Mittapalle, Kiran; Yagnavajjula, Madhu; Alku, Paavo

Classification of functional dysphonia using the tunable Q wavelet transform

Published in:
Speech Communication

DOI:
10.1016/j.specom.2023.102989

Published: 01/11/2023

Speech Communication 155 (2023) 102989

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Classification of functional dysphonia using the tunable Q wavelet transform

Mittapalle Kiran Reddy [a,*], Yagnavajjula Madhu Keerthana [a,b], Paavo Alku [a]

[a] Department of Information and Communications Engineering, Aalto University, Espoo, 02150, Finland
[b] Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, Kharagpur, 721302, India

ARTICLE INFO

Keywords:
Functional dysphonia
Tunable Q wavelet transform
Glottal features
MFCC
Convolutional neural network

ABSTRACT

Functional dysphonia (FD) refers to an abnormality in voice quality in the absence of an identifiable lesion. In this paper, we propose an approach based on the tunable Q wavelet transform (TQWT) to automatically classify two types of FD (hyperfunctional dysphonia and hypofunctional dysphonia) from a healthy voice using the acoustic voice signal. Using TQWT, voice signals were decomposed into sub-bands and the entropy values extracted from the sub-bands were utilized as features for the studied 3-class classification problem. In addition, the Mel-frequency cepstral coefficient (MFCC) and glottal features were extracted from the acoustic voice signal and the estimated glottal source signal, respectively. A convolutional neural network (CNN) classifier was trained separately for the TQWT, MFCC and glottal features. Experiments were conducted using voice signals of 57 healthy speakers and 113 FD patients (72 with hyperfunctional dysphonia and 41 with hypofunctional dysphonia) taken from the VOICED database. These experiments revealed that the TQWT features yielded an absolute improvement of 5.5% and 4.5% compared to the baseline MFCC features and glottal features, respectively. Furthermore, the highest classification accuracy (67.91%) was obtained using the combination of the TQWT and glottal features, which indicates the complementary nature of these features.

1. Introduction

Voice disorder is referred to as functional dysphonia when there are disturbances in voice quality without any obvious neurological, anatomical, or other organic difficulties affecting the larynx (Behlau et al., 2015; Reymond et al., 2006; Kiakojoury et al., 2014; Mumović et al., 2014). Functional dysphonia (FD) is the most frequent voice disorder among adults aged between 19 and 60 years (Martins et al., 2015). FD can manifest itself in two main forms, either as (i) hyperfunctional (hyperkinetic) dysphonia - the form associated with overuse of the laryngeal muscles and, occasionally, use of the false vocal folds, and (ii) hypofunctional (hypokinetic) dysphonia - the form associated with incomplete closure of the vocal folds due to reduced muscle tension (Reymond et al., 2006). Hyperfunctional dysphonia is common among people with voice-intensive occupations like teachers and singers, and is characterized by a tight, tense, loud and often deepened voice (Mumović et al., 2014). Conversely, hypofunctional dysphonia is mainly characterized by a breathy, low-pitched and weak voice. The goal of this study is to develop a 3-class classification system to perform automatic classification between a healthy voice and two types of FD voices (hyperfunctional and hypofunctional).

Typically, a voice disorder classification system can be developed using two approaches, namely, the traditional pipeline approach and the end-to-end approach. A system based on the traditional pipeline approach consists of two stages (feature extraction and classifier) (Reddy et al., 2021; Reddy and Alku, 2023). In the feature extraction stage, various features are extracted to capture the discriminative information present in voice signals. The extracted features are then used to train a classifier, such as a support vector machine (SVM) or convolutional neural network (CNN), to automatically distinguish healthy voices from disordered voices. In contrast, the end-to-end approach combines the feature extraction and classification steps into a single neural network that takes a voice signal (or its spectrogram) as input and generates the classification label as output (Narendra and Alku, 2020; Reddy et al., 2022, 2020). The use of end-to-end systems, however, is limited in the classification of disordered voices by the scarcity of training data in the study area. Because of the data scarcity, end-to-end systems cannot be trained effectively to learn the optimal feature mapping from the data (Narendra and Alku, 2020; Reddy et al., 2022). Hence, the traditional pipeline systems, which work well with smaller amounts of training data, are preferred in classification of voice disorders. It is worth noting that in most of the previous studies on automatic voice disorder classification, the focus is mainly on the binary classification problem (i.e., distinguishing a healthy voice from a disordered voice). However, the multi-class classification problem (i.e., classification between a healthy voice and several different voice disorders), which

* Corresponding author.
E-mail address: kiran.r.mittapalle@aalto.fi (K.R. Mittapalle).

https://doi.org/10.1016/j.specom.2023.102989
Received 9 June 2023; Received in revised form 30 August 2023; Accepted 21 September 2023
Available online 6 October 2023

would be more useful for clinical practitioners, has been investigated in only a few studies (Vaiciukynas et al., 2012; Behroozm and Almasganj, 2005; Kodrasi et al., 2021; Tirronen et al., 2023). This study focuses on using the traditional pipeline approach for developing classifiers to perform a 3-class classification task (healthy voice vs. hyperfunctional dysphonia vs. hypofunctional dysphonia).

In building effective traditional pipeline systems for classification of voice disorders, the selection of features is essential. In the literature, several acoustic features have been investigated, and the features studied can be grouped into four categories: (1) perturbation features (such as jitter and shimmer) (Silva et al., 2009; Vasilakis and Stylianou, 2009; Zhang et al., 2005); (2) spectral and cepstral features (such as Mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs), cepstral peak prominence (CPP) and perceptual linear prediction cepstral coefficients (PLPCCs)) (Reddy et al., 2021; Reddy and Alku, 2021; Wu et al., 2021; Fraile et al., 2011); (3) complexity features (such as the Hurst exponent and sample entropy) (Arias-Londoño et al., 2011; Arias-Londoño and Godino-Llorente, 2015); and (4) glottal features (such as time-domain and frequency-domain glottal source parameters) (Reddy and Alku, 2023; Narendra and Alku, 2020; Kadiri and Alku, 2019; Novotný et al., 2020). Among various feature types, cepstral features are widely used in the classification of voice disorders. The main advantage of a cepstral domain representation is that the features are less correlated, which is beneficial in building efficient machine learning (ML) models (Reddy and Alku, 2021; Kadiri and Alku, 2019). In addition, cepstral features can be computed without estimating the fundamental frequency (F_0) of the voice signal, which is beneficial compared to other feature types such as perturbation or glottal features whose computation depends on the extraction of F_0 (Reddy and Alku, 2021; Kadiri and Alku, 2019). Furthermore, cepstral features (such as MFCCs) have been shown to perform comparably to or better than the perturbation and complexity features (Kadiri and Alku, 2019; Gómez-García et al., 2019).

In this study, we propose to apply the tunable Q wavelet transform (TQWT) to voice signals for feature extraction. TQWT is a wavelet transform, which provides the ability to easily tune the Q-factor of the wavelet depending upon the oscillatory behavior of the signal (Selesnick, 2011; Reddy and Rao, 2019). TQWT results in the generation of more robust time-scale representations since it is based on the oscillatory behavior of the signal rather than on its frequency (Selesnick, 2011; Sakar et al., 2019). In disordered voices, possible disruptions in vocal folds may result in transient voice waveforms, and these abnormalities can be expected to be detected better with TQWT (Sakar et al., 2019). Motivated by this we propose utilizing the features derived using TQWT to capture the distinctive changes in voice signals for the classification of FD. The TQWT-based features have been previously used in speech tasks such as speech enhancement (Dash et al., 2021) and automatic detection of Parkinson's disease (PD) from speech (Sakar et al., 2019). However, as per our knowledge, this is the first study which utilizes TQWT in feature extraction in multi-class classification of voice disorders.

A few recent studies have analyzed the effectiveness of glottal source signals in discrimination of normal and disordered voices (Reddy et al., 2021; Reddy and Alku, 2021; Wu et al., 2021; Narendra and Alku, 2020; Reddy et al., 2020; Tirronen et al., 2023). These studies have shown that the glottal source waveform carries complementary voice quality-related information, and therefore combining the features derived from the glottal source with other features, such as MFCCs, can improve the classification performance. Therefore, in this work, we propose combining the features derived using TQWT with the glottal features to further enhance the multi-class classification performance. Taken together, the main goals of this work are as follows.

1. To study a multi-class classification problem in FD (healthy vs. hyperfunctional vs. hypofunctional) using a potential feature, TQWT, that has not been investigated before in the study area of pathological voice.
2. To study whether the multi-class classification performance could be improved by combining glottal features with the TQWT features.

The paper is organized as follows. Section 2 describes the VOICED database used for the classification task. Section 3 provides the details about the proposed FD classification system, the considered acoustic features, the CNN classifier and the evaluation criteria. The results are reported in Section 4. Finally, the conclusions of this study are provided in Section 5.

Table 1
Demographics of the participants considered in this study.

Health status	Age (in years)	Number of female subjects	Number of male subjects	Total
Healthy	18–34	21	7	28
	35–49	9	8	17
	≥50	6	6	12
	Total	**36**	**21**	**57**
Hyperkinetic	18–34	10	7	17
	35–49	16	7	23
	≥50	21	9	30
	Total	**47**	**23**	**70**
Hypokinetic	18–34	9	2	11
	35–49	10	2	12
	≥50	13	5	18
	Total	**32**	**9**	**41**

2. Database

In this work, we study pathological and healthy voices of the VOice ICar fEDerico II (VOICED) database (Cesari et al., 2018; Verde and Sannino, 2018). The publicly available VOICED database includes voice signals of 57 healthy speakers, 70 speakers with hyperfunctional dysphonia, and 41 speakers with hypofunctional dysphonia. All of the speakers were adults between 18 and 70 years of age. Subjects aged under 18 and over 70, or who had diseases, such as upper respiratory tract infections or neurological disorders, were excluded. The details about the number of male and female participants for each category are provided in Table 1. Each speaker produced the vowel [a] for 5 s without interruption. The collection of voice signals was performed in the medical room of the Institute of High Performance Computing and Networking (ICAR-CNR). Medical experts verified all the voice samples clinically according to the clinical protocol called SIFEL prepared by the Italian Society of Science and Bone Sciences. Based on the results of the medical (phoniatric) examination, the doctor diagnosed the presence or absence of a voice disorder. The collection process was carried out in a quiet and less dry environment with an ambient noise level of less than 30 dB. An m-health system called Vox4Health used the microphone on a Samsung Galaxy S4 to acquire the voice signals. All recordings were sampled at 8000 Hz with a resolution of 32 bits. In addition, any unexpected noise generated during the acquisition process was eliminated by applying an appropriate filter.

Fig. 1 shows examples of voice signals and corresponding spectrograms representing a healthy voice (left), a hyperfunctional dysphonic voice (middle) and a hypofunctional dysphonic voice (right). From Fig. 1(a), it can be seen that the time-domain waveform of a hyperfunctional dysphonic voice signal is more pressed and the time-domain waveform of a hypofunctional dysphonic voice signal is more rounded, compared to the healthy counterpart. Consequently, significant variations are present in the spectra of the three classes, especially in the harmonic structure shown in Fig. 1(b).

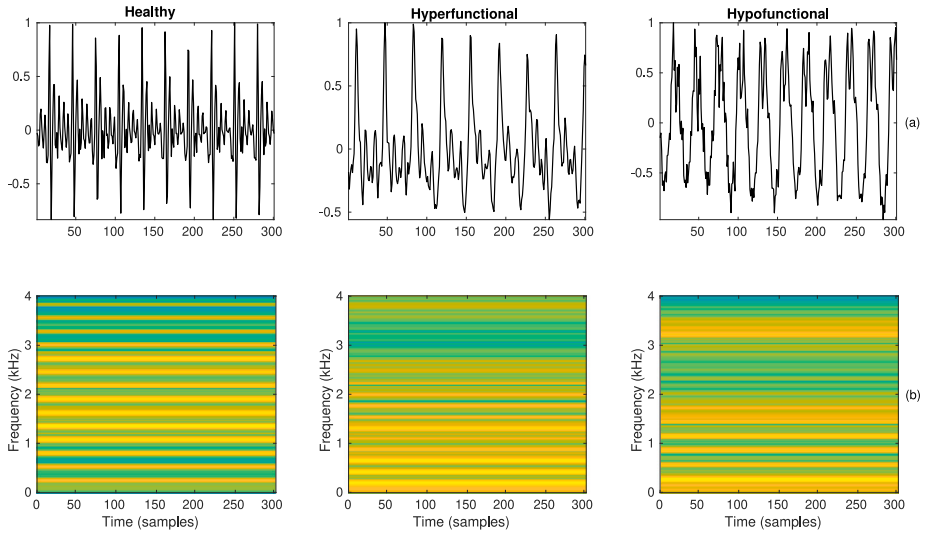


Fig. 1. An illustration of differences in voice signals (the vowel [a]) between a healthy female speaker (left), a female speaker with hyperfunctional dysphonia (middle) and a female speaker with hypofunctional dysphonia (right). The top panels show the time-domain voice signals and the bottom panels show the corresponding spectrograms.

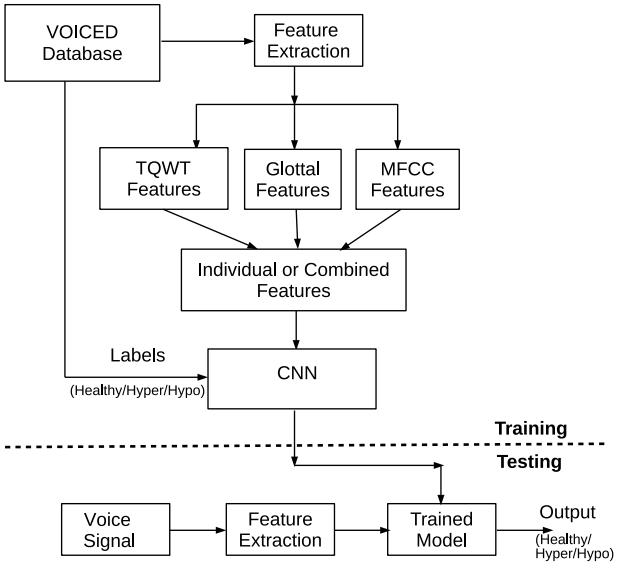


Fig. 2. Schematic block diagram of the proposed system for automatic classification of functional dysphonia.

2.1. Data pre-processing

Each voice signal included about 38000 discrete data samples of the recorded waveform, out of which the first 1000 were approximately equal to zero, and therefore these 1000 samples were deleted to avoid final classification errors. Furthermore, in order to increase the size of the database, each speaker's recording was divided into 9 chunks (each consisting of 4096 samples).

3. Architecture of the proposed system

The proposed system for the studied multi-class classification problem is shown in Fig. 2. During training, three types of features, namely, 13-dimensional MFCCs, 12-dimensional TQWT features, and 12-dimensional glottal features, are extracted from every voice signal present in the database. While the MFCC and TQWT features are extracted directly from the acoustic voice signals (described in Sections 3.1 and 3.3), the glottal features are extracted from the glottal flow waveforms (described in Section 3.2), which are estimated from acoustic voice signals by using the quasi-closed phase (QCP) glottal inverse filtering (GIF) algorithm (Airaksinen et al., 2014).

A CNN classifier is trained using the features extracted from the voice signals as input and the corresponding labels (healthy/hyper functional/hypofunctional) as output. Separate CNN classifiers are trained using individual features (MFCC, glottal, TQWT) and features sets where two individual features are combined (MFCC+TQWT, MFCC+glottal, TQWT+glottal). At the time of testing, the same set of speech features, which were used during training, are extracted from the test voice signals. The extracted features are given as input to the CNN classifier, which predicts the labels (healthy vs. hyperfunctional vs. hypofunctional). In this study, the classification systems developed using the widely used MFCCs and glottal features are considered baseline systems.

3.1. Extraction of the MFCC features

The variations in the spectra (as demonstrated by Fig. 1(b)) can be captured and represented in a compact form using MFCCs. Fig. 3 shows the steps involved in the MFCC feature extraction process. The input voice signal is first pre-emphasized and divided into several 30 ms frames using the Hamming window and a hop size of 10 ms. Next, the 512-point discrete Fourier transform (DFT) of each frame is computed. After this, a triangular filter bank consisting of 40 Mel-spaced filters is applied to the power spectrum. Finally, by computing the discrete cosine transform (DCT) of the logarithm of the filter bank's output, 13 MFCCs are obtained for each frame. The MFCCs computed from all frames are averaged to obtain a 13-dimensional feature vector for every voice signal.

3.2. Extraction of the glottal features

In recent studies, the glottal source signal, which is derived from the voice signal by using glottal inverse filtering (GIF), has been shown to carry complementary information related to voice disorders (Reddy et al., 2021; Reddy and Alku, 2021; Wu et al., 2021; Narendra and Alku, 2020; Reddy et al., 2020; Tirronen et al., 2023). In this study, we capture this information through a glottal feature vector consisting of 12 time- and frequency-domain parameters (Childers and Lee, 1991; Alku et al., 2002) computed from the glottal source waveform derived using the quasi-closed phase (QCP) GIF technique (shown in Fig. 2). The reason for choosing the QCP technique is that it has been shown

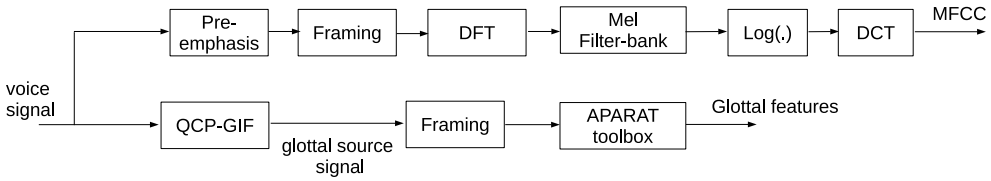


Fig. 3. Block diagram representation of the extraction of MFCC and glottal features. The DFT, DCT and log(.) blocks denote discrete Fourier transform, discrete cosine transform and logarithm operation, respectively.

Table 2
Time- and frequency-domain glottal parameters. For more details, see Childers and Lee (1991), Alku et al. (2002).

Time-domain glottal parameters	
OQ1	Open quotient, calculated from the primary glottal opening
OQ2	Open quotient, calculated from the secondary glottal opening
NAQ	Normalized amplitude quotient
AQ	Amplitude quotient
CIQ	Closing quotient
OQa	Open quotient, derived from the LF model
QOQ	Quasi-open quotient
SQ1	Speed quotient, calculated from the primary glottal opening
SQ2	Speed quotient, calculated from the secondary glottal opening
Frequency-domain glottal parameters	
H1H2	Difference between the first two glottal harmonics
PSP	Parabolic spectral parameter
HRF	Harmonic richness factor

to compute glottal source signals from non-modal voices better than several existing GIF techniques (Airaksinen et al., 2014). For complete details about the QCP algorithm, the reader is referred to Airaksinen et al. (2014). In total, 12 glottal parameters (listed in Table 2) are estimated using the APARAT toolbox (Airas et al., 2005), and they characterize various characteristics (e.g., different time-quotients and spectral tilt) of the glottal excitation. The glottal parameters are computed in 30 ms frames using a frame shift of 10 ms. While HRF and H1H2 are computed pitch-asynchronously once per frame, the remaining parameters are computed pitch-synchronously once per glottal cycle and then averaged over the frame. All the 9 time-domain parameters and PSP are expressed using a linear scale while H1H2 and HRF are expressed using the dB scale. The glottal parameters computed from all frames are finally averaged to obtain a 12-dimensional glottal feature vector for every voice signal.

3.3. Extraction of the TQWT features

TQWT is a wavelet transform, which enables easy tuning of the Q-factor of the wavelet depending upon the oscillatory behavior of the signal (Selesnick, 2011). The key parameters used for TQWT are the Q-factor Q, the redundancy r, and the decomposition levels J (Selesnick, 2011; Reddy and Rao, 2019). Usually, Q quantifies the number of oscillations that the wavelet exhibits. Typically, a wavelet with a Q-factor of 3 or larger consists of enough oscillatory cycles to process oscillatory signals (such as voice signals or electroencephalography signals) (Selesnick, 2011). On the other hand, when processing signals with little or no oscillatory behavior (such as a scan-line from a photographic image), the wavelet transform should have a low Q-factor (typically $Q = 1$) (Selesnick, 2011; Reddy and Rao, 2019). In this study, we use wavelets with a Q-factor of 2. The disordered voice signals embed transients in addition to oscillatory waveforms due to disruptions in vibration of the vocal folds. Therefore, by setting $Q = 2$ we hypothesize that TQWT can efficiently represent both oscillatory and transient behaviors. The parameter r can be interpreted as a measure of how much spectral overlap exists between adjacent band-pass filters (Selesnick, 2011; Reddy and Rao, 2019). The r value must be greater than 1, but increasing r has the effect of increasing overlap in the frequency domain of band-pass filters constituting TQWT (Reddy and Rao, 2019). Therefore, in this study, we set $r = 2$ to ensure the minimum overlap of band-pass filters.

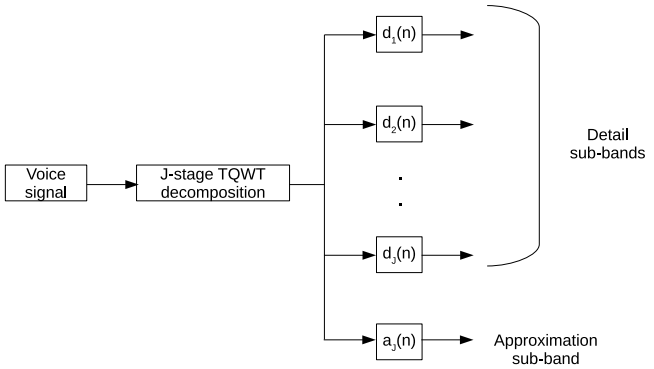
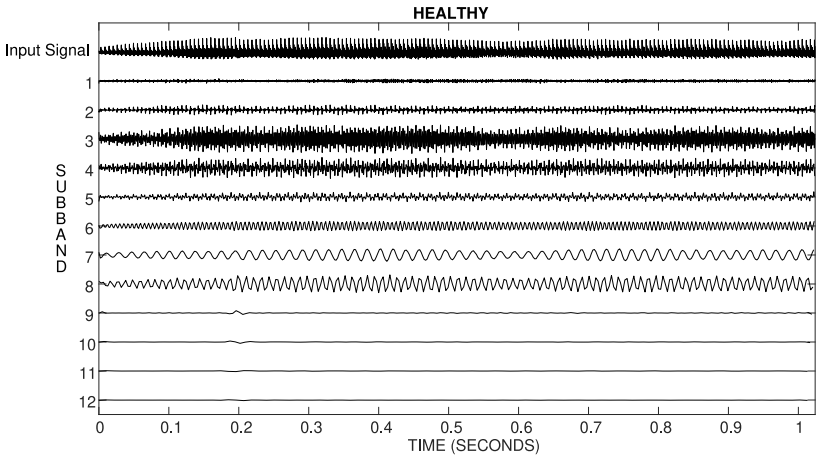
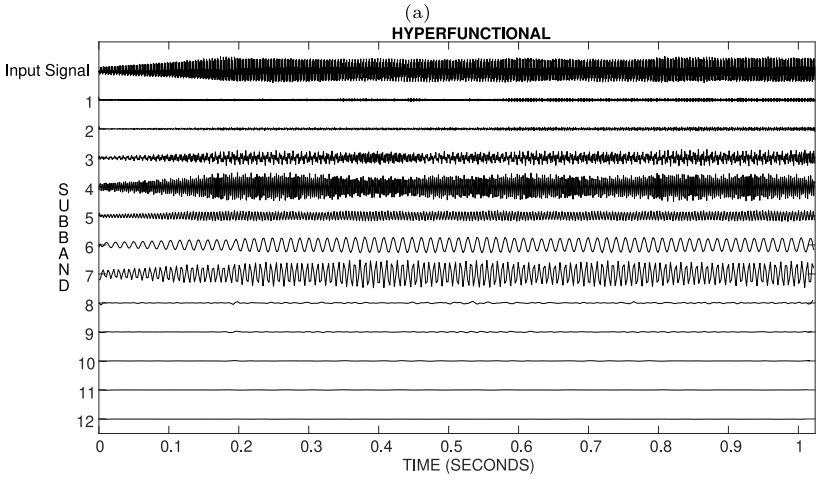


Fig. 4. Block diagram of TQWT with J-stage decomposition.

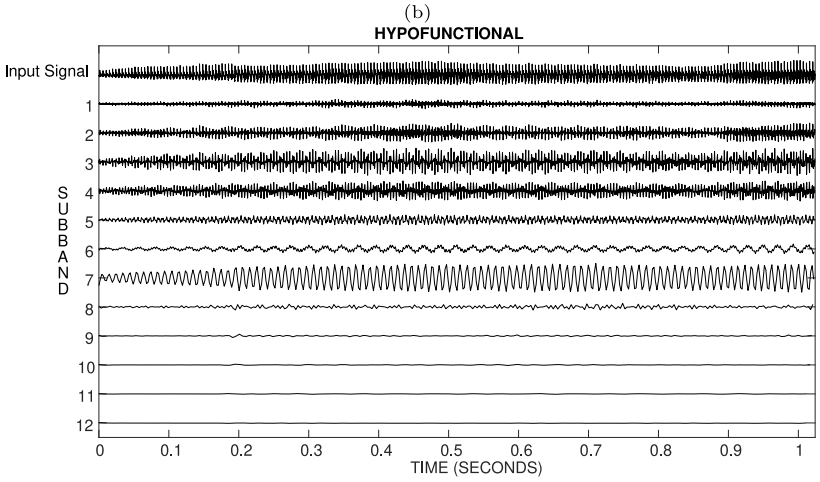
TQWT consists of a series of two-channel filter banks, with the low-pass output of each filter bank given as input to the succeeding filter bank (Selesnick, 2011; Reddy and Rao, 2019). Each output signal constitutes one sub-band of the wavelet transform. There will be $J + 1$ sub-bands, and these sub-bands are the low-pass filter output signal of the final filter bank (approximation sub-band) and the high-pass filter output signal of each filter bank (detail sub-bands). Fig. 4 demonstrates the J-stage TQWT decomposition. In this work, the value of J is chosen to be 11. Therefore, in total 12 sub-bands (11 detail sub-bands and 1 approximation sub-band) are obtained for each voice signal. Fig. 5 and Fig. 6 show the sub-bands obtained with the TQWT decomposition and the corresponding distribution of signal energy across the sub-bands for healthy and FD voice signals, respectively. Here, sub-band 1 corresponds to high frequencies while sub-band 12 corresponds to low frequencies. From the figures, it can be seen that the amount of energy in sub-bands 9–12 is negligible. This also justifies the decomposition of voice signals into only 12 sub-bands in the present work as sub-bands beyond 12 carry no significant information. Most importantly, a clear difference in the distribution of energy can be seen between the three classes. Compared to the healthy class, the relative energy in sub-band 8 is diminished in both of the FD classes. Furthermore, it can be seen that most of the energy of the healthy voice signal is concentrated on



(a)



(b)



(c)

Fig. 5. Illustration of the sub-bands computed with the TQWT decomposition for voice signals (the vowel [a]) produced by (a) a healthy female speaker, (b) a female speaker with hyperfunctional dysphonia, and (c) a female speaker with hypofunctional dysphonia.

sub-bands 3 and 4 indicating a smaller frequency leakage compared to the two FD voices. The hypofunctional voice signal has the largest frequency leakage, which can be attributed to breathy phonation.

After decomposing a signal with TQWT, log-energy entropy values of each sub-band are calculated to quantify how much information is carried in the relevant sub-band, and these entropy values are employed as features in classification of FD.

3.4. CNN

The classification experiments are performed using a one-dimensional convolutional neural network (1D-CNN). The input is fed to two sequential convolutional layers, each followed by the ReLU activation function and layer normalization. The first convolutional layer consists of 32 filters of size 3×3 and the second convolutional

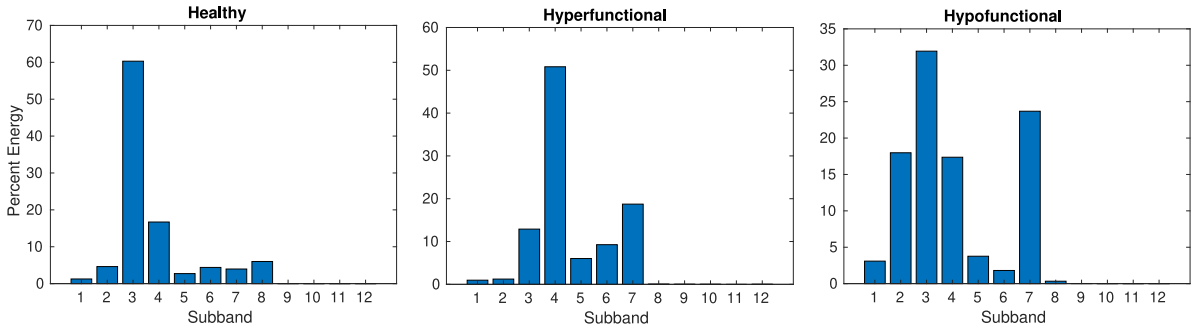


Fig. 6. Distribution of signal energy across sub-bands corresponding to the sub-bands shown in Fig. 5(a), 5(b) and 5(c).

Table 3
Results obtained for the multi-class classification task. ACC, PRE, REC and F1 refer to accuracy, precision, recall and F1, respectively. The numbers 0, 1 and 2 in the metrics refer to healthy voice, hyperfunctional dysphonia and hypofunctional dysphonia, respectively.

Feature	ACC (%)	PRE_0	REC_0	F1_0	PRE_1	REC_1	F1_1	PRE_2	REC_2	F1_2
MFCC	58.78	0.61	0.49	0.54	0.56	0.71	0.63	0.63	0.51	0.56
Glottal	59.80	0.57	0.55	0.56	0.58	0.66	0.62	0.67	0.55	0.60
TQWT	64.20	0.58	0.53	0.55	0.66	0.73	0.69	0.67	0.63	0.65
MFCC + TQWT	60.93	0.51	**0.69**	0.58	0.66	0.54	0.59	0.75	0.62	0.68
MFCC + Glottal	62.50	0.55	0.53	0.54	**0.67**	0.71	0.69	0.62	0.62	0.62
TQWT + Glottal	**67.91**	**0.69**	0.53	**0.60**	0.63	**0.78**	**0.70**	**0.77**	**0.70**	**0.74**

layer consists of 16 filters of size 3×3. The output of the convolutional layers is reduced to a single vector using a 1-D global average pooling layer. Finally, the resulting output is passed through a fully-connected layer with an output size matching the number of classes, followed by a softmax layer and a classification layer.

3.5. Evaluation scheme

The data from 80% of the speakers randomly selected from each class was used for training and the data from the remaining speakers was used for testing. 10% of the training data was used for validation. The validation dataset was used to give an unbiased estimate of model skill while tuning the CNN model's hyperparameters. No speaker used in training was used again in testing, thereby ensuring separation of speakers in the train and test set. The VOICED database has an imbalanced data distribution in terms of gender and class (see Table 1). Therefore, the training data was balanced with respect to class and gender by using the SMOTE (Synthetic Minority Over-Sampling Technique) algorithm (Chawla et al., 2002). SMOTE is the most popularly used oversampling technique where the synthetic samples are generated for the minority class. It focuses on the feature space to generate new minority instances between existing minority instances by utilizing linear interpolation. After balancing the data, the CNN was used for classification of FD. Four standard metrics, namely, accuracy, class-wise recall, class-wise precision, and class-wise F1-score were computed to evaluate the classification performance.

4. Results

The performance metrics obtained for all the features are shown in Table 3. From the table, it can be observed that in the case of the individual feature sets, the TQWT features provided the best performance in terms of accuracy (64.20%), F1-score for the hyperfunctional dysphonia class (0.69), and F1-score for the hypofunctional dysphonia class (0.65). The F1-scores obtained for the healthy class with all the individual features are comparable. Furthermore, combining the MFCC or TQWT features with glottal features has resulted in an improved classification performance. This indicates that there is complementary information between these feature sets. Overall, combining the TQWT features with the glottal features provided the best classification performance in terms of accuracy (67.91%), and F1-scores of 0.60, 0.70 and 0.74 for the healthy, hyperfunctional dysphonia and hypofunctional dysphonia classes, respectively. The results highlight that the distinctive changes in the time-frequency axis captured by TQWT are very effective in classification of normal and dysphonic voices. Altogether, the ability of the TQWT features to better capture abnormalities in voice signals combined with the ability of glottal features in characterizing the mode of phonation has resulted in an improved classification performance.

In order to demonstrate the behavior of the considered features, a one-way analysis of variance (ANOVA) was computed using MFCC, TQWT and glottal parameters extracted from the voice signals of 250 speakers from each of the three classes (healthy, hyperkinetic and hypokinetic). In general, a one-way ANOVA analysis tests the null hypothesis, i.e., it compares the means between the groups and determines whether any of those means are significantly different from each other. The results of ANOVA analysis are shown in Table 4. From the table, it can be seen that several MFCC, TQWT and glottal parameters show statistically significant differences ($p < 0.001$) between the three classes. Furthermore, the TQWT and glottal feature sets have a larger number of parameters showing statistically significant differences compared to the MFCC feature set. In conclusion, these results, which are based on statistical hypothesis testing, provide further evidence that the TQWT and glottal parameters have a better ability to distinguish the three classes compared to the MFCCs.

The confusion matrices for all the classification systems are shown in Fig. 7. It can be seen that similar performance is achieved for the healthy class with all the feature sets. The system developed with the combined features (TQWT+glottal) mainly increases the performance of the two FD classes. With the MFCC features, 90 hyperfunctional dysphonia samples are predicted correctly and 37 samples are misclassified as either healthy or hypofunctionally dysarthric. Similarly, 36 hypofunctional dysphonia samples are predicted correctly and 35 samples are mis-classified as either healthy or hyperfunctional. Compared to the MFCCs, the TQWT features improve the performance for both FD classes. With the TQWT features, 93 hyperfunctional dysphonia samples and 45 hypofunctional dysphonia samples are predicted correctly. The combination of the MFCC and glottal features provided better classification of the FD classes compared to MFCCs or glottal features alone. Overall, the combined feature set (TQWT+glottal) performed better than the other feature sets in classification of the two FD classes. With the system developed using TQWT+glottal features, 99 hyperfunctional dysphonia samples and 50 hypofunctional dysphonia samples are predicted correctly.

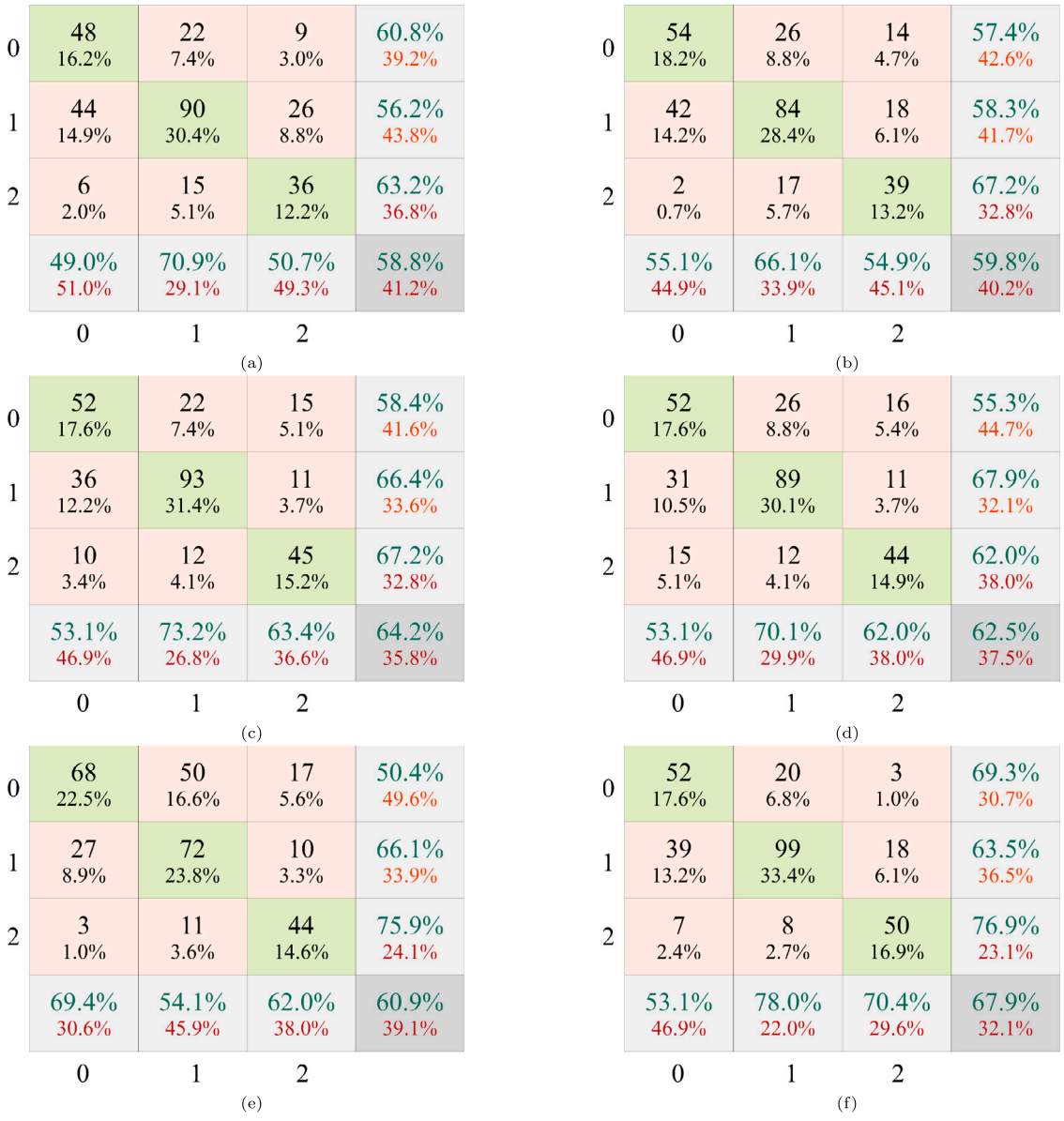


Fig. 7. Confusion matrices of the multi-class classification systems developed using (a) MFCC features, (b) glottal features, (c) TQWT features, (d) MFCC+glottal features, (e) MFCC+TQWT features and (f) TQWT+glottal features. The horizontal axis represents the true class, and the vertical axis represents the predicted classes. Class labels 0, 1, and 2 represent a healthy voice, hyperfunctional dysphonia, and hypofunctional dysphonia, respectively. The column on the far right of the plot shows the percentages of all the samples predicted to belong to each class that is correctly and incorrectly classified. The row at the bottom of the plot shows the percentages of all the samples belonging to each class that is correctly and incorrectly classified. The cell in the bottom right of the plot shows the overall accuracy. In the remaining cells, both the number of observations and the percentage of the total number of observations are shown in each cell.

5. Conclusions

In this paper, we investigated a 3-class classification task to automatically classify two FDs (hyperfunctional dysphonia and hypofunctional dysphonia) and healthy voices. The study proposed the use of log-energy entropy values that have been extracted from the sub-bands of TQWT as features for the considered multi-class classification task. Voice samples from the VOICED database were used in the experiments. Comparisons were made with two baseline features (MFCCs and glottal features) using CNN as the classifier.

The experimental results show that the TQWT features resulted in better classification performance compared to the widely-used MFCC features. This is because unlike in the MFCC extraction process, where the temporal information is lost due to windowing, the TQWT computation preserved temporal localization during the transform for the relevant sub-band Sakar et al. (2019). Hence, the time-frequency representation provided by TQWT can better characterize temporal abnormalities (like transients) of voice brought about by the voice excitation. The results also show that combining TQWT log-energy entropy values and glottal features resulted in improved classification performance, indicating the complementary nature of the TQWT and glottal features. We argue that this result is due to the glottal features' capability to quantify changes that are caused by the mode of vibration of the vocal folds (Reddy et al., 2021; Liu et al., 2023). In other words, while the TQWT features capture temporal abnormalities in voice signals, the glottal features provide supplementary information about the phonation mode to distinguish between voice signals generated using a modal vibration mode (as in a healthy voice), an adductive vibration mode (as in hyperfunctional dysphonia) and an abductive vibration mode (as in hypofunctional dysphonia).

Table 4
Results of one-way ANOVA analysis for the three feature sets (MFCC, TQWT, Glottal). SBE denotes Sub-band Entropy. The value after '_' indicates the sub-band number or MFCC feature number.

MFCC			TQWT			Glottal		
Feature	F-stats	*P-value*	Feature	F-stats	*P-value*	Feature	F-stats	*P-value*
MFCC_1	6.7086	0.0013	SBE_1	0.3575	0.6997	OQ1	25.2997	<0.001
MFCC_2	15.4389	<0.001	SBE_2	18.52	<0.001	OQ2	40.1684	<0.001
MFCC_3	19.1721	<0.001	SBE_3	19.5005	<0.001	NAQ	18.3784	<0.001
MFCC_4	50.5907	<0.001	SBE_4	41.9387	<0.001	AQ	6.6607	0.0014
MFCC_5	15.4720	<0.001	SBE_5	43.4599	<0.001	ClQ	20.2033	<0.001
MFCC_6	4.9943	0.0072	SBE_6	10.62	<0.001	OQa	7.7413	<0.001
MFCC_7	5.2634	0.0055	SBE_7	16.8794	<0.001	QOQ	49.5496	<0.001
MFCC_8	5.2485	0.0056	SBE_8	5.4096	0.0048	SQ1	28.6661	<0.001
MFCC_9	10.1602	<0.001	SBE_9	38.8733	<0.001	SQ2	33.5650	<0.001
MFCC_10	4.8894	0.0079	SBE_10	15.5555	<0.001	H1H2	8.1368	<0.001
MFCC_11	13.1761	<0.001	SBE_11	7.9737	<0.001	PSP	2.8697	0.0578
MFCC_12	4.9358	0.076	SBE_12	0.6794	0.5075	HRF	5.8979	0.0030
MFCC_13	16.8161	<0.001						

In conclusion, the study shows that combining features based on the tunable Q wavelet transform with glottal features constitutes an effective feature extraction approach in the studied 3-class problem. It is to be noted that the VOICED database considered in this study includes only one type of speaking task, production of the vowel [a]. Unlike vowels, continuous speech provides richer information related to the voice disorders. Therefore, the effectiveness of the combined features (TQWT + glottal) extracted from continuous speech in improving the classification performance needs to be studied. Furthermore, there was no information available about the severity of the voice disorders in the VOICED database. Therefore, the potential of the proposed features in the early detection of voice disorders, which is one of the most important applications in speech-based biomarking in general, could not be investigated in this study. In addition, previous studies (see, e.g., Narendra et al., 2019) have shown that glottal source features cannot be extracted robustly from telephone-quality speech. This might restrict the use of the proposed combination of the TQWT and glottal features in such remote monitoring applications where voice signal is recorded by phone and transmitted through the telephone network. However, developing a robust approach for glottal source extraction from telephone quality speech can aid in overcoming this limitation of the proposed approach.

CRediT authorship contribution statement

Mittapalle Kiran Reddy: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Yagnavajjula Madhu Keerthana:** Visualization, Writing – review & editing. **Paavo Alku:** Conceptualization, Validation, Visualization, Supervision, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research was supported by the Academy of Finland (project no. 330139), Aalto University (the MEC program for India) and Tata Consultancy Services India (the TCS Research Scholar Program).

References

Airaksinen, M., Raitio, T., et al., 2014. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. IEEE/ACM Trans. Audio, Speech, Lang. Process. 22 (3), 596–607.
Airas, M., et al., 2005. A toolkit for voice inverse filtering and parametrisation. In: Proc. INTERSPEECH. pp. 2145–2148.
Alku, P., et al., 2002. Normalized amplitude quotient for parameterization of the glottal glow. J. Acoust. Soc. Am. 112 (2), 701–710.
Arias-Londoño, J.D., Godino-Llorente, J.I., 2015. Entropies from Markov models as complexity measures of embedded attractors. Entropy 17 (6), 3595–3620.
Arias-Londoño, J.D., Godino-Llorente, J.I., et al., 2011. Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. IEEE Trans. Biomed. Eng. 58 (2), 370–379.
Behlau, M., Madazio, G., Oliveira, G., 2015. Functional dysphonia: strategies to improve patient outcomes. Patient Relat Outcome Meas. 6, 243–253.
Behroozm, R., Almasganj, F., 2005. Comparison of neural networks and support vector machines applied to optimized features extracted from patients' speech signal for classification of vocal fold inflammation. In: Proc. IEEE Int. Symposium on Signal Processing and Inforamtion Technolology. pp. 844—849.
Cesari, U., et al., 2018. A new database of healthy and pathological voices. Comput. Electr. Eng. 68, 310–321.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16 (1), 321–357.
Childers, D.G., Lee, C.K., 1991. Vocal quality factors: analysis, synthesis, and perception. J. Acoust. Soc. Am. 90 (5), 2394–2410.
Dash, T.K., Solanki, S.S., Panda, G., 2021. Multi-objective approach to speech enhancement using tunable Q-factor-based wavelet transform and ANN techniques. Circuits Systems Signal Process. 40, 6067–6097.
Fraile, R., Godino-Llorente, J.I., et al., 2011. Spectral analysis of pathological voices: sustained vowels vs running speech. In: Proceedings of the Seventh International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications.
Gómez-García, J.A., Moro-Velázquez, L., Godino-Llorente, J.I., 2019. On the design of automatic voice condition analysis systems, part II: Review of speaker recognition techniques and study on the effects of different variability factors. Biomed. Signal Process. Control 48, 128–143.
Kadiri, S.R., Alku, P., 2019. Analysis and detection of pathological voice using glottal source features. IEEE J. Sel. Top. Sign. Proces. 14 (2), 367–379.
Kiakojoury, K., Dehghan, M., Hajizade, F., Khafri, S., 2014. Etiologies of dysphonia in patients referred to ENT clinics based on videolaryngoscopy. Iran. J. Otorhinolaryngol. 76 (26), 169–174.
Kodrasi, I., et al., 2021. Automatic and perceptual discrimination between dysarthria, apraxia of speech, and neurotypical speech. In: Proc. IEEE International Conference on Acoustics Speech and Signal Processing. pp. 7308–7312.
Liu, Y., et al., 2023. Automatic assessment of Parkinson's disease using speech representations of phonation and articulation. IEEE/ACM Trans. Audio Speech Lang. Proc. 31, 242–255.
Martins, R.H.G., et al., 2015. Voice disorders: Etiology and diagnosis. J. Voice 30 (6), 761.e1—761.e9.
Mumović, G., Veselinović, M., Arbutina, T., Škrbić, R., 2014. Vocal therapy of hyperkinetic dysphonia. Serbian Arch. Med. 142 (11–12), 656–662.
Narendra, N.P., Airaksinen, M., Story, B., Alku, P., 2019. Estimation of the glottal source from coded telephone speech using deep neural networks. Speech Commun. 106, 95–104.
Narendra, N.P., Alku, P., 2020. Glottal source information for pathological voice detection. IEEE Access 8, 67745–67755.
Novotný, M., Dušek, P., Daly, I., Růžiška, E., Rusz, J., 2020. Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with Parkinson's disease: Correlation between acoustic speech characteristics and non-speech motor performance. Biomed. Signal Process. Control 57, 101818.

Reddy, M.K., Alku, P., 2021. A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation. IEEE Access 4 (9), 135953-135963.
Reddy, M.K., Alku, P., 2023. Exemplar-based sparse representations for detection of Parkinson's disease from speech. IEEE/ACM Trans. Audio Speech Lang. Proc. 31, 1386–1396.
Reddy, M.K., Alku, P.A., Rao, K.S., 2020. Detection of specific language impairment in children using glottal source features. IEEE Access 8, 15273–15279.
Reddy, M.K., Helkkula, P., et al., 2021. The automatic detection of heart failure using speech signals. Comput. Speech Lang. 69, 101205.
Reddy, M.K., Keerthana, Y.M., Alku, P., 2022. End-to-end pathological speech detection using wavelet scattering network. IEEE Signal Process. Lett. 29, 1863–1867.
Reddy, G.R.S., Rao, R., 2019. Oscillatory-plus-transient signal decomposition using TQWT and MCA. J. Electron. Sci. Technol. 17 (2), 135–151.
Reymond, H., Colton, K., Casper, R., 2006. Understanding voice problems. A Physiological Perspective for Diagnosis and Treatment, 3rd Ed. Lippincott Williams and Wilkins.
Sakar, C.O., et al., 2019. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. Appl. Soft Comput. 74, 255–263.
Selesnick, I.W., 2011. Wavelet transform with tunable Q-factor. IEEE Trans. Signal Process. 59 (8), 3560–3575.
Silva, D.G., Oliveira, L.C., Andrea, M., 2009. Jitter estimation algorithms for detection of pathological voices. EURASIP J. Adv. Signal Process. 2009, 1–9.
Tirronen, S., Kadiri, S.R., Alku, P., 2023. Hierarchical multi-class classification of voice disorders using self-supervised models and glottal features. IEEE Open J Signal Process. 4, 80–88.
Vaiciukynas, E., et al., 2012. Exploring similarity-based classification of larynx disorders from human voice. Speech Commun. 54 (5), 601–610.
Vasilakis, M., Stylianou, Y., 2009. Voice pathology detection based on short term jitter estimations in running speech. Folia Phoniatr. Logop. 61 (3), 153–170.
Verde, L., Sannino, G., 2018. XVOICED database. Available online: https://physionet.org/content/voiced/1.0.0/.
Wu, Y., Zhou, C., Fan, Z., et al., 2021. Investigation and evaluation of glottal flow waveform for voice pathology detection. IEEE Access 9, 30–44.
Zhang, Y., Jiang, J.J., et al., 2005. Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis. J. Voice 19 (4), 519–528.