International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 1, February 2022, pp. 946~956
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i1.pp946-956

Classification of three pathological voices based on specific features groups using support vector machine

Muneera Altayeb[1], Amani Al-Ghraibah[2]
[1]Department of Electronics and Communications Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan
[2]Department of Medical Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan

Article Info

Article history:

Received Feb 10, 2021
Revised Aug 3, 2021
Accepted Aug 20, 2021

Keywords:

Discrete wavelet transform
Mel frequency cepstral-coefficients
Support vector machine
Voice disorders

ABSTRACT

Determining and classifying pathological human sounds are still an interesting area of research in the field of speech processing. This paper explores different methods of voice features extraction, namely: Mel frequency cepstral coefficients (MFCCs), zero-crossing rate (ZCR) and discrete wavelet transform (DWT). A comparison is made between these methods in order to identify their ability in classifying any input sound as a normal or pathological voices using support vector machine (SVM). Firstly, the voice signal is processed and filtered, then vocal features are extracted using the proposed methods and finally six groups of features are used to classify the voice data as healthy, hyperkinetic dysphonia, hypokinetic dysphonia, or reflux laryngitis using separate classification processes. The classification results reach 100% accuracy using the MFCC and kurtosis feature group. While the other classification accuracies range between~60% to~97%. The Wavelet features provide very good classification results in comparison with other common voice features like MFCC and ZCR features. This paper aims to improve the diagnosis of voice disorders without the need for surgical interventions and endoscopic procedures which consumes time and burden the patients. Also, the comparison between the proposed feature extraction methods offers a good reference for further researches in the voice classification area.

Corresponding Author:

Muneera Altayeb
Department of Electronics and Communications Engineering, Faculty of Engineering, Al-Ahliyya Amman University
Al-Saro, Al-Salt, Amman, Jordan
Email: m.altayeb@ammanu.edu.jo

1. INTRODUCTION

Speech is considered as one of the most important means of communication among humans. Therefore, when any defect occurs in the speech system, this considered as an impediment in communication among people. Difficulty in speech may arise due to imbalance in the speech or auditory system [1].

Many researchers in literature have studied speech disorders and vocal pathology by analyzing and classifying samples of patient's voices. The purpose was to help patients with pathological problems and to monitor the progress of the vocal therapy pathway and to minimize the use of traditional diagnostic pathologies of vocal pathology. Researchers developed many diagnosis methods for observations of vocal folds by means of laparoscopic tools. However, these techniques are risky, time consuming, discomfort and require expensive resources [2], [3]. From the in Ankışhan work [4], a new approach for detection of pathological voice disorders was developed with minimum parameters. The preprocessing step has been carried out since the recording of the sound data. The sound data is re-modeled with the calculated

coefficients using two different models: linear predictive coding (LPC) and Mel frequency cepstral coefficients (MFCC). Then, the recorded speech is divided into two different signal types, clean and residual signals. The signals that modeled from the coefficients are called clean data and are removed from the recorded data to generate the residual data. The modeled signals were first separated into sub-frames, then the characteristics (features) of each frame were extracted, the features that are extracted in their study are: jitter, shimmer, skewness, kurtosis, entropy, and largest lyapunov exponents (LLEs). Finally, a pathological classification was made depending on these features, where 30% of the data were randomly selected as testing data and 70% were randomly selected as training data. The classification performance in their work was very good. Using the ten features, the accuracy rate of training data was 100%, however, the estimated accuracy for testing data was 99.56% [4].

A detection of pathological voices was also developed by Fang [5], using cepstrum vectors and a deep learning approach. This study retrospectively collected 60 normal voice samples and 402 pathological voice samples of 8 common clinical voice disorders in a voice clinic of a tertiary teaching hospital. They extracted MFCCs from 3-second samples of a sustained vowel. The performances of three machine learning algorithms, namely, deep neural network (DNN), support vector machine (SVM), and Gaussian mixture model (GMM), were evaluated based on a fivefold cross-validation. Collective cases from the voice disorder database of Massachusetts Eye and Ear Infirmary (MEEI) were used to verify the performance of the classification mechanisms. The experimental results demonstrated that DNN outperforms GMM and SVM. Its accuracy in detecting voice pathologies reached 94.26% and 90.52% in male and female subjects, based on three representative MFCC features. When applied to the MEEI database for validation, the DNN also achieved a higher accuracy (99.32%) than the other two classification algorithms. They concluded that stacking several layers of neurons with optimized weights, the proposed DNN algorithm can fully utilize the acoustic features and efficiently differentiate between normal and pathological voice samples [5].

Panek *et al.* [6] created an acoustic analysis assessment in detecting four major speech diseases: excessive dysfunction, dysfunction, laryngitis, vocal cord paralysis. At the beginning, 28 acoustic parameters were evaluated by examination. The analysis of the speech signal was performed by extracting many features, namely: fundamental frequency, jitter and shimmer coefficients, energy, zeroth, first, second, third-order moment, kurtosis, power factor, 1, 2 and 3-formant amplitude, 1, 2 and 3-formant frequency, maximum and minimum values of the signal and 10 MFCCs. The classification consisted of results from the analysis for each patient. It was analyzed using three methods: Principal component analysis (PCA), kernel principal component analysis (KPCA) and an auto-associative neural network (NLPCA). Ten-fold cross-validation was used, where the data was divided into 10 subsets; 10% of the data was used as a testing set, and the remaining 90% was representing as a training set. The analysis was completed individually for each vowel at different intonations, separately for men and women for each pathology and each vowel at a different pitch. The aim of their research was to perform a classification that can distinguish between healthy and pathological voices [6].

The novelty of this work lies in extracting new features from healthy voices and three different pathological voice samples followed by several classification processes to classify the voice samples as healthy or pathology voices using specific feature groups which contain a combination of the extracted features. The extracted features are mainly three different MFCC feature groups and wavelet features group. Also, the discrete wavelet transform (DWT) method is unique and has not been used before in the mean of voice disorder classification.

2. PROPOSED METHOD

The main steps of this work began with extracting features from the voice data, then use the feature groups to build an automated system using SVM which classify the input data as normal or pathology voices. The features of the voice samples were extracted using: MFCCs, zero-crossing rate (ZCR) and discrete wavelet transform (DWT) in addition to other statistical features which are: skewness, kurtosis, and entropy. Figure 1 shows the block diagram of the proposed work.

This paper is organized as follows: section 1 contains a brief description of the data used in this work; section 2 presents the methodology of this work. Section 3 presents the results of classification. Section 4 contains a conclusion of all the work featured in this paper.

2.1. Database

The registered database which was used in this study contains voices of healthy and pathological people besides much information about each patient. The voice signal acquisitions were performed in the Hospital University of Naples "Federico II", at the medical room of the "Institute of High-Performance Computing and Networking" [7].

The recordings were made using a mobile microphone which was held at about 20 cm away from the patient. Each signal consists of a recording of vocalization of the vowel \\a\\ five seconds in length without any interruption of other sounds. The recording signals were sampled at 8000 Hz and their resolution was 32-bits. All samples were recorded in less than 30 dB of background noise and room humidity was greater than 30-40%. Each recording was filtered to remove any noise accidentally added during the acquisition [8]. Table 1 shows the details of the number of voice samples of each case used in this work.

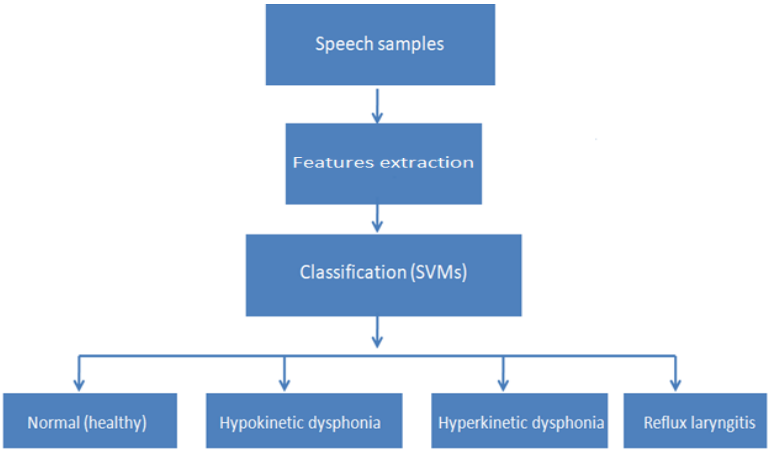

Figure 1. Block diagram of the proposed work

Table 1. Patient database

Patient's status	Female	Male	Total
Healthy	33	21	54
Hyperkinetic dysphonia	29	19	48
Hypokinetic dysphonia	40	22	62
Reflux laryngitis	18	20	38

2.2. Feature extraction

Many methods were used in literature to extract features from human voices. This paper proposes three different methods beside the statistical analysis to extract six feature groups. The first two methods MFCCs and ZCR are applied in the time and frequency domains, also other features are extracted using DWT. In the first step, the data were filtered using proper filter then blocked into 38 frames, then delta, delta-delta MFCC, ZCR and other statistical features were extracted from each frame for all voice samples. DWT was carried on this work based on 5-level decomposition of the voice signals to extract 5 features: one from each level. After preparing the features, classification processes were performed using SVM based on six feature groups and based on a combination of the most significant features. In the following sections, an explanation of each feature extraction method is proposed followed by details of the features groups that were used at each classification process.

2.2.1. Mel frequency cepstral coefficients (MFCC)

MFCCs are types of cepstral representation of the signal, where the frequency bands are distributed depending on the Mel-scale. The MFCCs are basically include windowing the signal, applying the discrete Fourier transformation (DFT), followed by Mel filter banks, taking the logarithm of all filter bank energies, then applying the discrete cosine transformation (DCT). The steps involved in the MFCC features extraction are summarized in Figure 2 [9].

2.2.2. Pre-emphasis

Pre-emphasis refers to filtering which emphasizes the higher frequencies, the modeling of the sound signals and extracting of the features from the modeled data which are all performed after some preprocessing steps. The speech signal S(n) is sent to a high-pass filter given by (1),

$$\check{S}(n) = S(n) - \beta S(n-1) \quad (1)$$

where, S(n) is input signal, βis a constant and it is around 0.97 in this work, and Š(n) is the signal after filtration.

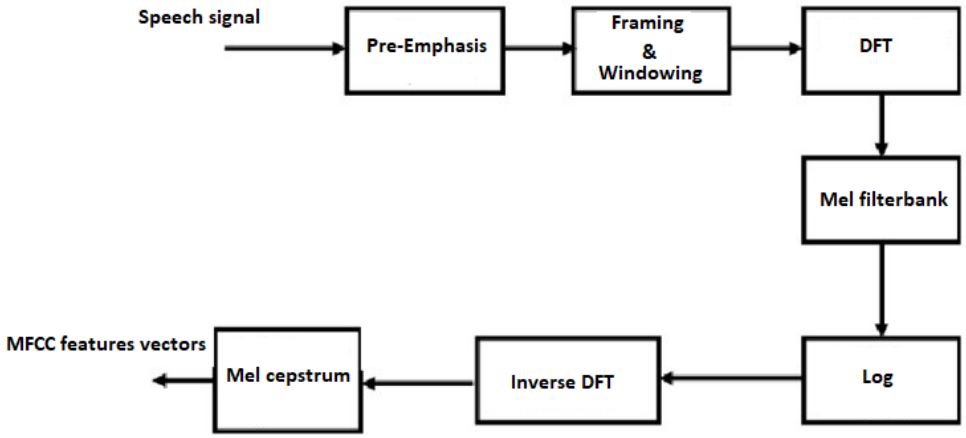

Figure 2. MFCC features extraction [10]

2.2.3. Framing and windowing

The speech signal is time-varying or non-stationary signal, therefore speech analysis is always performed by broken the signal into possibly overlapping frames, so that the speech signal is constant [11]. In this step, the continuous 1D voice signals are blocked into 38 frames of N=2000 samples, with next frames separated by L=512 samples, the adjacent frames are overlapped by N-L samples around 74%. Windowing is done to enhance the harmonics and smooth the edges at the beginning and ending points of the frame. Mainly hamming window represented by w(n) multiplies by the input signal represented x(n). The hamming window amplitude is shown in Figure 3, the output signal represented by (2):

$$x_i(n) = x(n)w(n - m_i), \qquad i = 0, \ldots, K - 1 \tag{2}$$

where K is the number of frames and m_i is the number of samples by which the window is shifted in order to yield the *i-th* frame then taking the DFT of the resulting signal $x_i(n)$ [9].

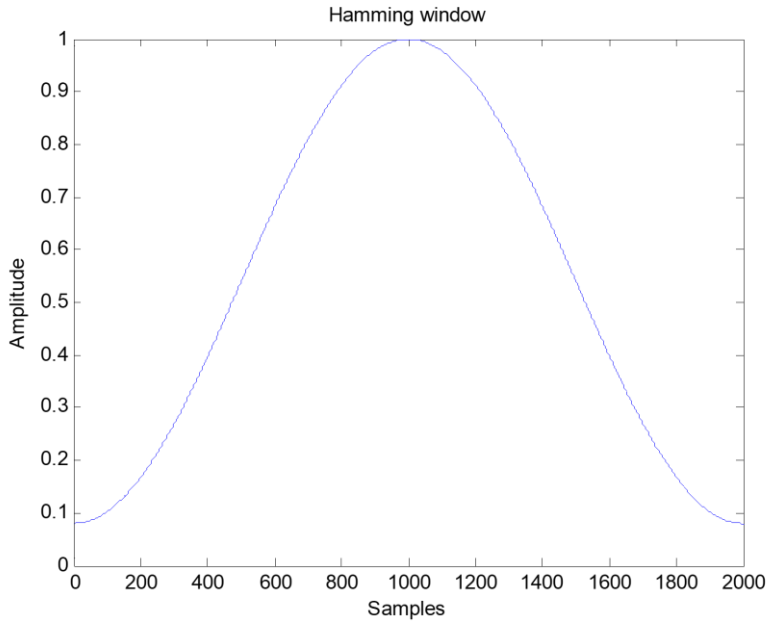

Figure 3. Hamming window

2.2.4. Mel-spectrum

The resulting spectrum of Fourier transformed signal is given as input to a Mel-scale filter bank that consists of 24 filters as shown in Figure 4. A Mel is a unit of measure based on the human ears perceived frequency where human ears are not sensitive enough to detect sounds below 1000 Hz when frequency warping process occurs, the coefficients of each short time Fourier transformed (STFT) are multiplied by the corresponding filter gain. A popular formula to convert f in hertz into f_{mel} is given in (3) [12]-[14].

$$\mathrm{f_{mel}} = 2595 \log_{10} (1 + \frac{f}{700}) \quad (3)$$

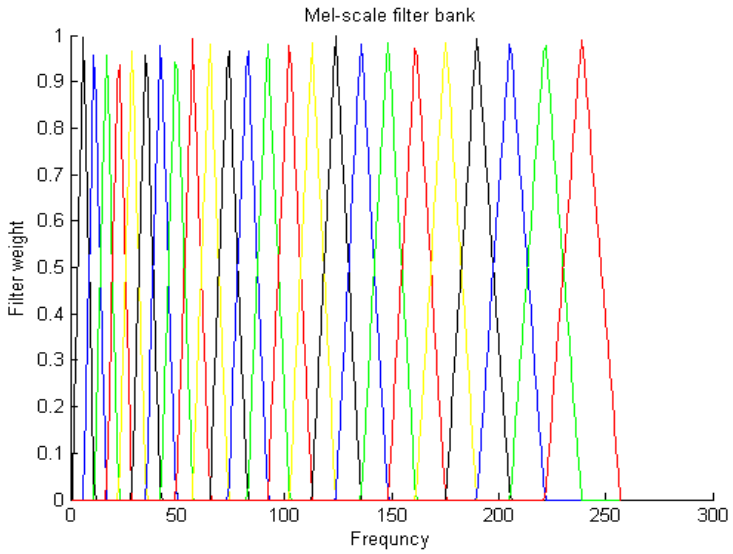

Figure 4. Mel-scale filter bank

The DCT applied to the transformed Mel frequency coefficients produced a set of MFCC campestral coefficients. The cepstral coefficients are usually referred to as stationary features, the extra information about the dynamic features of the signal is obtained by computing first derivative of cepstral coefficients and it is called delta coefficients, the second order derivative is called delta-delta coefficients [15]. Figure 5(a)-(d) clearly shows the differences in the values of MFCC coefficients for the healthy signal compared with, hyperkinetic pathology signal, hypokinetic pathology signal, and reflux pathology signal. These obvious differences in the MFCC coefficients will be very helpful in the classification process, where the x-axis represents the number of MFCCs extracted from the input signal and the y-axis represents the feature values for each frame. In this work, 12 MFCC coefficients were used which are from the 2nd to the 13th coefficients and the rest were discarded. The lower order coefficients contain most of the information about the overall spectral shape according to the feature values shown in Figure 5 as we can observe the difference in MFCCs for the four cases.

2.2.5. Data clustering using K-means

K-means clustering is popular in signal processing field, it can be used to cluster the extracted features from speech signals, and it is applied to relatively large sets of data. K-means separate the data into spherical clusters by selecting 'k' number of distinct clusters then finding a set of cluster centers. The common Euclidean distance between the point and each cluster center is computed and the sum of error (SSE) is then calculated using (4) [16],

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2 (m_i, x) \quad (4)$$

where, x is a data point in cluster C_i, K is the number of clusters, mi corresponds to the center (mean) of the cluster and dist is the Euclidean distance. Here, the MFCCs coefficients were computed for about 38 frames

of voice samples with 12 coefficients related to each frame. Vector quantization based on K-means clustering was done with respect to the cluster index to reduce the size of the feature vector for each voice signal. Thus, vector of 12 features were extracted from each voice signal and used in further classification processes.

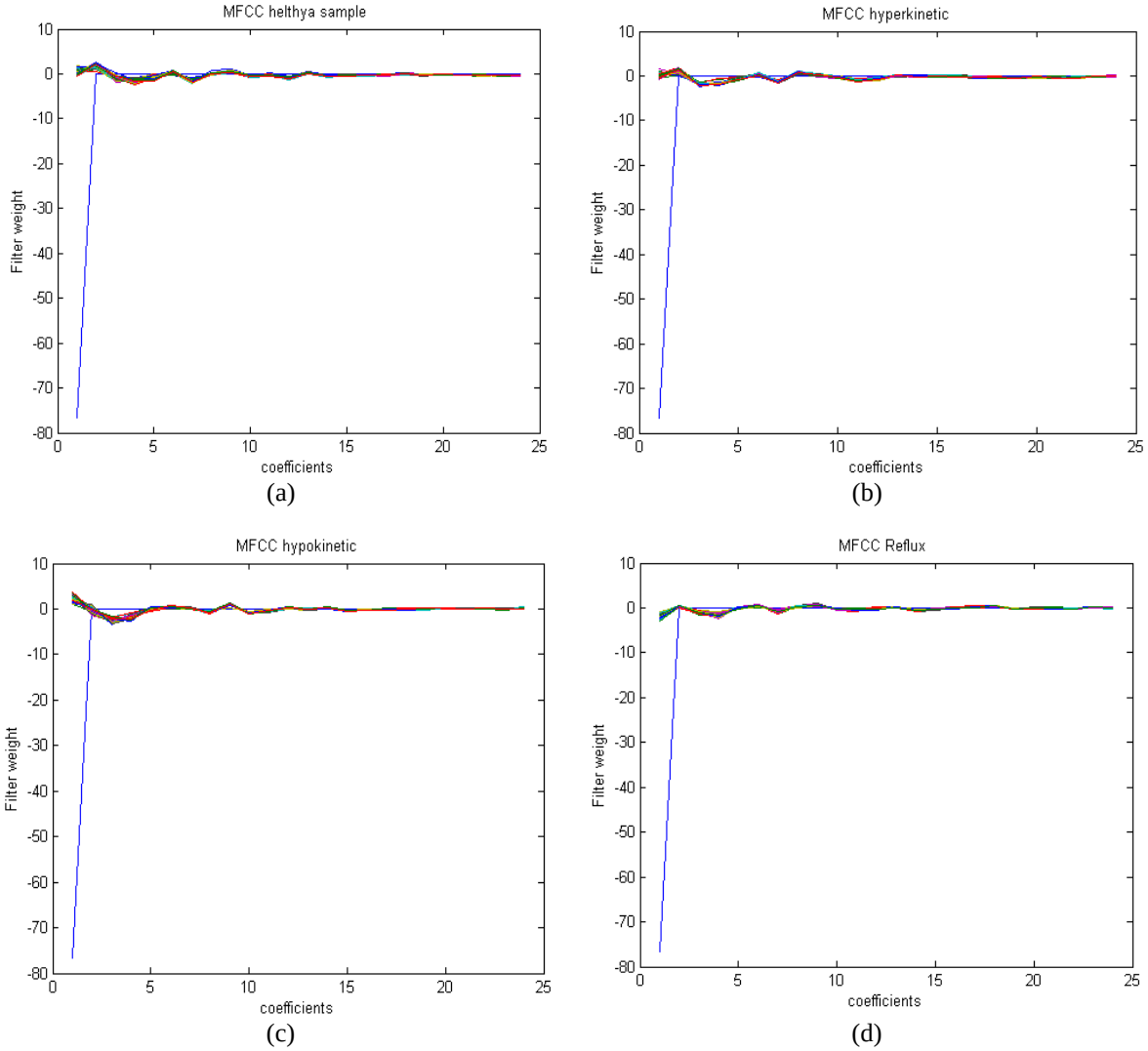

Figure 5. MFCCs extracted from the signals; (a) healthy signal, (b) hyperkinetic pathology signal, (c) hypokinetic pathology signal, (d) reflux pathology signal

2.3. Statistical feature extraction

In this study, the filtered signals from the preprocessing step were separated into frames, which were then used to calculate the 24 statistical features: 6-skewness, 6-kurtosis, 6-entropy and 6-ZCR.

2.3.1. Skewness

Skewness is a measure of the distortion asymmetry of the probability distribution and for any signal $\hat{s}(t)$ and it is defined as the standardized third moment of this signals which is given by (5) [4]:

$$\text{Skew}\left[\hat{S}(t)\right] = E\left[\frac{\hat{S}-\mu}{\sigma}\right]^3 \quad (5)$$

where, *E* represents the expected operator, μ is the mean of the signal and σis the standard deviation of the signal.

2.3.2. Kurtosis

The kurtosis of the signal is defined as the standardized fourth moment of the signals $\hat{S}(t)$ and it is a measure of the combined weight of a distribution's tails relative to the center of the distribution, and it is given by (6) [4]:

$$\text{Skew}\left[\hat{S}(t)\right] = \frac{E[(\hat{S}-\mu)^4]}{(E[(\hat{S}-\mu)^2])^2} \quad (6)$$

where, E represents the expected operator, μ is the mean of the signal.

2.3.3. Entropy

Entropy is the probability distribution of the signal $\hat{S}(t)$, or is the average level of information or uncertainty. The relational probability of the events $\hat{S}_i$(where i=1, 2, 3, …, k) called the self-probability $h(p_i)$ and it is defined in (7). The Entropy (H) is defined in (8) and it is the weights of K numbers of self-information values [17].

$$h(p_i) = log_2\left(\frac{1}{p_i}\right) \quad (7)$$

$$H = -\sum_{i=1}^{k} p_i log_2\left(\frac{1}{p_i}\right) \quad (8)$$

2.3.4. Zero-crossing rate (ZCR)

The zero-crossing rate (ZCR) of an audio frame is the rate of sign changes of the signal during the frame. The ZCR is defined in the (9) [18],

$$Z(i) = \frac{1}{2W_l}\sum_{n=1}^{W_l} |sgn[\hat{S}_i(n)] - sgn\left[\hat{S}_i(n-1)\right]| \quad (9)$$

where, W_l is the length of the frame and $sgn(\cdot)$ is the sign function. Usually, ZCR is used to separate signal as voiced and unvoiced, but here this method is used to extract features which will help to classify the voice signal as normal or abnormal voices according to the voice signal nature. ZCR can be interpreted as a measure of the noisiness of a signal; it usually returns higher values in the case of a noisy signal [18].

2.4. Discrete wavelet transform (DWT)

The voice signal is decomposed into N levels using DWT, where N must be a strictly positive integer chosen to be five levels in this paper. In the first step of the DWT-based analysis, the DWT of the voice signal $s(t)$ produces two sets of coefficients: approximation coefficients cA1, and detail coefficients cD1. These vectors are obtained by convolving the signal s with the low-pass filter Lo_D for approximation, and with the high-pass filter Hi_D for detail, followed by dyadic decimation (down sampling) as shown in the block diagram in Figure 6, where the length of each filter is equal to 2N [19], [20]. The next step splits the approximation coefficients cA_1 into two parts following same scheme in the first step by replacing *s* by cA_1, and hence producing cA_2 and cD_2; and steps continue as such *N* times. Following these steps, the wavelet decomposition of the voice signal s(t) (analyzed at level *N*=5) results in the structure: [*cA5, cD5, ..., cD1*] as shown in Figure 7.

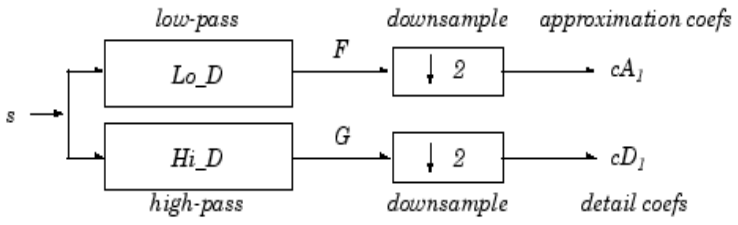

Figure 6. The first step of DWT

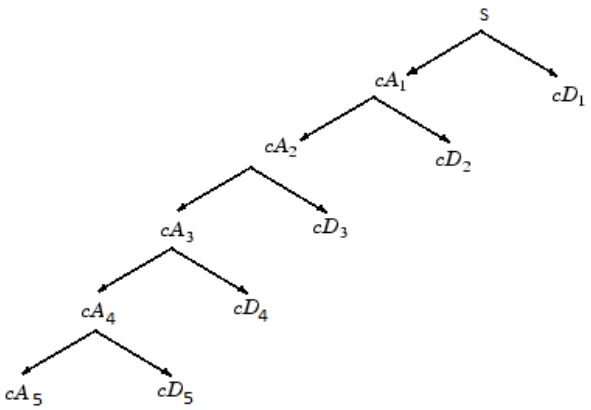

Figure 7. The general structure of DWT of 5 levels

3. CLASSIFICATION RESULTS

Support vector machines (SVMs) are state-of-the-art classifiers, SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, according to the SVM methodology, a kernel function is used in order to map the feature vectors to the 'kernel space'. In this work, 10-fold cross-validation on the training data were used to create the model. 10 samples from each case are kept for testing while the remaining samples were used in training. The target variable corresponds to a decision that input data x belongs to normal voice (class 0) or abnormal voice (class 1) [21]-[23].

The classification process was performed six times using the following feature groups: Group 1 includes delta-MFCC features alone, group 2 has delta-delta MFCC and Kurtosis features, group 3 includes delta-delta MFCC and skewness features, while group 4 has delta-delta MFCC and ZCR features together. Also, group 5 has the delta-delta MFCC plus the entropy features. The five DWT features are used as group 6 in the last classification process. Section 6 provides details about the feature groups and the resultant classification accuracies [21], [22].

3.1. Performance evaluation

The total numbers of speech samples used in this work are 202 for the evaluation purpose of which 148 are pathology while 54 are normal voices. The terms used in the confusion matrix as shown in Table 2 can briefly be described as: true positive (TP): true decisive system classified as true; true negative (TN): false event detected as false; false positive (FP): the event is false and discriminated as true; and false negative (FN): true event classified as false [24]-[26].

Table 2. Confusion matrix

Confusion matrix		
	Normal	Pathology
Normal	TP	FP
Pathology	FN	TN

Also, accuracy (AC) is defined as the probability that the classification by the system is correct and it is given by (10) [20]:

$$AC = \frac{TP + TN}{(TP + FP + TN + FN)} * 100 \quad (10)$$

The sensitivity (true positive rate (TPR)) and specificity (true negative rate (TNR)) are also calculated from the confusion matrix using (11), and (12) respectively [20]:

$$TPR = \frac{TP}{TP+FN} \quad (11)$$

$$TNR = \frac{TN}{TP+FN} \quad (12)$$

3.2. Classification using delta, delta-delta MFCC, ZCR and other statistical features

Here, delta-MFCC, delta-delta MFCC, ZCR and other statistical features are used to create five feature groups which are named by F1: Delta MFCC features, F2: Related to delta-delta MFCC and kurtosis features, F3: Delta-delta MFCC and skewness features, F4: Delta-delta MFCC with ZCR features, while F5: Delta-delta MFCC plus entropy features. Where the number of normal data=54, hyperkinetic=64, hypokinetic=45, reflux=38 samples. In each case 10 samples are kept for testing and the remaining samples are used in training. Tables 3, 4 and 5 show the result of the classification process including training-, testing-accuracy, TNR and TPR using each feature group and repeated for the three different pathological cases versus the normal voice signal.

The voice classifications results are shown in Tables 3, 4 and 5 where the confusion matrix is used to envision the performance. The maximum results of the model-, test- accuracy, TNR and TPR were found using the features group 3 (F3), as they reached 100% in the classification of all pathological cases. Another good accuracy was found when classifying the data as Hyperkinetic or normal voices using features group 4 (F4) in Table 4, also classifying the data as Hypokinetic or normal using features group 1 (F1) as seen in Table 4. Very good results were performed using the first feature group (F1) and fifth group (F5) in the case of classifying the data as Reflux or normal data as shown ion Table 5. From the table, it can be noticed that some of the features are consuming lower accuracy than others, for example the features group 4 (F4) gives test accuracy less than or equal to 50% in most of the cases.

In order to reveal the best feature combinations and to obtain the highest accuracy in classification, Table 6 shows the result using a combination of the best three feature groups at each pathological case found in Table 3. The first column in Table 6 shows the results of classifying the data as normal or hyperkinetic using delta-delta MFCC, skewness, and ZCR features together. It is found that the test-, train-accuracies, TNR and TPR are all 100% accuracy. The same results are found when classifying the data as hypokinetic or normal using delta-delta MFCC, skewness and delta-MFCC together. The delta-delta MFCC, entropy and delta-MFCC features combination are used in the third pathological case (Reflux) and the result accuracies were very good.

Table 3. Feature combinations and accuracy obtained for hyperkinetic pathology vs normal cases

	Normal F1 Hyper F1	Normal F2 Hyper F2	Normal F3 Hyper F3	Normal F4 Hyper F4	Normal F5 Hyper F5
Train Accuracy	58%	61%	100%	67%	55%
Test Accuracy	50%	65%	100%	70%	50%
TNR	57.89%	58.33%	100%	64.28%	0
TPR	58.03%	62.5%	100%	68.97%	55%

Table 4. Feature combinations and accuracy obtained for hypokinetic pathology vs normal cases

	Normal F1 Hypo F1	Normal F2 Hypo F2	Normal F3 Hypo F3	Normal F4 Hypo F4	Normal F5 Hypo F5
Train Accuracy	70.37%	70.37%	100%	64.19%	58.024%
Test Accuracy	80%	55%	100%	50%	40%
TNR	74.42%	68.42%	100%	68.18	59.01%
TPR	65.78%	75%	100%	59.45%	55%

Table 5. Feature combinations and accuracy obtained for reflux pathology vs normal cases

	Normal F1 Reflux F1	Normal F2 Reflux F2	Normal F3 Reflux F3	Normal F4 Reflux F4	Normal F5 Reflux F5
Train Accuracy	98.65%	66.22%	100%	67.57%	79.73%
Test Accuracy	100%	50%	100%	45%	80%
TNR	97.82%	68.52%	100%	69.09%	81.25%
TPR	100%	60%	100%	63.16%	76.92%

Table 6. Best feature combinations and accuracy obtained

	Normal/Hyper ΔΔMFCC+ Skewness+ ZCR	Normal/Hypo ΔΔMFCC+ Skewness+ ΔMFCC	Normal/Reflux ΔΔMFCC + Entropy+ΔMFCC
Train Accuracy	100%	100%	97.3%
Test Accuracy	100%	100%	100%
TNR	100%	100%	100%
TPR	100%	100%	93.54%

3.3. Classification using DWT features

Discrete wavelet transform (DWT) analysis is used to extract 5 features; one energy feature is extracted from each one of the five wavelet levels. Table 7 presents the accuracy obtained by Wavelet features. The Table 7 shows that the classification train accuracy (model accuracy) are ranged between 60% and 70% in all pathological cases, while test accuracies are ranged between 80% and 90% which are good results I in comparison with some of the results seen in Table 3. To the best of our knowledge, the DWT features are not widely used in voice classification field, but in this work, the results show that they are better than some of the MFCC and ZCR features which are widely used in this area. The Wavelet features present good results and could be used in classifying the voice as normal or pathological voices with good accuracy.

Table 7. Wavelet features and accuracy obtained

Wavelet features	Hyperkinetic	Hypokinetic	Reflux
Train Accuracy	70%	60.49%	67.57%
Test Accuracy	90%	80%	90%
TNR	74.19%	61.02%	66.67%
TPR	68.12%	59.09%	72.73%

4. DISCUSSION

Classification of pathological voices using machine learning has significant benefits for patient assessment and improvement computer-aided systems, as there are many previous researches in this field that apply various methods of feature extraction and classification algorithms. In this paper, the proposed method for detecting and classifying vocal disorder is compared with the methods found in previous studies [4]-[6], [25] which show that related voices could be classified into normal/pathological depends on sounds features and the accuracy of the classification algorithms. On the other hand, the result accuracy demonstrated in this research is shown to be superior to earlier researches. The accuracy rate was 99.56% in [4], the accuracy rate was 94.26% in [5], the accuracy rate was between 90-100% in [6], and the accuracy rate was 97.9% in [2]. However, in this study, the accuracy rate not only increased to around 100%, but also the methods presented are able to classify the related voices into four different classes (normal, hyperkinetic, hypokinetic, reflux), which is important in voice diseases diagnostic field.

5. CONCLUSION

This paper explores and compares several voice features extraction methods which are used to classify the voices as normal or pathological voices. Three different abnormal cases were studied: hyperkinetic, hypokinetic, and reflux. Three different methods were used to extract features which are: MFCC, ZCR, DWT and a related statistical feature are found using skewness, kurtosis, and entropy. The purpose of this work is to classify the voice dataset and compare the classification results using different feature groups where the classification process was repeated. The classification processes were all done using SVM and the train-, test- accuracies, TNR and TPR are calculated from the resultant confusion matrix in each case. The best classification results were reached using the feature group that includes delta-delta MFCC and skewness features, as it gave 100% accuracy in all cases. A combination of some of the delta-delta MFCC and ZCR features are also gave very good accuracy. The DWT features are not commonly used in voice classification, but in this paper, the results show that they are better than the delta-delta MFCC and ZCR features which are widely used in this area and can be used to classify the voice as normal or pathological with good accuracy. In the future research, other methods, or a combination of classification methods than SVM may be used to enhance the results where lower accuracies were found also classification of other diseases that cause temporary vocal impairments, such as COVID-19.

REFERENCES

[1] A. Visave, P. Kachare, A. Jeyakumar, A. N. Cheeran, and G. Bachher, "Vocal features for glottal pathology detection using BPNN," *International Journal of Computer Applications*, vol. 118, no. 17, pp. 1-6, 2015, doi: 10.5120/20834-3571.

[2] V. Sellam and J. Jagadeesan, "Classification of normal and pathological voice using SVM and RBFNN," *Journal of Signal and Information Processing*, vol. 5, no. 1, 2014, doi: 10.4236/jsip.2014.51001.

[3] D. Pravena, S. Dhivya, and A. D. Devi, "Pathological voice recognition for vocal fold disease," *International Journal of Computer Applications*, vol. 47, no. 13, pp. 31-37, 2012, doi:10.5120/7250-0314.

[4] H. Ankışhan, "A new approach for detection of pathological voice disorders with reduced parameters," *Electrica*, vol. 18, no. 1, pp. 60-71, 2018, doi: 10.5152/iujeee.2018.1810.

[5] S. H. Fang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, vol. 33, no. 5, pp. 634-641, 2019, doi: 10.1016/j.jvoice.2018.02.003.

[6] D. Panek, A. Skalski, J. Gajda, and R. Tadeusiewicz, "R Acoustic analysis assessment in speech pathology detection," *International Journal of Applied Mathematics and Computer Science*, vol. 25, no. 3, pp. 631-643, 2015, doi: 10.1515/amcs-2015-0046.
[7] U. Cesari, G. D. Pietro, E. Marciano, C. Niri, G. Sannino, and L. Verde, "A new database of healthy and pathological voices," *Computers and Electrical Engineering*, vol. 68, pp. 310-321, 2018, doi: 10.1016/j.compeleceng.2018.04.008.
[8] A. L. Goldberger *et al*., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101 , no. 23, pp. e215-e220, 2000, doi: 10.1161/01.cir.101.23.e215.
[9] T. Giannakopoulos and A. Pikrakis, "Introduction to audio analysis, a MATLAB approach," 1st ed., Academic Press, 2014.
[10] A. Sithara, A. Thomas, and D. Mathew, "Study of MFCC and IHC feature extraction methods with probabilistic acoustic models for speaker biometric applications," *Procedia computer science*, vol. 143, pp. 267-276, 2018, doi: 10.1016/j.procs.2018.10.395.
[11] O. K. Hamid, "Frame blocking and windowing speech signal," *Journal of Information, Communication, and Intelligence Systems*, vol. 4, no. 5, 2018.
[12] F. S. Cabral, H. Fukai, and S. Tamura, "Feature extraction methods proposed for speech recognition are effective on road condition monitoring using smartphone inertial sensors," *Sensors*, vol. 19, no. 16, pp. 3481, 2019, doi: 10.3390/s19163481.
[13] T. Grzywalski *et al.*, "Parameterization of Sequence of MFCCs for DNN-based voice disorder detection," *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 5247-5251, doi: 10.1109/BigData.2018.8622012.
[14] D. Prabakaran and S. Sriuppili, "Speech processing: MFCC based feature extraction techniques-an investigation," In *Journal of Physics: Conference Series*, vol. 1717, 2021, doi: 10.1088/1742-6596/1717/1/012009.
[15] A. Al Bashit and D. Valles, "MFCC-based houston toad call detection using LSTM," *2019 IEEE International Symposium on Measurement and Control in Robotics (ISMCR)*, 2019, pp. D3-3-1-D3-3-6, doi: 10.1109/ISMCR47492.2019.8955667.
[16] X. Li, M. Yao, and W. Huang, "Speech recognition based on k-means clustering and neural network ensembles," *2011 Seventh International Conference on Natural Computation*, 2011, pp. 614-617, doi: 10.1109/ICNC.2011.6022159.
[17] A. Papoulis and S. U. Pillai, "Probability, random variables, and stochastic processes," 4th ed., Tata McGraw-Hill Education, 2002.
[18] D. S. Shete, S. B. Patil, and S. B. Patil, "Zero crossing rate and energy of the speech signal of devanagari script," *IOSR-JVSP*, vol. 4, no. 1, pp. 1-5, 2014.
[19] G. Tzanetakis, G. Essi, and P. R. Cook, "Audio analysis using the discrete wavelet transform," In *Proceedings of the WSES International Conference Acoustics and Music: Theory and Applications (AMTA 2001)*, 2001.
[20] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *Journal of computing*, vol. 2, no. 3, 2010.
[21] S. Suthaharan, "Machine learning models and algorithms for big data classification," *Support vector machine*, Springer, pp. 207-235, 2016.
[22] B. Scholkopf, A. J. Smola, and F. Bach, "Learning with kernels: support vector machines, regularization, optimization, and beyond," The MIT Press., 2018.
[23] U. E. Akpudo and J. Hur, "Intelligent solenoid pump fault detection based on MFCC Features, LLE and SVM," *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2020, pp. 404-408, doi: 10.1109/ICAIIC48513.2020.9065282.
[24] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17 no. 1, pp. 168-192, 2021, doi: 10.1016/j.aci.2018.08.003.
[25] J. Dey, M. S. Bin Hossain and M. A. Haque, "An ensemble SVM-based approach for voice activity detection," *2018 10th Int. Conference on Electrical and Computer Engineering (ICECE)*, 2018, pp. 297-300, doi: 10.1109/ICECE.2018.8636745.
[26] B. Sabir, F. Rouda, Y. Khazri, B. Touri, and M. Moussetad, "Improved algorithm for pathological and normal voices identification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 1, pp. 238-243, doi: 10.11591/ijece.v7i1.pp238-243.

BIOGRAPHIES OF AUTHORS

Muneera Altayeb earned her BSc in Computer Engineering in 2007, and MSc in Communications Engineering from the University of Jordan in 2010. She is currently Assistant Dean of the Faculty Engineering/Al-Ahliyya Amman University (AAU) and she is a lecturer in the Department of Electronics and Communications Engineering at (AAU) since 2015. Her research interests are focused on the following areas: digital signals & image processing and machine learning. She can be contacted at email: m.altayeb@ammanu.edu.jo.

Amani Al-Ghraibah earned her BSc in Biomedical Engineering in 2008 from Jordan University of Science and Technology/Jordan. She got her Ph.D. and MSc in Electrical and Computer Engineering from New Mexico State University/USA in 2012 and 2015, respectively. She is currently an assistant professor in the Medical Engineering department at Al-Ahliyya Amman University. Her research interest is in the fields of digital signals & images processing and pattern recognition. She can be contacted at email: a.ghraibah@ammanu.edu.jo.