INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
RAICHUR, INDIA, 584101

PHONATION CLASSIFICATION USING SELF SUPERVISED MODELS

Minor Project Report
VI Semester

Submitted By:
PAVAN KUMAR CS22B1042
MANVENDRA SINGH CS22B1054

Under the Guidance of
Dr. KIRAN REDDY MITTAPALLE
Department of Computer Science and Engineering

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Background

Three different forms of voiced sounds—breathy, neutral, and pressed—are produced by phonation, which uses vocal fold vibrations and sub-glottal pressure. These are essential for speaking and singing. These kinds help with speech pathology and emotion analysis, but they are difficult to categorize because of minute acoustic variations. While they rely on manual feature extraction, traditional approaches such as MFCCs or TQWT with entropy features and neural networks had moderate results (e.g., 81.67% accuracy for spoken speech). A strong substitute is provided by self-supervised models (HuBERT, Wav2Vec 2.0), which use unlabeled audio data to learn high-dimensional, contextual embeddings. This effort surpasses conventional methods by using these models in conjunction with classifiers (SVM, MLP, Random Forest) to increase the accuracy of phonation type categorization for speaking voices.

1.2. Objectives

Using self-supervised models (HuBERT, Wav2Vec 2.0 Base, Wav2Vec 2.0 Large) for feature extraction and classifiers (SVM, MLP, Random Forest), a system for classifying phonation types (breathy, neutral, pressed) in speaking voice signals is being developed with the goal of achieving higher accuracy than conventional methods and improving applications in speech analysis and pathology.

1.3. Problem statement

Due to subtle acoustic differences and the limitations of traditional feature extraction methods like MFCCs and TQWT, which rely on manual engineering and lack contextual robustness, it is difficult to classify phonation types (breathy, neutral, and pressed) in speaking voice signals. The classification accuracy is moderate (e.g., 81.67% for speaking voice). For applications in speech analysis and pathology, a data-driven strategy that uses sophisticated classifiers and self-supervised models to extract rich, contextual characteristics is required in order to increase accuracy and generalizability.

2. LITERATURE REVIEW

2.1. Existing Work on Phonation Type Classification

There are three different types of phonation—breathy, neutral, and pressed—that are essential for speaking and singing. Phonation is the act of creating voiced sounds by means of vocal fold vibrations triggered by sub-glottal pressure. In order to help applications like emotion analysis, singing style identification, occupational voice care, and speech pathology, numerous studies have investigated the automatic classification of different phonation kinds. Initially, glottal inverse filtering (GIF) was used to estimate the glottal source waveform from speech signals in order to extract features for classification [3, 5, 8, 12, 23]. In order to enhance performance, especially for singing voices, more recent research has turned to characteristics that are directly taken from acoustic signals, such as harmonic amplitudes, formant frequencies, and cepstral coefficients [5, 8, 9, 16].The application of the Tunable Q-factor Wavelet Transform (TQWT) for feature extraction is a noteworthy development, since it has demonstrated exceptional performance in phonation type classification in both speech and singing [15].

2.2. Methods, Techniques, Technologies, and Innovations

- Glotal source features: Conventional techniques frequently use GIF to estimate the glottal flow waveform, extracting frequency-domain features like H1-H2 and Harmonic Richness Factor (HRF) in addition to time-domain features like Amplitude Quotient (AQ), Normalized Amplitude Quotient (NAQ), Closed Quotient (CIQ), and Open Quotient (OQ1) [1, 3]. These characteristics record differences in the forms of glottal pulses, which vary depending on the kind of phonation—asymmetric for pushed, smooth for breathy [3]. But when it comes to singing voices, GIF has trouble with source-filter coupling, which lowers classification accuracy [5, 9].

- Auditory Signal characteristics: Researchers have looked into characteristics that are directly derived from auditory signals in order to overcome the limits of GIF. In singing voice classification, studies have used harmonic amplitudes, formant frequencies, cepstral peak prominence, and harmonic-to-noise ratios, which perform better than glottal characteristics [8]. Although Mel-Frequency Cepstral Coefficients (MFCCs) have been applied extensively to speech, their efficacy in singing has been shown to be restricted [5, 8]. In order to enhance speech and singing categorization, Single-Frequency Filtering-based Cepstral Coefficients (SFFCCs) were suggested; they outperformed MFCCs and glottal characteristics [5, 8].

- Wavelet-Based Methods: One notable advancement is the application of wavelet transformations, especially TQWT, which divides speech signals into sub-bands according to their oscillatory characteristics rather than their frequency [9, 17, 22]. One study classified phonation types using a Feed Forward Neural Network (FFNN) after using TQWT to extract Shannon wavelet entropy information from sub-bands [9]. This method outperformed six state-of-the-art features, such as MFCCs and SFFCCs, with 91% accuracy for singing and 82% accuracy for speaking voices [9]. A two-layer wavelet scattering network was used in another investigation to extract characteristics, and it performed better for classifying singing voices [14].

2.3. Drawbacks and Limitations

- Glottal Source Features: The source-filter coupling causes GIF-based features to perform poorly in singing voices, which restricts their use [5, 8, 9].
- Acoustic Features: Although MFCCs and SFFCCs enhance categorization, they frequently miss minute oscillatory variations among phonation types, resulting in incorrect classifications, especially between neutral and breathy voices [5, 8, 14].
- TQWT-Based Methods: Despite its effectiveness, TQWT depends on domain-specific expertise for feature extraction and necessitates manual Q-factor tweaking, which may not generalize across a variety of datasets [15]. For real-time applications, wavelet decomposition's processing complexity may potentially be a hindrance.
- Common Problems: Current features may not adequately reflect the continuum of laryngeal changes, as most studies find confusion between breathy and neutral sounds, and occasionally between neutral and pressured voices.

3. PROPOSED WORK AND METHODOLOGY

3.1. Proposed Work

Using self-supervised models for feature extraction and machine learning classifiers, the proposed study seeks to create an automated system for categorizing spoken voice signals into three phonation types: breathy, neutral, and pressed. This methodology uses the contextual and resilient feature representations from self-supervised models (HuBERT, Wav2Vec 2.0 Base, and Wav2Vec 2.0 Large) to increase classification accuracy, in contrast to standard methods that rely on manually created features like MFCCs or TQWT. To distinguish between different phonation types, the collected features are input into classifiers such as Random Forest, Multilayer Perceptron (MLP), and Support Vector Machine (SVM). To guarantee strong performance, the system is assessed using stratified k-fold cross-validation. The objective is to outperform the previous benchmark accuracy of 81.67% for speaking voice categorization.

3.2. Methodology

3.2.1. Dataset Description

The dataset includes recordings of the eight Finnish vowels (/a/, /e/, /i/, /o/, /u/, /y/, /ae/, /ø/) in three different phonation types: breathy, neutral, and pressed. It comprises 792 isolated vowel samples. Eleven speakers, five of whom were men and six of whom were women, ages 18 to 48, produced the recordings. A total of 792 samples were produced by repeating each vowel three times for each phonation type (3 repetitions × 3 phonation types × 8 vowels × 11 speakers). To provide high-quality, noise-free audio data, the recordings were made at a sampling rate of 44.1 kHz in an anechoic environment. The phonation types in the collection are balanced, with 264 samples (792 ÷ 3) representing the breathy, neutral, and pressed types.This balance reduces bias toward any phonation type by guaranteeing that the classifiers are trained and assessed on an equal amount of samples per class. The recording settings in the anechoic chamber remove reverberation and background noise, producing steady and regulated acoustic data that can be used for accurate feature extraction and categorization.

3.2.2. Feature Extraction

After obtaining high-dimensional, contextual embeddings from the audio signals by feature extraction using self-supervised models, the features are validated and stored.

Self-Supervised Models

- HuBERT: Previously trained on extensive unlabeled audio data, HuBERT (Hidden-Unit BERT) is a transformer-based model. By learning strong contextual representations of speech, it forecasts clustered units from masked audio segments. It processes raw audio using a convolutional neural network (CNN) encoder, then models dependencies using a transformer encoder.
- Wav2Vec 2.0 Base: This model uses a transformer to capture contextual dependencies after a CNN processes raw audio waveforms. Compact representations are produced by pre-training it with a contrastive loss to separate real audio portions from distractions.
- Wav2Vec 2.0 Large: Richer feature representations are made possible by Wav2Vec 2.0 Large, a more potent version of Wav2Vec 2.0 Base with 24 transformer layers and a higher parameter count.

Audio Processing and Feature Extraction Pipeline

1. Audio Loading: To standardize the input for model processing, audio files from the breathy, neutral, and pushed phonation type directories are loaded using the librosa library at a sampling rate of 16 kHz.
2. Model Processing: The processor of the corresponding self-supervised model (such as HuBERT or Wav2Vec 2.0) is used to process the loaded audio. The model receives the raw waveform without any feature extraction done by hand.
3. Hidden State Extraction: A designated layer of the model (for example, layer 5 for HuBERT) is used to extract hidden states, or embeddings. For every audio sample, a fixed-length representation is obtained by calculating the mean of these properties throughout the time dimension.
4. Feature Validation: Verify the extracted features to make sure they make up a 1D array. In order to prevent invalid inputs, check for zeros or NaN values. Make sure the feature dimension corresponds to the model's anticipated size.
5. Label Assignment: Features and labels are stacked into arrays, and each sample is given a label (0: neutral, 1: breathy, 2: pushed).
6. Pickle File Storage With the format features/task_type/model_name_l, the extracted features (X) and associated labels (Y) are merged and saved to a pickle file. This guarantees that processed data is efficiently stored and retrieved for further classification.

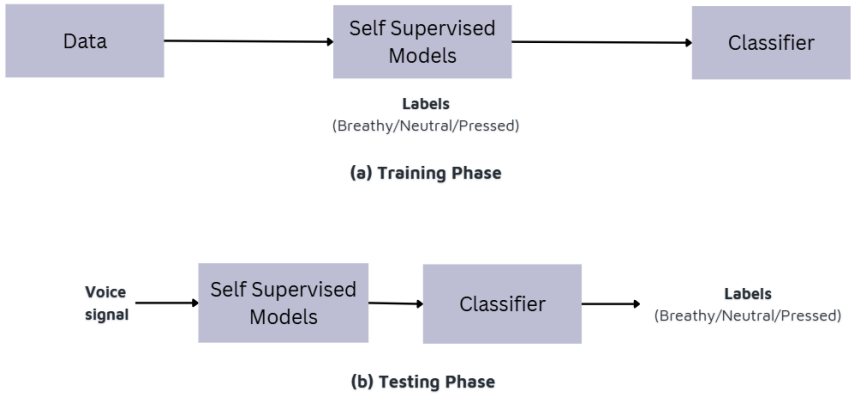
3.2.3. Classification

The classification pipeline is designed to evaluate the performance of the extracted features using multiple classifiers and robust validation techniques.

Classification Pipeline

1. Data Loading: The pickle file created during feature extraction is used to load the feature data and labels.
2. Stratified K-Fold Cross-Validation: Using stratified k-fold cross-validation, the dataset is divided into training and test sets (usually k=10). Class balance is maintained using stratification, which guarantees that each fold has the same percentage of phonation types as the initial dataset.
3. Feature Standardization: Features are standardized (e.g., scaled to zero mean and unit variance) to normalize the input for classifiers, enhancing convergence and performance.
4. Classifier Training and Prediction:
    - The Support Vector Machine (SVM) handles non-linear class borders by using a radial basis function (RBF) kernel with optimized hyperparameters (C=100, gamma='auto', probability=True).
    - Multilayer Perceptron (MLP): A neural network with several hidden layers used to model intricate feature connections is called a multilayer perceptron (MLP).
    - Random Forest: An ensemble technique that enhances robustness and decreases overfitting by combining several decision trees.
5. Performance Evaluation: Performance evaluation involves calculating accuracy for each fold and calculating the mean and standard deviation of accuracy for all folds to evaluate overall performance.
6. Storage of Results: For study and comparison, results are stored to a JSON file together with model specifics, hyperparameters, and performance metrics.

3.2.4. Proposed System



4. RESULTS

4.1. Experimental Setup

Using a 10-fold cross-validation (CV) technique, the dataset was split into ten equal sections at random for the tests. To guarantee that there was no overlap between training and testing samples in any fold, 90% of the data was set aside for training and the remaining 10% for testing. by using the same experimental setup as previous research.

4.2. Results

Using a variety of classifiers and self-supervised models for feature extraction, the suggested approach for identifying the three phonation types—pressed, breathy, and neutral—in speaking voice data produced noteworthy results. The maximum classification accuracy of **91.916% ± 2.84%** was attained by the HuBERT model, which used an SVM with a radial basis function (RBF) kernel and features taken from its fifth layer. The hyperparameters `C=100`, `gamma='auto'`, and `probability=True` were used to optimize the SVM, allowing for reliable handling of non-linear class borders. Additionally, using a Random Forest classifier on features from HuBERT layer 0 yielded a strong performance with an accuracy of **87.11% ± 3.98%**, using `n_estimators=300`, `max_leaf_nodes=75`, and `max_depth=10`. Similarly, the MLP classifier achieved an accuracy of **89.51% ± 4.47%** on features from HuBERT layer 5, with a hidden layer configuration of `[256, 64, 32]`, `learning_rate_init=0.001`, `early_stopping=True`, `max_iter=500`, and using the `adam` solver. This result notably exceeds the previous benchmark of **81.67%** for speaking voice classification using conventional TQWT-based features[15].

Table 4.1: Classification Performance of Different Classifiers with HuBERT Features

Classifier	Accuracy (%)	Std. Deviation (%)
Multilayer Perceptron (MLP)	89.506	4.47
Random Forest	87.107	3.97
SVM	91.916	2.84

4.3. Confusion Matrix Analysis

The HuBERT + SVM model's confusion matrix shows how well the model classified different phonation types (pressed, breathy, and neutral). Correct classifications are shown by diagonal values in the normalized matrix, while misclassifications are indicated by off-diagonal values.

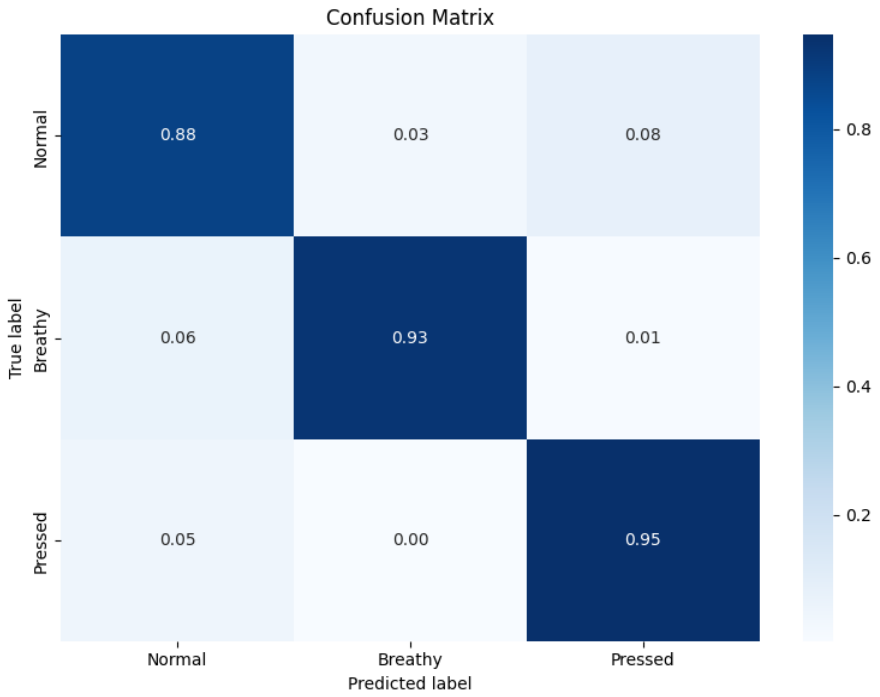


Figure 4.1: Confusion Matrix for HuBERT + SVM Model Across Phonation Types

4.4. Classification Accuracy across Layers

For phonation type classification (i.e., differentiating between two phonation kinds), the plot displays the SVM classification accuracy over layers of three self-supervised models: HuBERT, Wav2Vec 2.0 Base (Wav2Vec2-BASE), and Wav2Vec 2.0 Large (Wav2Vec2-LARGE). The y-axis shows the classification accuracy as a percentage (%), while the x-axis shows the layer number (ranging from 1 to 25).

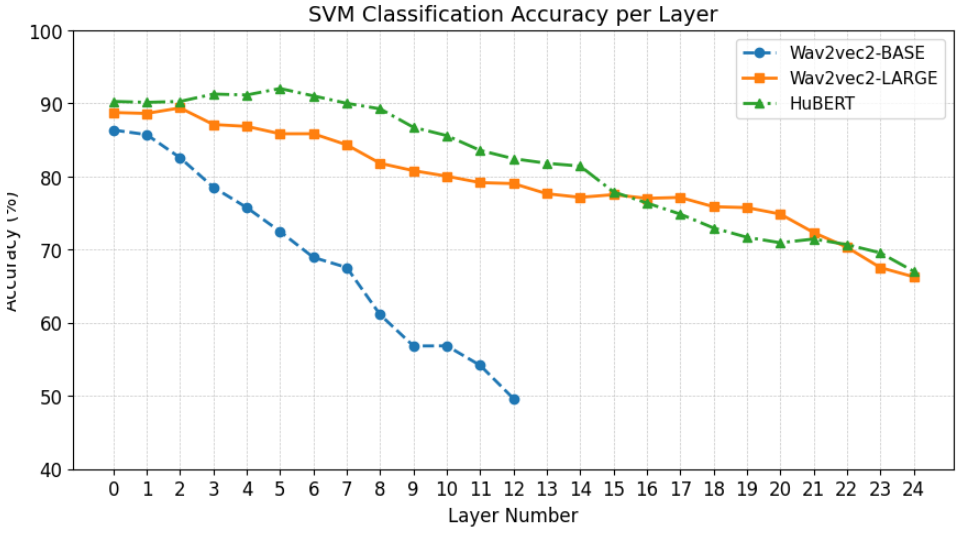


Figure 4.2: SVM Classification Accuracy Across Layers of Self-Supervised Models (HuBERT, Wav2Vec2-BASE, Wav2Vec2-LARGE)

5. CONCLUSION

This experiment effectively illustrated how self-supervised models can classify speech signals into three phonation types: breathy, neutral, and pushed. The system produced high-dimensional, contextual embeddings by utilizing the potent feature extraction capabilities of models such as HuBERT and Wav2Vec 2.0 (Base and Large), which considerably outperformed conventional hand-engineered features like MFCCs and TQWT-based representations. When combined with HuBERT embeddings, the SVM with RBF kernel outperformed the other classifiers by achieving the greatest accuracy of 91.92%, significantly exceeding the previous benchmark of 81.67%.

The findings highlight the promise of self-supervised learning for speech-related classification tasks, particularly in fields with a large amount of unlabeled audio and little labeled data. Additionally, this method creates new opportunities for useful applications in emotion analysis, speech therapy, and vocal health monitoring. To further improve this system's resilience and usefulness, future research can include real-time processing capabilities, assess performance across a range of languages and accents, and integrate more phonation types.

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27]

References

[1] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types." in *INTERSPEECH*, 2007, pp. 1410–1413.

[2] A. Akbari and M. K. Arjmandi, "An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features," *Biomedical Signal Processing and Control*, vol. 10, pp. 209–223, 2014.

[3] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *the Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[4] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[5] D. N. Gowda and M. Kurimo, "Analysis of breathy, modal and pressed phonation based on low frequency spectral density." in *INTERSPEECH*, 2013, pp. 3206–3210.

[6] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[7] M. Ito, "Politeness and voice quality-the alternative method to measure aspiration noise," in *Proc. Speech Prosody*, 2004, pp. 213–216.

[8] S. R. Kadiri and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (sffcc)." in *Interspeech*, 2018, pp. 441–445.

[9] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Analysis and classification of phonation types in speech and singing voice," *Speech Communication*, vol. 118, pp. 33–47, 2020.

[10] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform." in *Interspeech*, 2011, pp. 177–180.

[11] ——, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1170–1179, 2013.

[12] K. R. Mittapalle, H. Pohjalainen, P. Helkkula, K. Kaitue, M. Minkkinen, H. Tolppanen, T. Nieminen, and P. Alku, "Glottal flow characteristics in vowels produced by speakers with heart failure," *Speech Communication*, vol. 137, pp. 35–43, 2022.

[13] K. R. Mittapalle, M. K. Yagnavajjula, and P. Alku, "Classification of functional dysphonia using the tunable q wavelet transform," *Speech Communication*, vol. 155, p. 102989, 2023.

[14] K. R. Mittapalle and P. Alku, "Classification of phonation types in singing voice using wavelet scattering network-based features," *JASA Express Letters*, vol. 4, no. 6, 2024.

[15] ——, "Tunable q-factor wavelet transform-based features in the classification of phonation types in the singing and speaking voice," *Journal of Voice*, 2024.

[16] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing," *Journal of New Music Research*, vol. 42, no. 2, pp. 171–186, 2013.

[17] G. R. S. Reddy and R. Rao, "Oscillatory-plus-transient signal decomposition using tqwt and mca," *Journal of Electronic Science and Technology*, vol. 17, no. 2, pp. 135–151, 2019.

[18] M. K. Reddy, P. Helkkula, Y. M. Keerthana, K. Kaitue, M. Minkkinen, H. Tolppanen, T. Nieminen, and P. Alku, "The automatic detection of heart failure using speech signals," *Computer Speech & Language*, vol. 69, p. 101205, 2021.

[19] M. K. Reddy, Y. M. Keerthana, and P. Alku, "End-to-end pathological speech detection using wavelet scattering network," *IEEE Signal Processing Letters*, vol. 29, pp. 1863–1867, 2022.

[20] J.-L. Rouas and L. Ioannidis, "Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings," in *interspeech*, vol. 2016, 2016, pp. 150–154.

[21] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul, and H. Apaydin, "A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform," *Applied Soft Computing*, vol. 74, pp. 255–263, 2019.

[22] I. W. Selesnick, "Wavelet transform with tunable q-factor," *IEEE transactions on signal processing*, vol. 59, no. 8, pp. 3560–3575, 2011.

[23] J. Sundberg, "Objective characterization of phonation type using amplitude of flow glottogram pulse and of voice source fundamental," *Journal of Voice*, vol. 36, no. 1, pp. 4–14, 2022.

[24] I. Titze, "Fluctuations and perturbations in vocal output," *Principles of voice production*, pp. 209–306, 1994.

[25] E. Vilkman, "Voice problems at work: a challenge for occupational safety and health arrangement," *Folia phoniatrica et logopaedica*, vol. 52, no. 1-3, pp. 120–125, 2000.

[26] M. K. Yagnavajjula, K. R. Mittapalle, P. Alku, P. Mitra *et al.*, "Automatic classification of neurological voice disorders using wavelet scattering features," *Speech Communication*, vol. 157, p. 103040, 2024.

[27] I. Yanushevskaya, C. Gobl, and A. N. Chasaide, "Voice quality and f0 cues for affect expression: implications for synthesis." in *Interspeech*, 2005, pp. 1849–1852.