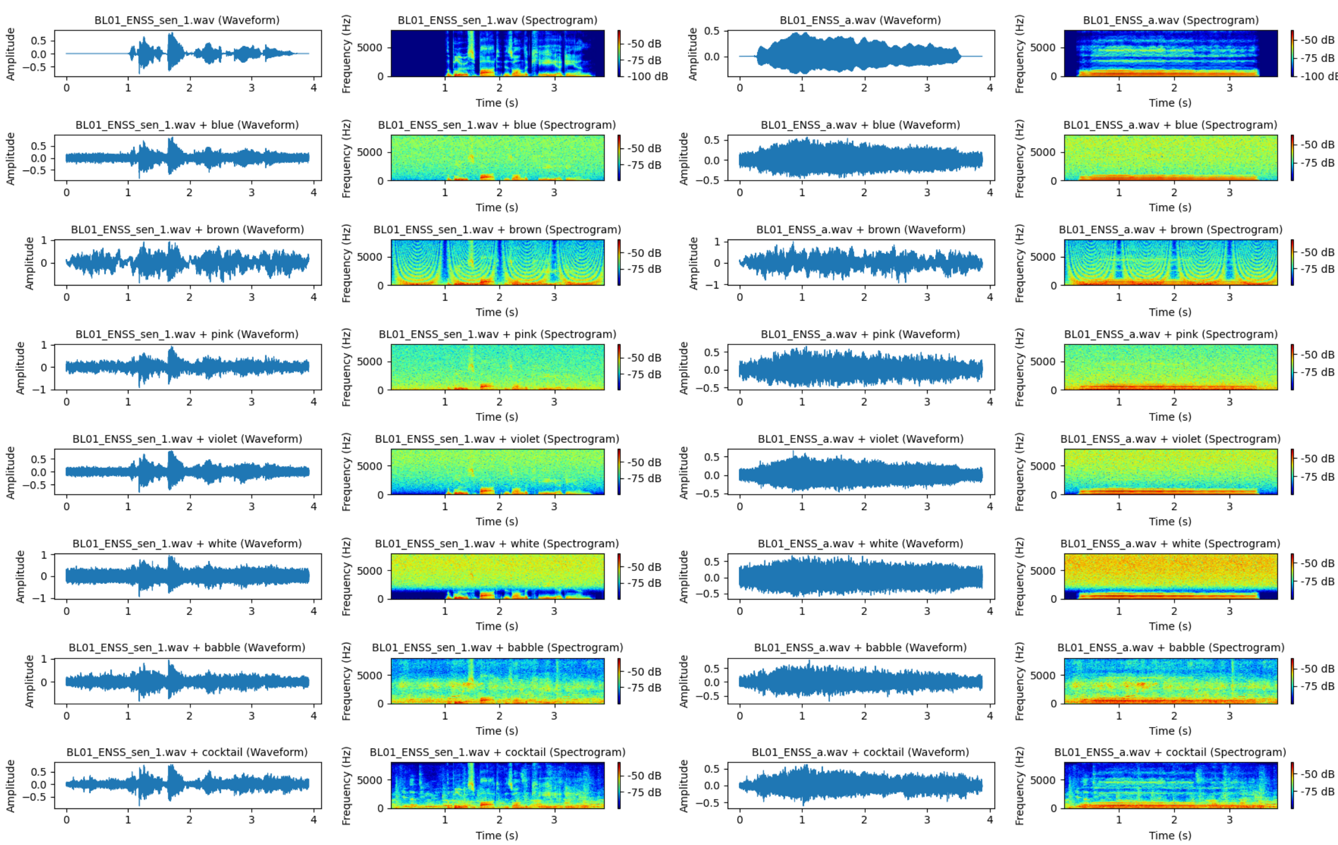


Fig. 2. Waveforms and Spectrograms of audio samples in PVQD-A and PVQD-S

C. Model Architecture and Training

Both VOQANet and VOQANet+ use the same regression backbone: a three-layer fully connected neural network with batch normalization, dropout, and ReLU activation:

$$H = \sigma(W_3 \cdot (\sigma(W_2 \cdot (\sigma(W_1 X + b_1)) + b_2)) + b_3), \quad (4)$$

where H is the latent representation, and $\sigma(\cdot)$ denotes the ReLU function. To make the network focus on the salient components of the learned features, attention-based pooling is applied after the last hidden layer to calculate a weighted summary of the feature representation. The attention module first projects H into an intermediate space using a non-linear transformation, and then calculates the attention weights over the feature dimension:

$$\alpha = \mathrm{softmax}(W_{\text{attn}} \cdot \tanh(W_h H + b_h) + b_{\text{attn}}). \quad (5)$$

The attended feature vector $\mathbf{z}$ is obtained as follows,

$$\mathbf{z} = \sum_{t=1}^{T} \alpha_t \cdot H_t, \quad (6)$$

where T is the length of the feature vector sequence, and H_t represents the feature vector at position t in the sequence. Finally, $\mathbf{z}$ is fed into the final regression layer to predict the VQA score.

To prioritize clinically significant deviations, we use the Weighted Mean Squared Error (WMSE) loss function [31]:

$$\mathcal{L}_{\text{WMSE}} = \frac{1}{N} \sum_{i=1}^{N} \left(1 + \frac{Y^i}{Y_{\max}} \cdot \beta\right) \cdot (\hat{Y}^i - Y^i)^2, \quad (7)$$

where N is the number of training samples; β is a hyperparameter that controls the degree of emphasis on higher severity levels; Y^i and $\hat{Y}^i$ are the predicted and ground-truth ratings of sample i, respectively; and $Y_{\max}$ is the maximum ground-truth score among the N training samples.

IV. Experiments

A. Dataset

The audio samples used in this study come from the Perceptual Voice Quality Database (PVQD) provided by The Voice Foundation [3]. This dataset contains 296 audio recordings, each featuring one person's sustained /a/ and /i/ vowels and continuous speech sentences defined by the CAPE-V protocol, providing a mixture of controlled and naturalistic speech material. The audio files are stored in WAV format, encoded with 16-bit resolution and a sampling rate of 44.1 kHz, ensuring high-fidelity acoustic analysis. In addition to raw audio, the dataset provides rich metadata including the speaker's age, gender, diagnosis, and expert perceptual ratings using the CAPE-V and GRBAS frameworks. CAPE-V scores are continuous, ranging from 0 to 100, and are used to assess dimensions such as overall severity, breathiness, and strain. GRBAS scores use a 4-point ordinal scale (0–3) to describe Grade, Roughness, Breathiness, Asthenia, and Strain [32]. Each recording was independently rated by two qualified speech-language pathologists, and the final perceptual score for each dimension was the average of the two raters to ensure reliability and reduce subjective bias. Table I summarizes the score distribution of each dimension of the two scales.

TABLE I
DESCRIPTIVE STATISTICS OF CAPE-V AND GRBAS RATINGS IN THE PVQD DATASET.

Scale	Attribute	Mean	Median	Mode	Min	Max
CAPE-V (0–100)	Severity	29.4	19.5	19.3	0.33	98.67
	Roughness	20.7	13.7	9.7	0.17	84.83
	Breathiness	19.8	12.2	5.0	0.00	99.50
	Strain	21.1	12.2	4.5	0.12	96.83
	Pitch	16.3	9.3	0.5	0.00	99.17
	Loudness	18.7	8.8	0.7	0.00	99.17
GRBAS (0–3)	Grade	1.0	0.8	0	0	3
	Roughness	0.8	0.7	0	0	3
	Breathiness	0.7	0.4	0	0	3
	Asthenia	0.6	0.2	0	0	3
	Strain	0.8	0.5	0	0	3

TABLE II
DEMOGRAPHIC INFORMATION OF SPEAKERS IN VOICE SAMPLES.

	Female/Male		Age (years)	
	Samples	Percentage (%)	Mean ±	Range
Training	143/83	63.3/36.7	46.31 ± 22.04	14-93
Testing	42/15	73.7/26.3	47.56 ± 21.04	18-90

TABLE III
DISTRIBUTION OF TRAINING AND TESTING SAMPLES FOR UTTERANCE- AND PATIENT-LEVEL ASSESSMENTS.

Evaluation Type	PVQD-A		PVQD-S	
	Training	Testing	Training	Testing
Utterance-Level Evaluation	226	57	1352	339
Patient-Level Evaluation	226	57	226	57

B. Data Split for Training and Testing

To prevent data leakage and ensure generalizability, the PVQD dataset was split at the patient level, meaning that each speaker was exclusively assigned to either the training set or the test set. This ensured that the model was evaluating voice quality on unseen speakers rather than memorized speech patterns, thus validating its ability to generalize beyond the training set. This approach follows that used in [16] to ensure consistency with previous studies. The PVQD dataset originally contained 296 recordings, but 13 corrupted files were excluded, leaving a total of 283 valid samples. Specifically, 226 samples were used for training and 57 samples for testing. Table II provides the demographic distribution of speakers in the training and test sets, including gender ratio and age range. For a more comprehensive evaluation, vowel segments and continuous speech segments were extracted from each recording. Therefore, the PVQD dataset was divided into two subsets: PVQD-A (vowel-only subset), which contains /a/ vowel segments; and PVQD-S (speech-based subset), which consists of continuous speech segments. In this way, vowel phonation (PVQD-A) and continuous speech (PVQD-S) were evaluated independently. The number of samples in each subset is shown in Table III. Since each recording contains one /a/ vowel and multiple continuous speech segments, there

TABLE IV
NOISE TYPES AND SNR LEVELS USED FOR TRAINING, SEEN-TEST, AND UNSEEN-TEST.

	Configuration	Details
Training	SNR Noise Type	-5 dB, 0 dB, 5 dB, 10 dB White noise, Pink noise, Cafeteria babble, & Cocktail party
Testing (Seen)	SNR Noise Type	-5 dB, 0 dB, 5 dB, 10 dB White noise, Pink noise, Cafeteria babble, & Cocktail party
Testing (Unseen)	SNR Noise Type	0 dB, 5 dB Brown noise Baby cry, & Laughter

are more samples of continuous speech than /a/ vowel.

C. Feature Extraction

All signals were resampled to 16 kHz before feature extraction. Two types of features were extracted. First, SFM embeddings from a pre-trained model (WavLM or Whisper) were used to capture phonetic and prosodic information. These embeddings were computed for both PVQD-A and PVQD-S to evaluate their effectiveness across different speech units. Second, handcrafted acoustic features Jitter, Shimmer, and HNR were extracted using the Praat toolkit [34] through the Parselmouth interface [33], [35].

D. Model Training

All models were trained for 100 epochs using the AdamW optimizer (learning rate = 0.002, weight decay = 1e–5). Each model was trained and tested independently on the PVQD-A and PVQD-S subsets, and the performance on the two subsets is shown separately.

E. Evaluation Criteria

Model performance was evaluated using two metrics: Root Mean Squared Error (RMSE) and Pearson Correlation Coefficient (PCC). RMSE quantifies the average squared difference between the model output and the ground-truth perceptual score, with lower values indicating better performance. PCC measures the linear correlation between predicted and actual scores, with values closer to 1.0 indicating greater consistency with human ratings.

To reflect both fine-grained prediction accuracy and clinical relevance, we used both utterance-level and patient-level assessments. For patient-level scoring, we average the predictions from all utterances of a single speaker to arrive at a final perceptual rating. This approach mimics real-world clinical scenarios, where judgments are typically based on multiple utterances. This dual framework allows for a more comprehensive evaluation of the model's prediction performance, and the results are closely aligned with clinical practice.

TABLE V
PERFORMANCE COMPARISON OF VOQANET AND BASELINE MODELS.

Model	Feature	PVQD-A		PVQD-S	
		RMSE ↓	PCC ↑	RMSE ↓	PCC ↑
CAPE-V Prediction					
Lin [16]	HF	15.22	0.69	-	-
Lin [16]	MFCC+MS	14.76	0.64	-	-
Lin [16]	Waveform	17.09	0.48	-	-
Lin [16]	W2V2 (Last)	17.09	0.55	-	-
Lin [16]	HuBERT (Last)	18.14	0.49	-	-
Lin [16]	WavLM (Last)	20.23	0.33	-	-
Lin [16]	Whisper (Last)	15.67	0.62	-	-
VOQANet	Whisper (Last)	10.514	0.803	10.546	0.843
VOQANet	WavLM (Last)	**9.955**	**0.838**	**9.756**	**0.847**
GRBAS Prediction					
VOQANet	Whisper (Last)	**0.380**	0.793	0.352	0.819
VOQANet	WavLM (Last)	**0.380**	**0.795**	**0.322**	**0.833**

F. Noise Robustness Setup

We introduced a noise-augmented version of the PVQD dataset to evaluate the robustness of the model under adverse acoustic conditions. Table IV summarizes the noise types and signal-to-noise ratio (SNR) levels used for training, seen test scenarios, and unseen test scenarios. Each set contains the original clean utterances. The training set was augmented with four noise types: white noise, pink noise (colored), cafeteria babble, and cocktail party (background), with SNRs of -5 dB, 0 dB, 5 dB, and 10 dB, respectively. This combination includes both stationary and non-stationary interference commonly used in speech robustness studies. To evaluate generalization, we introduced unseen noise types into the evaluation process: brown noise (low-frequency), baby cry, and laughter, the latter two representing non-speech vocalization of emotional expression. These noise types were chosen to reflect real-world clinical and telehealth environments, where voice assessment may be performed in varied background settings. Fig. 2 illustrates the waveforms and spectrograms of the PVQD-A and PVQD-S subsets under different noise types. The selected SNR levels (–5 dB to 10 dB) cover a wide range of environments from challenging to moderate noise, aligning with previous research in the field of ASR and speech enhancement, ensuring that the robustness evaluation reflects the real-world clinical environment.

This experimental design ensures that the models are evaluated across a wide range of input conditions, including clean, noisy, vowel, and sentence-based speech, to comprehensively evaluate their generalizability, interpretability, and robustness.

V. RESULTS AND DISCUSSION

A. Comparison of VOQANet with the Baseline

Table V shows the utterance-level evaluation results of VOQANet with different features (including the embeddings of the last layer of WavLM or Whisper) on CAPE-V and GRBAS prediction. We compare VOQANet with the methods of Lin et al. [16], which include traditional or neural regressors, using handcrafted features (HF), traditional signal transformations (mel-frequency cepstral coefficient and modulation spectrum

TABLE VI
PERFORMANCE COMPARISON OF VOQANET MODELS USING DIFFERENT SFMS AND REPRESENTATIONS.

Feature	SFM	PVQD-A		PVQD-S	
		RMSE ↓	PCC ↑	RMSE ↓	PCC ↑
CAPE-V Prediction					
Last	Whisper	10.514	0.803	10.546	0.843
	WavLM	9.955	0.838	9.756	0.847
WS	Whisper	**9.770**	0.854	9.933	0.863
	WavLM	9.891	**0.865**	**9.209**	**0.870**
GRBAS Prediction					
Last	Whisper	0.380	0.793	0.352	0.819
	WavLM	0.380	0.795	0.322	0.833
WS	Whisper	0.370	0.800	0.354	0.822
	WavLM	**0.369**	**0.809**	**0.318**	**0.845**

(MFCC+MS)), raw waveform, and features extracted by pre-trained models (e.g., Wav2Vec2 (W2V2), HuBERT, WavLM, and Whisper). Lin et al. only evaluated their methods on CAPE-V prediction on the PVQD-A subset. Their results show that traditional features (including HF and MFCC+MS) outperform SFM features. SFM features performed worse than expected, except for the representation of the last layer of Whisper (Whisper (Last)). The lowest RMSE is 14.76 and the highest PCC is 0.69. However, their best results are significantly worse than those of VOQANet, suggesting that these traditional features are not sufficient to model the complex acoustic patterns relevant to perceptual voice assessment. Moreover, their models were trained using data augmentation with noise injection, whereas our VOQANet models in this experiment was trained on the original PVQD-A dataset without data augmentation. Despite this difference, VOQANet still performs significantly better than these baselines. On PVQD-A, VOQANet reduces RMSE to 10.514 (Whisper (Last)) and 9.955 (WavLM (WS)), and the latter is more than 30% lower than the best baseline. PCC is also improved substantially, exceeding 0.8 for both embeddings.

VOQANet with WavLM (Last) outperforms VOQANet with Whisper (Last) on CAPE-V prediction on the PVDQ-S subset, suggesting that WavLM is particularly well suited to capture prosodic and phonetic nuances in continuous speech. Similar improvements were observed for GRBAS prediction on both sunsets. While the numerical gains in GRBAS prediction may appear more subtle in absolute terms compared to CAPE-V prediction due to the narrower range of GRBAS (0-3), these gains are meaningful from a clinical perspective.

B. Comparison of SFM Types and Representations

In this study, WavLM and Whisper are used as backbone SFMs because of their excellent performance in speech quality assessment tasks [30]. Another study also highlighted the growing role of self-supervised learning models in non-intrusive speech assessment [37]. To further analyze the effectiveness of SFM-based representations, we compare two adaptation strategies: last-layer features (Last) and weighted-sum aggregation (WS) in Whisper and WavLM. This experiment adopts utterance-level evaluation. As shown in Table VI, the weighted-sum representation consistently outperforms the last-layer representation in all configurations. For example, for GRBAS prediction on the PVQD-A subset, WavLM WS yields lower RMSE (0.369 vs. 0.380) and higher PCC (0.809 vs. 0.795) than its last-layer counterpart. This is consistent with previous findings in the speech processing literature [17], [18], which show that intermediate transformer layers encode diverse and complementary representations, some of which better capture prosodic and phonatory features critical to the perception of voice quality.

From Table VI, we can also see that WavLM outperforms Whisper in CAPE-V and GRBAS predictions on both the PVQD-A and PVQD-S subsets. In particular, on the PVQD-S subset, WavLM achieves lower RMSE (9.209 vs. 9.933) and higher PCC (0.870 vs. 0.863) than Whisper when using WS features for CAPE-V prediction. The performance gap is even more pronounced in GRBAS prediction, with WavLM WS achieving a PCC of 0.845 on PVQD-S, compared to 0.822 for Whisper WS. These findings confirm that WavLM's denoising pretraining and fine-grained acoustic modeling have advantages in disordered voice settings, especially when handling longer or more naturalistic speech. Overall, this experiment confirms that WavLM (WS) provides the most informative and powerful representation and will be used in subsequent experiments.

C. Comparison between VOQANet and VOQANet+

While SFM (e.g., WavLM) embeddings are effective in capturing high-level acoustic and prosodic information, they are not explicitly optimized for clinically salient voice quality traits. To address this limitation, VOQANet+ incorporates handcrafted acoustic features (jitter, shimmer, and HNR, or JSH for short). These features have been shown to correlate with perceptual voice quality dimensions and are widely adopted in clinical voice analysis and speech processing systems due to their ability to reflect phonatory stability and noise characteristics [38]. Combined with WavLM embeddings, these handcrafted features provide interpretable low-level signal-based information that complements deep learning-based representations.

As shown in Table VII, VOQANet+ consistently achieves the lowest RMSE and highest PCC across all tasks and evaluation levels, outperforming VOQANet with SFM alone. The performance improvement is particularly prominent at the patient level, where predictions are aggregated over utterances from the same speaker to mimic real-world assessment. For example, in GRBAS prediction on PVQD-S, VOQANet+ improves the patient-level PCC from 0.867 to 0.874 and reduces RMSE from 0.297 to 0.289. These results indicate that adding handcrafted features helps to supplement low-level acoustic cues, such as irregularities in frequency or amplitude, that may not be adequately captured by SFM embeddings alone.

Fig. 3 visualizes the CAPE-V scores predicted by VOQANet versus the actual scores, while Fig. 4 does so for GRBAS. Fig. 5 visualizes the CAPE-V scores predicted by VOQANet+ versus the actual scores, while Fig. 6 does so for GRBAS. In each figure, the x-axis denotes the ground-truth scores assigned by expert raters, and the y-axis represents

TABLE VII
PERFORMANCE COMPARISON OF VOQANET AND VOQANET+.

Method	Feature	Utterance-Level				Patient-Level			
		PVQD-A		PVQD-S		PVQD-A		PVQD-S	
		RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑
CAPE-V Prediction									
VOQANet	Whisper (WS)	9.770	0.854	9.933	0.863	11.473	0.848	10.212	0.870
	WavLM (WS)	9.891	0.865	9.209	0.870	9.720	0.864	7.765	0.901
VOQANet+	Whisper (WS) + JSH	9.304	0.868	9.922	0.866	9.790	0.862	9.350	0.875
	WavLM (WS) + JSH	**8.594**	**0.877**	**8.720**	**0.883**	**9.042**	**0.878**	**7.356**	**0.908**
GRBAS Prediction									
VOQANet	Whisper (WS)	0.370	0.800	0.354	0.822	0.342	0.813	0.324	0.855
	WavLM (WS)	0.369	0.809	0.318	0.845	0.337	0.826	0.297	0.867
VOQANet+	Whisper (WS) + JSH	0.367	0.822	0.344	0.835	0.349	0.828	0.343	0.858
	WavLM (WS) + JSH	**0.364**	**0.830**	**0.307**	**0.854**	**0.332**	**0.839**	**0.289**	**0.874**

TABLE VIII
ROBUSTNESS EVALUATION OF VOQANET AND VOQANET+ UNDER SEEN AND UNSEEN NOISY CONDITIONS.

Evaluation Type	Method	Seen				Unseen			
		PVQD-A		PVQD-S		PVQD-A		PVQD-S	
		RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑
CAPE-V Prediction									
Utterance-Level	VOQANet (WavLM (WS))	9.981	0.832	10.279	0.843	10.555	0.807	10.627	0.809
	VOQANet+ (WavLM (WS) + JSH)	9.453	0.844	9.318	0.852	10.393	0.809	10.579	0.811
Patient-Level	VOQANet (WavLM (WS))	9.948	0.844	8.429	0.878	11.326	0.828	9.066	0.868
	VOQANet+ (WavLM (WS) + JSH)	9.381	0.852	8.265	0.888	10.096	0.832	8.625	0.881
GRBAS Prediction									
Utterance-Level	VOQANet (WavLM (WS))	0.365	0.778	0.375	0.836	0.381	0.774	0.402	0.807
	VOQANet+ (WavLM (WS) + JSH	0.362	0.785	0.320	0.841	0.373	0.779	0.326	0.836
Patient-Level	VOQANet (WavLM (WS))	0.365	0.817	0.299	0.858	0.452	0.802	0.336	0.832
	VOQANet+ (WavLM (WS) + JSH)	0.386	0.827	0.293	0.865	0.348	0.819	0.313	0.855

the model's predicted scores. The top row (e.g., Fig. 3(a–b)) corresponds to utterance-level predictions, and the bottom row (e.g., Fig. 3(c–d)) shows patient-level predictions obtained by averaging the prediction scores of different utterances of the same speaker. Each point represents an utterance (or speaker), the red line shows the regression fit, and the shaded area is the 95% confidence interval. VOQANet+ shows tighter clustering near the diagonal, especially in the patient-level plots (e.g., Fig. 3(d) vs. Fig. 5(d) and Fig. 4(d) vs. Fig. 6(d)), which indicates stronger prediction consistency and lower variance. These findings further demonstrate that handcrafted features can enhance the model's ability to estimate perceptual ratings, especially in cases of higher severity, where expert judgments tend to be more variable. In summary, by combining black-box deep learning-based representations with interpretable acoustic features, VOQANet+ improves the clinical relevance, trustworthiness, and robustness of voice quality prediction.

D. Utterance-Level Evaluation vs. Patient-Level Evaluation

In clinical settings, auditory perceptual judgments are often made by listening to multiple utterances from a single speaker. To reflect this practice, we evaluate model performance at both the utterance level and the patient level by averaging the predictions for all utterances from the same speaker. As shown in Table VII and visualized in scatter plots (Figs. 3, 4, 5, 6), patient-level evaluation achieves higher PCC and lower RMSE than utterance-level evaluation. For example, when using VOQANet+ (with WavLM (WS) + JSH features) for CAPE-V prediction on PVQD-S, PCC improved from 0.883 at the utterance level to 0.908 at the patient level, and RMSE decreased from 8.720 to 7.356.

To further investigate the contribution of different speech types, we compare the results of VOQANet+ (with WavLM (WS) + JSH features) on the PVQD-A (vowel-based) and PVQD-S (sentence-based) subsets. As shown in Fig. 7, sentence-based predictions on PVQD-S consistently outperform vowel-based predictions on PVQD-A, especially at the patient level. This supports the hypothesis that longer continuous speech utterances provide richer prosodic and articulatory information, allowing the model to make more accurate and stable predictions. These results suggest that while sustained vowels are still clinically useful, sentence-level inputs can provide more contextual acoustic dynamics, such as stress, pitch variation, and connected phonation, which are highly informative for complex perceptual dimensions such as strain or roughness.

Moreover, the advantage of sentence-based predictions is further amplified when handcrafted features are incorporated into the model (see the comparison of VOQANet and VOQANet+ in Table VII). This suggests a synergistic effect between rich SFM-derived representations and domain-informed acoustic markers, especially in the capture of the characteristics of voice disorders. These findings also have important implications for robustness. As we will explore in the next subsection, sentence-level input enables VOQANet+

to maintain stronger performance under seen and unseen noisy conditions. This resilience further highlights the clinical value of incorporating continuous speech into the automated voice assessment framework.

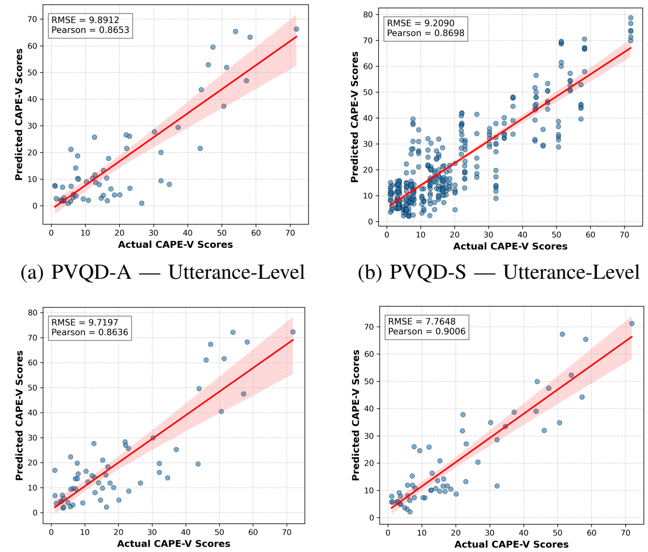
E. Robustness to Seen and Unseen Noise

To examine the generalizability of the models under adverse acoustic conditions, we tested the robustness of VOQANet and VOQANet+ under various noisy conditions on the noise-augmented PVQD dataset. To improve the robustness of dysphonic voice detection, recent approaches have focused on generating acoustic feature embeddings that are sensitive to vocal quality and robust across different corpora [39]. This emphasis on real-world acoustic variability and noise resilience stems from recent research in voice disorder modeling, which highlights the importance of domain robustness for clinical applicability [40]. As shown in Table VIII, both models were evaluated for seen (i.e., white, pink, babble, and cocktail party noise used during training) and unseen (i.e., baby cry, laughter, and brown noise not included in training) noise types at multiple SNR levels from -5 dB to 10 dB. On both CAPE-V prediction and GRBAS prediction tasks, VOQANet+ consistently outperforms VOQANet under noisy conditions, demonstrating greater resilience to both seen and unseen disturbances. For example, for patient-level GRBAS prediction on PVQD-S under unseen noise, VOQANet+ achieves a PCC of 0.855, compared to 0.832 for VOQANet, while reducing RMSE from 0.336 to 0.313. Similarly, for utterance-level CAPE-V prediction on PVQD-S, VOQANet+ improves PCC from 0.809 to 0.811 and reduces RMSE from 10.627 to 10.579. While numerically modest, these improvements are meaningful in noisy clinical or telehealth scenarios, where the reliability of scoring under real-world background conditions is critical.

This improvement highlights that handcrafted features (JSH), rooted in perturbation measures and periodicity detection, provide complementary information that is robust to noise and remains stable even when deep SFM embeddings degrade. By capturing signal-level voice irregularities that are less susceptible to spectral masking or background interference, VOQANet+ benefits from richer and more noise-resistant representations. Moreover, VOQANet+ shows less performance degradation from seen to unseen noise than VOQANet, especially in GRBAS prediction. This highlights the advantage of combining interpretable clinically meaningful features with learned embeddings to form a more adaptable model that maintains prediction quality even when faces with out-of-distribution acoustic environments. The results confirm that VOQANet+ provides a more robust solution for deployment in noisy real-world environments.

VI. Conclusions

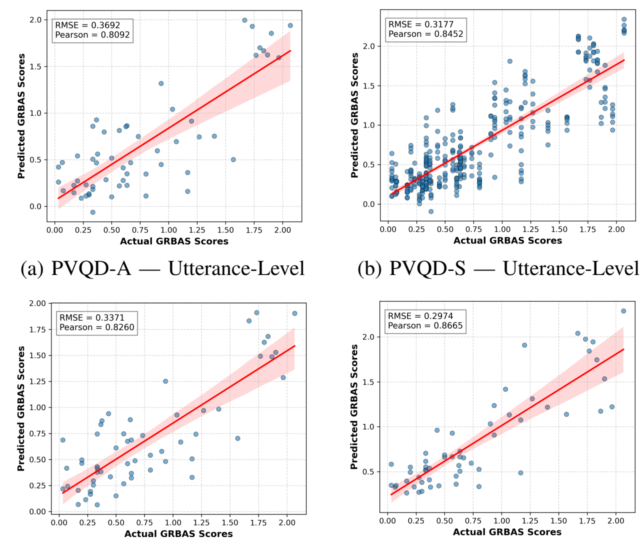
This study introduces VOQANet, a deep learning-based framework with an attention mechanism for automated perceptual voice quality assessment. VOQANet uses SFM to extract representations from speech input, effectively capturing high-level acoustic and prosodic information from raw waveforms.



(a) PVQD-A — Utterance-Level (b) PVQD-S — Utterance-Level

(c) PVQD-A — Patient-Level (d) PVQD-S — Patient-Level

Fig. 3. Scatter plots of CAPE-V scores predicted by VOQANet with WavLM (WS) features versus actual scores. The top row shows utterance-level predictions on (a) PVQD-A and (b) PVQD-S, and the bottom row shows patient-level predictions on (c) PVQD-A and (d) PVQD-S.



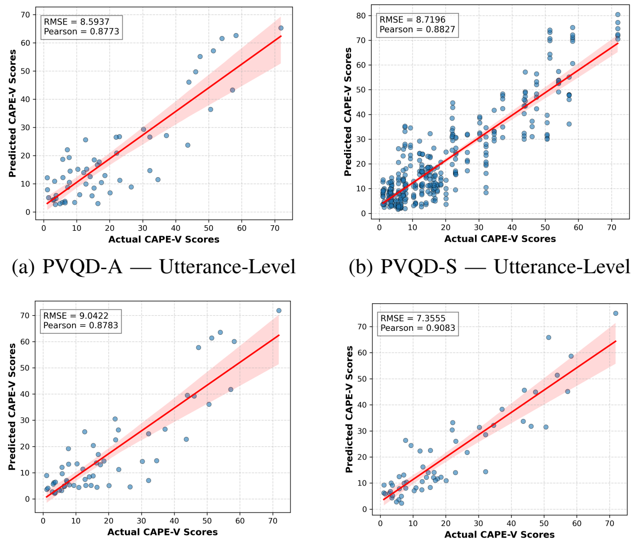
(a) PVQD-A — Utterance-Level (b) PVQD-S — Utterance-Level

(c) PVQD-A — Patient-Level (d) PVQD-S — Patient-Level

Fig. 4. Scatter plots of GRBAS scores predicted by VOQANet with WavLM (WS) features versus actual scores. The top row shows utterance-level predictions on (a) PVQD-A and (b) PVQD-S, and the bottom row shows patient-level predictions on (c) PVQD-A and (d) PVQD-S.

It achieves strong predictive performance on both CAPE-V and GRBAS rating scales, demonstrating the utility of SFM embeddings in modeling perceptual voice characteristics.

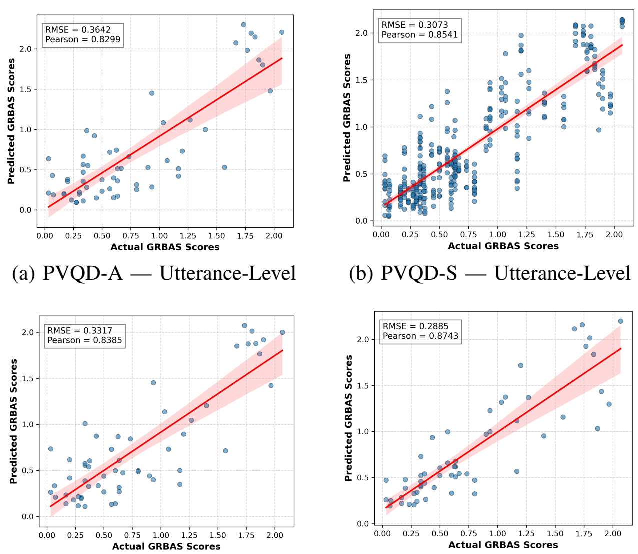
To further enhance clinical relevance and model robustness, we propose VOQANet+, which incorporates handcrafted acoustic features (jitter, shimmer, and HNR) and SFM embeddings. Through a comprehensive evaluation on the PVQD dataset, we identify several key findings: (1) WavLM outperforms Whisper in voice quality prediction, especially when



(a) PVQD-A — Utterance-Level (b) PVQD-S — Utterance-Level

(c) PVQD-A — Patient-Level (d) PVQD-S — Patient-Level

Fig. 5. Scatter plots of CAPE-V scores predicted by VOQANet+ with WavLM (WS) features and prosodic features (Jitter, Shimmer, and HNR). The top row shows utterance-level predictions on (a) PVQD-A and (b) PVQD-S, and the bottom row shows patient-level predictions on (c) PVQD-A and (d) PVQD-S.

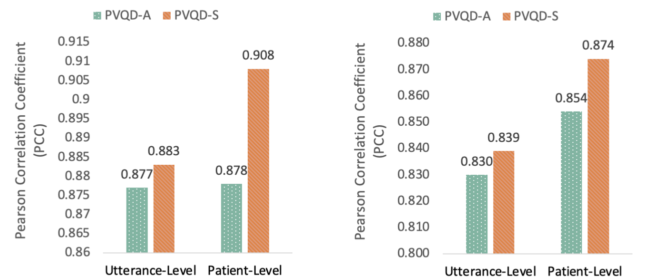


(a) PVQD-A — Utterance-Level (b) PVQD-S — Utterance-Level

(c) PVQD-A — Patient-Level (d) PVQD-S — Patient-Level

Fig. 6. Scatter plots of GRBAS scores predicted by VOQANet+ with WavLM (WS) features and prosodic features (Jitter, Shimmer, and HNR). The top row shows utterance-level predictions on (a) PVQD-A and (b) PVQD-S, and the bottom row shows patient-level predictions on (c) PVQD-A and (d) PVQD-S.

using weighted-sum (WS) embeddings that aggregate across multiple layers; (2) VOQANet+ consistently improves performance over VOQANet in both utterance-level and patient-level evaluations, reflecting its better stability and alignment with clinical assessments; and (3) VOQANet+ demonstrates greater resilience under seen and unseen noise conditions, showing greater robustness and generalizability in acoustically challenging environments.



(a) CAPE-V Prediction (b) GRBAS Prediction

Fig. 7. Comparison of the performance of VOQANet+ (with WavLM (WS) + JSH features) for vowel-based predictions on PVQD-A and sentence-based predictions on PVQD-S.

By combining the strengths of pre-trained SFM representations and clinically interpretable acoustic features, VOQANet+ provides a robust and interpretable foundation for real-world voice quality assessment applications. Future research directions may include exploring multi-task learning to jointly predict individual CAPE-V (or GRBAS) dimensions or perceptual subscales, as well as cross-lingual generalization and domain adaptation to enable broader deployment across clinical settings and languages.

References

[1] S.-S. Wang, C.-T. Wang, C.-C. Lai, and Y. Tsao, “Continuous Speech for Improved Learning Pathological Voice Disorders,” *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 3, pp. 25-33, 2022.

[2] W. Ariyanti, T. Hussain, J.-C. Wang, C.-T. Wang, S.-H. Fang, and Y. Tsao, “Ensemble and Multimodal Learning for Pathological Voice Classification,” *IEEE Sensors Letters*, pp. 1-4, 2021.

[3] P. R. Walden, “Perceptual voice qualities database (PVQD): database characteristics,” *Journal of Voice*, pp. 875.e15-875.e23, 2022.

[4] J. Kreiman, B. Gerratt, M. Ito, “When and why listeners disagree in voice quality assessment tasks,” *The Journal of the Acoustical Society of America*, pp. 2354-64, vol. 122, 2007.

[5] G. B. Kempster, B. R. Gerrat, K. V. Abbott, J. B. Kraemer, and R. E. Hillman, “Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol,” *American Journal of Speech-Language Pathology*, pp. 124-132, 2009.

[6] M. Hirano and K. R. McCormick, “Clinical examination of voice,” *The Journal of the Acoustical Society of America*, pp. 1273-1273, 1986.

[7] A. Sasou, “Automatic identification of pathological voice quality based on the GRBAS categorization,” in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1243-1247, 2017.

[8] T. Pommée, D. Mbagira, and D. Morsomme, “French-language adaptation of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V),” *Journal of Voice*, 2024. doi:10.1016/j.jvoice.2024.03.011.

[9] E. Ertan-Schlüter, H. Okur, M. Yildiz, and H. Sözen, “The Turkish version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V): A reliability and validity study,” *Journal of Voice*, vol. 34, no. 6, pp. 965.e13–965.e22, 2020. doi:10.1016/j.jvoice.2019.05.001.

[10] L. M. T. Jesus, A. Barney, P. S. Couto, H. Vilarinho, and A. Correia, “Voice quality evaluation using CAPE-V and GRBAS in European Portuguese,” in *Proc. MAVEBA*, pp. 61-64, 2009.

[11] K. Kondo, M. Mizuta, Y. Kawai, et al., “Development and validation of the Japanese version of the Consensus Auditory-Perceptual Evaluation of Voice,” *Journal of Speech, Language, and Hearing Research*, vol. 64, pp. 4754–4761, 2021. doi:10.1044/2021_JSLHR-21-00269.

[12] M. Hirano, *Clinical Examination of Voice*. Vienna, Austria: Springer-Verlag, 1981.

[13] T. McAllister, C. Nightingale, G. Moya-Gale, A. Kawamura, and L. Ramig, “Crowdsourced perceptual ratings of voice quality in people with Parkinson’s disease before and after intensive voice and articulation therapies: Secondary outcome of a randomized controlled trial,” *Journal of Speech, Language, and Hearing Research*, vol. 66, pp. 1–22, Apr. 2023.

[14] S. Xie, N. Yan, P. Yu, M. L. Ng, L. Wang, and Z. Ji, “Deep Neural Networks for Voice Quality Assessment based on the GRBAS Scale,” in *Interspeech*, pp. 2656–2660, 2016.

[15] J. M. Miramont, M. A. Colominas, and G. Schlotthauer, “Emulating perceptual evaluation of voice using scattering transform based features,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1892-1901, 2002.

[16] Y. Lin, W.-H. Tseng, L.-C. Chen, C.-T. Tan, and Y. Tsao, “Lightly weighted automatic audio parameter extraction for the quality assessment of consensus auditory-perceptual evaluation of voice,” *in Proc. ICCE*, pp. 1-6, 2023.

[17] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, pp. 1505-1518, 2022.

[18] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Slakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 3451-3460, 2021.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” *in Proc. International Conference on Machine Learning*, 2022.

[20] J. P. Teixeira, C. Oliveira, and C. Lopes, “Vocal acoustic analysis – jitter, shimmer and HNR parameters,” *Procedia Technology*, vol.9, pp. 1112-1122, 2013.

[21] J. Kreiman and B. R. Gerratt, “Perceptual assessment of voice quality: Past, present, and future,” *Perspectives on Voice and Voice Disorders*, pp. 62-67, vol. 20, 2010.

[22] K. F. Nagle, “Clinical use of the CAPE-V scales: Agreement, reliability, and notes on voice quality,” *Journal of Voice*, 2022.

[23] Y. Liu, X. Zhang, and L. Wang, “Adapting Whisper for low-data-resource speaker verification,” *Speech Communication*, vol. 150, pp. 123–135, 2024.

[24] G. Elbanna, Z. Mostaani, and M. Magimai-Doss, “Predicting heart activity from speech using data-driven and knowledge-based features,” in *Proc. Interspeech*, 2024.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *ICASSP*, pp. 5206-5210, 2015.

[26] S. Liu, et al., “Audio self-supervised learning: A survey,” *Pattern*, 2022.

[27] L. V. Staden and H. Kamper, “A Comparison of Self-Supervised Speech Representations As Input Features For Unsupervised Acoustic Word Embeddings,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 927-934, 2020.

[28] G. Elbanna, A. Biryukov, N. Scheidwasser-Clow, L. Orlandic, P. Mainar, M. Kegler, P. Beckmann, and M. Cernak, “Hybrid handcrafted and learnable audio representation for analysis of speech under cognitive and physical load,” *in Proc. Interspeech*, 2022.

[29] R. B. Fujiki and S. L. Thibeault, “The relationship between auditory-perceptual rating scales and objective voice measures in children with voice disorders,” *American Journal of Speech-Language Pathology*, vol. 30, no. 1, pp. 228–238, 2021. doi:10.1044/2020_AJSLP-20-00188.

[30] H.-T. Chiang, S.-W. Fu, H.-M. Wang, Y. Tsao, and J. H. L. Hansen, “Multi-objective non-intrusive hearing-aid speech assessment model,” *J. Acoust. Soc. Am.*, vol. 156, no. 5, pp. 3574-3587, Nov. 2024.

[31] K.-S. Lu, A. Ortega, D. Mukherjee, and Y. Chen, “Perceptually Inspired Weighted MSE Optimization Using Irregularity-Aware Graph Fourier Transform,” in *ICIP*, pp. 3384-3388, 2020.

[32] T. Kojima, S. Fujimura, K. Hasebe, Y. Okanoue, O. Shuya, R. Yuki, K. Shoji, R. Hori, Y. Kishimoto, and K. Omori, “Objective assessment of pathological voice using artificial intelligence based on the GRBAS scale,” *Journal of Voice*, pp. 561-566, 2024.

[33] D. R. Feinberg, “Parselmouth Praat Scripts in Python,” 2018.

[34] P. Boersma and V. van Heuven, “Speak and unspeak with praat,” *Glot Int.*, vol. 5, no. 9-10, pp. 341-347, 2001.

[35] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing parselmouth: A python interface to praat,” *J. Phon.*, vol. 71, pp. 1-15, 2018.

[36] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” *CoRR*, 2014.

[37] R.E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, “Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54-70, 2023.

[38] E. Keller, “The Analysis of Voice Quality in Speech Processing,” *Springer-Verlag*, pp. 54-73, 2005.

[39] J. Zhang, J. Liss, S. Jayasuriya, and V. Berisha, “Robust vocal quality feature embeddings for dysphonic voice detection,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1–12, 2022.

[40] 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, Oct. 2023.