Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions

Ahmed Al-nasheri, Ghulam Muhammad, Mansour Alsulaiman, and Zulfiqar Ali

Digital Speech Processing Group, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
a.alnashari@yahoo.com, {ghulam, msuliman, zuali}@ksu.edu.sa

Summary

Objectives and background: Automatic voice-pathology detection and classification systems effectively contribute to the assessment of voice disorders, which helps clinicians to detect the existence of any voice pathologies and the type of pathology from which patients suffer in the early stages. This work concentrates on developing an accurate and robust feature extraction for detecting and classifying voice pathologies by investigating different frequency bands using correlation functions. In this paper, we extracted maximum peak values and their corresponding lag values from each frame of a voiced signal by using correlation functions as features to detect and classify pathological samples. These features are investigated in different frequency bands to see the contribution of each band on the detection and classification processes.

Material and Methods: Various samples of sustained vowel /a/ of normal and pathological voices were extracted from three different databases. The used database in this study represent three different languages: Arabic Voice Pathology Database (AVPD; Arabic), Saarbruecken Voice Database (SVD; German), and the Massachusetts Eye and Ear Infirmary (MEEI; English). A support vector machine (SVM) was used as a classifier. We also performed *t-test* to investigate the significant differences in mean of normal and pathological samples.

Results: The best achieved accuracies in both detection and classification were varied depending on the band, the correlation function, and the database. The most contributive bands in both detection and classification were between 1000 ~ 8000 Hz. In detection, the highest acquired accuracies when using cross correlation were 99.809%, 90.979%, and 91.168% in the MEEI, SVD, and APVD databases, respectively. However, in classification the highest acquired accuracies when using cross correlation were: 99.255%, 98.941%, and 95.188% in the three databases, respectively.

Keywords

Voice pathology detection; voice pathology classification; frequency investigation; Arabic Voice Pathology Database(AVPD), Saarbruecken Voice Database (SVD), Massachusetts Eye and Ear Infirmary (MEEI).

Introduction

Recently, lifestyle comes with an increased risk of pathological voice problems. About 25 percent of the population are engaged in works that are “vocally demanding.” For these individuals, either their jobs require excessive vocalization or their work environments force them to speak above a high noise level. Examples of professionals with heavy vocal demands include teachers, lawyers, auctioneers, aerobics instructors, singers, actors and manufacturing supervisors. As a consequence, working on digital processing of speech signals was found to provide a noninvasive analytical technique that is considered to be an effective assisting tool to medical doctors when identifying voice disorders specifically in their early stages. Voice pathologies affect the vocal folds, producing irregular vibrations due to the malfunctioning of many factors contributing to vocal vibrations. Vocal fold pathologies exhibit variations in the vibratory cycle of the vocal folds due to their incomplete closure. Voice disorders also affect the shape of the vocal tract (supra-glottal) and produce irregularities in spectral properties [1]. It is well known that there is no infralaryngeal (tracheobronchial tree) effect on the vocal tract during the production of a vowel if we consider that the voicing source has an infinite resistance. However, an accurate detailed analysis must realize that the infralaryngeal structures do an influence on the vocal tract, the articulatory configuration in the vocal tract interacts with the articulation in the vocal folds [38]. Upon on that, supplemental vocal tract-related information is predictable to help in detecting the characteristics of the vocal folds, essentially during phonation [39]. In addition, voice disorders affect vocal-fold vibration differently depending on the type of disorder and the location of the disease in the vocal folds, making them produce different basic tones. Vocal folds' vibration depends on several factors such as mucus present on the vocal folds tissue, stiffness, tension, muscles in the larynx, closing and opening of the folds, etc. These factors are affected differently for various voice pathologies. Due to the position and the size of the pathologies, vocal folds closing behaves differently during the vibration. Therefore, the vibration varies from one type of pathology to another. This vibration produces glottal source excitation frequencies, as well as affects the supra-glottal (the bottom part of the vocal tract) area, which in turn contributes to the frequency of the output voice signal.

The number of dysphonic patients has increased significantly, and in the United States alone approximately 7.5 million people have vocal difficulties [2]. It has been found that 15% of all visitors to King Abdul Aziz University Hospital, in Riyadh, Saudi Arabia, complain of a voice disorder [3]. The impact of voice problems on teaching professionals is significantly greater than for non-teaching professionals. Studies revealed that in the United States, the prevalence of voice disorders during a lifetime is 57.7% for teachers and 28.8% for non-teachers [4]. Approximately 33% of male and female teachers in the Riyadh area suffer from voice disorders [5]. The Communication and Swallowing Disorders Unit at King Abdul Aziz University Hospital examines a high volume of voice disorder cases (almost 760 cases per annum) in individuals with various professional and etiological backgrounds. The use of computers to detect or identify pathological problems in speech, a non-invasive method, is advancing over time. In the last decade, much research has been done on the automatic detection of vocal-fold disorders, which continues to require further investigation due to the lack of standard automatic diagnostic approaches/equipment for voice disorders. Detection of pathology is the first crucial step to diagnose and manage voice disorders correctly. Objective assessment, including acoustical analysis, is independent of human bias and can assist clinicians in making decisions. We firmly believe that clinicians have the final decision regarding medical diagnosis, and an objective assessment can only be used as an assistive tool. On the other hand, subjective measurement of voice quality is based on individual experience, which may vary. Automatic voice-pathology detection can be accomplished by various types of long-term and short-term signal analyses. Long-term parameters can be derived from the acoustic analysis [6], [7] of speech, and short-term parameters can be calculated using linear predictive coefficients [8], [9], linear predictive cepstral coefficients [10], Mel-frequency cepstral coefficients (MFCC) [11], [12], and so on. Different pattern-matching techniques, such as a Gaussian mixture model [13], [14], hidden Markov model [15], support vector machine (SVM) [16], artificial neural networks [17], and so on have been used to differentiate between disordered and normal samples. Multiple long-term acoustic features, namely pitch, shimmer, jitter, amplitude perturbation quotient (APQ), pitch perturbation quotient, harmonic-to-noise ratio, normalized noise energy, voice-turbulence index, soft-phonation index (SPI), frequency amplitude tremor, and glottal-to-noise excitation ratio are frequently used to diagnose voice pathology (referenced in [14] as [2]-[12]). Furthermore, jitter and shimmer capture vocal-fold vibratory characteristics for both pathological and normal people, and both parameters are widely used for clinical research purposes [18]. Seven acoustic parameters, including shimmer and jitter, are extracted by means of an iterative residual-signal estimator in Rosa et al. [19], and jitter provided 54.8% accuracy of detection for 21 pathologies. Thirty-three different long-term acoustic parameters with their definitions, derived from the Multi-Dimensional Voice Program (MDVP) [20], are listed in Arjmandi et al. [21]. Twenty-two acoustic parameters were selected from the list extracted from voice samples in the Massachusetts Eye and Ear Infirmary (MEEI) database. Fifty

dysphonic patients and 50 normal persons were used for detection. The 22 parameters were calculated for each sample and fed to six different classifiers to compare their accuracies. Two feature-reduction techniques were also used before applying classification methods. Binary classifier SVM showed the best results compared with other classifiers, with a recognition rate of 94.26%. In Wang et al. [22], MFCC and six acoustic parameters (jitter, shimmer, NHR, SPI, APQ, and Relative Average Perturbation) were extracted, with the results compared with those of the NN-based voice pathology detection system [23]. Sáenz-Lechón et al. compared their proposed parameters based on wavelet transform with some of the MDVP parameters to discriminate between pathological and normal voices [24]. To ensure the reliability of the acoustic MDVP parameters, some of them were compared with the same parameters extracted using Praat; results showed no significant difference between the two computer software approaches [25]. Recently, MPEG-7 audio descriptors and multi-directional, regression-based features have been used in voice-pathology detection, with good accuracy [26, 27]. Another recent study investigated the most discriminative frequency region for voice-pathology detection [28].

Correlation functions are considered as one of the common methods for extracting various characteristics from speech signals. They are known as a domain that has certain good properties that can be used as features. The methods based on correlation function applied on a short section of voice signal can provide substantial information that enables us to estimate the vocal tract transfer function. For example, these methods result in many peak values with periodicity as the same periodicity as of the input signal. Therefore, to examine the periodicity of the signal, it is common to examine its autocorrelation function. This indicates that the correlation function of a periodic signal is also periodic. Consequently, finding pitch, fundamental frequency, etc. of the signal will be possible by using these methods. In many researches it is observed that the normal voice has more periodicity than the pathological one, and therefore performing correlation functions on these types of classes will provide an excellent allusion that can be used to discriminate between normal and pathological voices. For instance, Von Leden, Moore, and Timke observed that the pathological samples have a strong tendency for frequent and rapid changes in the regularity [29]. In addition, Lieberman found that pathological voices tend to show unusually large cycle-to-cycle fluctuations in the fundamental period [30]. In this work, we performed different forms of correlation functions such as autocorrelation on the signal itself frame by frame, cross correlation between two successive frames in the same signal, and cross correlation between two successive filters frame by frame. It is preferable to use a short segment of the voice signal instead of the whole signal, because the noise tends to be cancelled out in the autocorrelation process in this short segment [31].

As we observe, every voice disorder produces different frequencies depending on the type of voice disorder and its location on the vocal fold, as we described before. Consequently, observing the frequency band is

very important to see which frequency band contributes more to the detection and classification of voice disorders. For instance, in [37] the authors found that the lower frequencies between 0 ~ 3000 Hz are more suitable for discriminating dysphonic voices than the higher frequencies. In addition, Fraile et al. in [36] found that the power in bands between 2000 and 6400 Hz is significantly less stable in dysphonic voices.

In this paper, we mainly focus on developing a computationally less expensive method for voice pathology detection. Specifically, we concentrate on extracting a feature set having low dimension. In the proposed method, the input voice is passed through a bank of band pass filters, and each filter output is divided into overlapping blocks. Correlation functions are applied to extract peak and lag to be stored as features. In order to detect and classify voice pathology, the proposed method is evaluated using three different databases that have three voice disorders in common: (i) the MEEI [32]; (ii) the Saarbruecken Voice Database (SVD) [33]; and (iii) the Arabic Voice Pathology Database (AVPD).

Materials and Methods

Data

In this study, we used three different databases (MEEI, SVD, and AVPD), and we chose only three types of pathological voices — (1) vocal fold cyst; (2) unilateral vocal fold paralysis; and (3) vocal fold polyp — because only these pathologies are common in all three databases. The number of samples in each database is shown in Table 1, where the numbers of male and female speakers are shown, respectively, inside parentheses. The three used databases are each described below.

Table 1: Normal and pathological samples from three different databases.

Database	**Normal**	**Pathological**			
		Cysts	**Paralysis**	**Polyp**	**Total**
AVPD	169 (102,67)	25 (8, 17)	56 (35, 31)	46 (26, 20)	127
MEEI	53 (19, 34)	10 (6, 4)	71 (39, 32)	20 (11, 9)	101
SVD	266 (130, 136)	6 (1, 5)	212 (73, 139)	45 (26, 19)	263

Massachusetts Eye and Ear Infirmary (MEEI) Voice Disorder Database

This database was developed by the MEEI Voice and Speech Lab and includes more than 1,400 voiced samples of the sustained vowel /a/ and the first part of the Rainbow Passage. It is commercialized by Kay Elemetrics [32] and was recorded in two different environments. The sampling frequency for normal samples was 50 kHz, while that of the pathological samples was 25 kHz or 50 kHz. It is used in most studies of voice-pathology detection and classification even though it has many disadvantages, such as the different environments and sample frequencies used to record normal and pathological voices. In this database, many tools were used to evaluate voice condition, including stroboscopy, acoustic aerodynamic measures, and a physical examination of the neck and mouth (this information is provided by Kay Elemetrics). In the CD Kay Elemetrics provides, we filtered the filenames according to the three diseases; if there were multiple pathologies for a file, we ignored that file. For normal speakers, we chose all available 53 samples. We selected only sustained vowel /a/ samples.

Saarbruecken Voice Database (SVD)

The SVD is a freely downloadable database [33], recorded by the Institute of Phonetics of Saarland University. This database contains sustained vowels /a/, /i/, and /u/ with different intonations (normal, low, high, and low-high-low), along with a spoken sentence in German “Guten Morgen, wie geht es Ihnen?” which translates into English as “Good morning, how are you?” These attributes make it a good database for researchers to conduct experiments. All recorded voices in the SVD database were sampled at 50 kHz with 16-bit resolution. This database is new, and thus very few studies of voice-pathology detection have been done using it. We downloaded the files from the website mentioned in [33] using the criteria of the three diseases. We selected only sustained vowel /a/ samples produced at normal pitch.

Arabic Voice Pathology Database (AVPD)

The voice and speech samples in this database were collected in different sessions at the Communication and Swallowing Disorders Unit [3] of King Abdul Aziz University Hospital in Riyadh, Saudi Arabia, by experienced phoneticians in a sound-treated room using a standardized recording protocol. This database collection was one of the major tasks of the ongoing project funded by the National Plans for Science and Technology, Saudi Arabia, over the duration of two years. The protocol of the database was designed to avoid the various shortcomings of the MEEI database [24]. AVPD database has recordings of sustained vowels as well as the speech of patients who have vocal-fold pathologies, along with the same recordings

of persons with normal speech. Normal and pathological vocal folds were determined after clinical assessment using a laryngeal stroboscope. In case of pathology, the perceptual severity of voice disorders was rated on a scale of 1–3, where 3 represents the most severe case. A severity rating was associated with each sample based upon the consensus of a panel of three expert medical doctors. The recording has different types of texts: (1) three sustained vowels with onset and offset information; (2) isolated words including Arabic digits and some other common words; and (3) continuous speech. The selected text was carefully chosen to cover all Arabic phonemes. All speakers recorded three utterances of each vowel /a/, /u/, and /i/, while isolated words and continuous speech were recorded once to avoid burdening patients. The sampling frequency in the database was 50 kHz, and the speech was recorded using the computerized speech lab program. The voice disorders recorded in this database were evaluated and validated by different specialist doctors at King Abdul Aziz University Hospital. Among the recorded samples, only recordings of patients with vocal-fold cysts, vocal-fold polyps, and unilateral vocal-fold paralysis pathologies were included in this study. We selected only sustained vowel /a/ samples.

Proposed Method

The main aim of the current study is to extract robust and reliable features that can contribute to a detection and classification of voice pathologies, and to investigate the effect of different frequency regions (bands) on the detection and classification processes using these features. We used a correlation function to extract the peaks and their corresponding lag values from the voiced signals to represent the features, which are used to discriminate between normal and pathological classes. As we can see from Figures 1 and 2, which illustrate the proposed method, the voice signal is fed to a filter bank, which is composed of eight band pass filters.

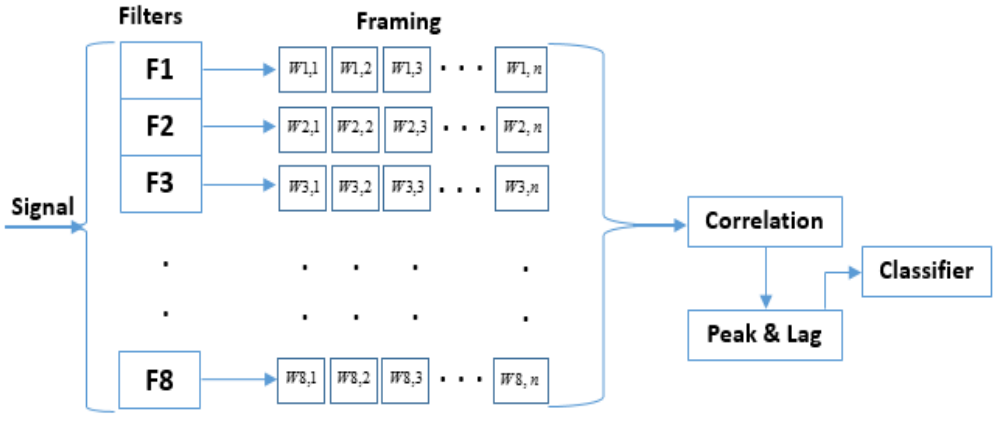

Figure 1: Block diagram of the proposed method

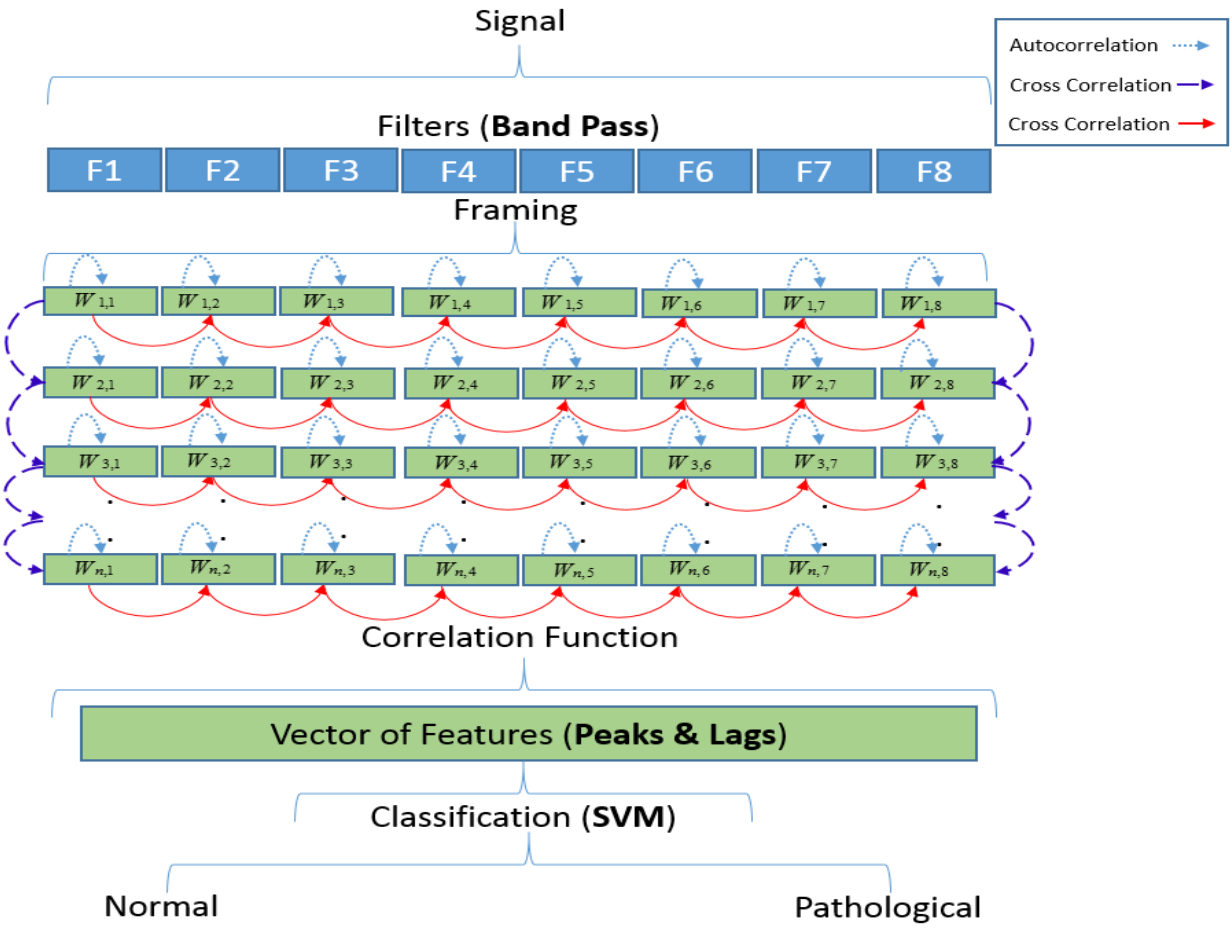

Figure 2: Detailed block diagram of the proposed method

These filters represent the band pass of finite impulse response (FIR) filters with center frequencies spaced on an octave scale. The center frequencies are 31.25, 93.75, 187.5, 375, 750, 1.5K, 3K, 6K, and 10K Hz. Although we also performed experiments with the Mel scale, the octave scale showed better performance in detection and classification. The reason behind using filter banks is to analyze the voiced signal in different frequency bands. The filter bands are slightly overlapping, and their frequency magnitude responses are shown in Figure 3. The output of each filter is divided into frames with a specific size of 40 ms with an overlap of 50% (20 ms). Extracting peak and lag values can be achieved by applying different forms of the correlation function. For instance, (1) the autocorrelation function was applied frame by frame in the same filter, (2) cross correlation was applied between two successive filters' frames, and (3) cross correlation was applied between two successive frames in the same filter. In each form of the correlation function, we chose the maximum peak value and its corresponding lag. The proposed method will be described below.

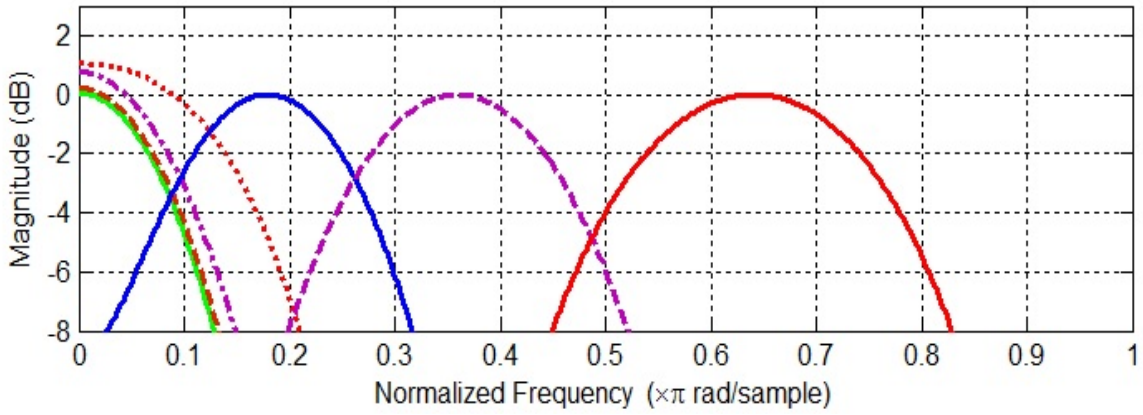

Figure 3: Frequency responses of the eight FIR filters used in the proposed method

Correlation Function

As illustrated in Figures 1 and 2, the correlation function in different forms is applied individually frame by frame and then the highest peak value (after 2 ms lag, because we assume that the first harmonic appears after that duration) and its corresponding lag are taken to represent finally a feature vector. The idea behind using autocorrelation is that for a normal sustained voice the peak will be high and the lag value will be inside the first half of the autocorrelation. Moreover, the normal sustained voice will be more harmonic if it is compared with the pathological sustained voice, and therefore using the correlation function is beneficial. From these two reasons, we believe that the peaks and lags will contribute in discriminating between the normal and pathological samples. The autocorrelation (*AC*) function of a signal (*s*) in a frame can be computed as follows:

$$AC(\tau) = \sum_{n=0}^{N-\tau-1} s(n)s(n+\tau) \qquad (1)$$

Where $0 \le \tau \le L-1$, L is the maximum lag value, N is the number of samples in a frame, and τ is the lag. The cross correlation between two successive frames (*cross frame correlation*: *CFC*) can be computed as follows:

$$CFC(\tau) = \sum_{n=0}^{N-\tau-1} s_i(n)s_{i+1}(n+\tau) \qquad (2)$$

Where *i* represents the frame number.

In addition, the cross correlation between two successive filters (*filter cross correlation*: *FCC*) can be computed as follows:

$$FCC(\tau) = \sum_{n=0}^{N-\tau-1} s(n)g(n+\tau) \qquad (3)$$

Where *s(n)* and *g(n)* represent the frames in filter *j* and filter *j*+1, respectively.

Setup of the Experiments

First, for every database we down sampled the selected sustained vowel /a/ samples to 25 kHz to ensure the same sampling frequency for all the samples. Second, we performed three experiments on each database individually to extract the peaks and their corresponding lags depending on the type of correlation function. When the features for each database were ready, we performed 100 (=36+36+28) experiments on each database to detect pathology based on individual filters and on their combination. In case of autocorrelation and cross frames correlation, 36 experiments were performed for each case: eight experiments for each filter and the rest for the combination between filters by combining two successive filters, three successive filters, and so on until eight combined filters were achieved. However, in case of cross correlation between two successive filters, we performed 28 experiments with the same previous scenario. These experiments were performed by using two features (peak *and* lag) and they were repeated on an individual basis (peak *or* lag) to see the contribution of each feature separately on the detection process. In addition to these experiments, we performed 21 experiments on each database for the classification process by choosing the best different cases that achieved the best detection accuracy from cross correlation between successive filters (*FCC*), because this function better performed than the other two functions. We only used peak *and* lag values to perform the experiments of the classification. Finally, to obtain the *p-value*, we performed the *t-test* between two classes of normal and pathological samples on the three different databases separately with the following null hypothesis: “there is no significant difference between the two classes.”

Results

The results of the performed experiments for pathology detection and classification are expressed in terms of accuracy (ACC: the ratio between correctly detected samples and the total number of samples), sensitivity (SN: the proportion of pathological samples that are positively identified), specificity (SP: the proportion of normal samples that are negatively identified), and the area under the Receiver Operating Characteristic (ROC) curve, called the Area Under Curve. These terms can be calculated using the following distinct equations:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

$$SN = \frac{TP}{TP + FN} \qquad (5)$$

$$SP = \frac{TN}{FP + TN} \qquad (6)$$

where true negative (*TN*) means that the system detects a normal subject as a normal subject, true positive (*TP*) means that the system detects a pathological subject as a pathological subject, false negative (*FN*) means that the system detects a pathological subject as a normal subject, and false positive (*FP*) means that the system detects a normal subject as a pathological subject.

To verify the validity of the extracted features from the three different databases in the detection and classification process of pathological and normal samples, various experiments were performed. To ensure accuracy, various experiments were performed individually for each filter and their combination (10 folds and 100 experiments on each database, which equals 1000 runs). The achieved accuracies varied from one database to another with the same correlation function that was used to extract the features. Moreover, the obtained accuracies in the same database were also varied depending on the number of features that were used to carry the experiments. Two features were used in each filter, but this number is increased in case of the combination between filters (number of combined filters multiplied by two). In case of the detection process, Table 2 shows the best achieved accuracies on each database by using three different correlation functions. As we can see from Table 2, the accuracies vary from one database to another and these accuracies are varied in the same database depending on the used correlation function. For example, the achieved accuracies in case of autocorrelation were different than those in case of cross filter correlation and cross frame correlation. In general, as it is shown in Table 2, the highest acquired accuracies are 99.809%, 93.85%, and 90.979% in the three databases MEEI, AVPD, and SVD, respectively. To see the performance of each individual filter and the combined filters on the detection and classification, various experiments were performed, but we report the best individual filter and best combination that have the

highest results. For instance, Table 3 shows the best obtained accuracies by performing the experiments on an individual filter, and on combined successive number of filters in case of using the autocorrelation function.

Table 2: Best detection accuracies in the three different databases using various correlation functions

Methods	Database	Accuracies %		
		SN	SP	ACC
Autocorrelation	MEEI	87.511	93.090	99.673
	SVD	88.690	88.707	88.696
	AVPD	90.324	93.052	91.690
Cross correlation between filters	MEEI	98.571	99.545	**99.809**
	SVD	90.201	91.720	**90.979**
	AVPD	85.552	94.796	91.168
Cross correlation between frames	MEEI	87.511	93.090	99.736
	SVD	84.723	88.101	86.413
	AVPD	93.463	94.237	**93.850**

Moreover, Table 4 also shows the highest acquired results from the experiments that use cross correlation between two successive filters to extract the features from the mentioned databases. The highest achieved accuracies in this case are 99.81%, 91.17%, and 90.98% in the MEEI, AVPD, and SVD databases, respectively. These accuracies were achieved when we combined more than one filter as shown in this table.

Table 3: Best accuracy on different filter number from the three used databases using autocorrelation function to extract the features

Number of	Best Filter performance	Best Accuracy %

	MEEI	SVD	AVPD	MEEI	SVD	AVPD
1	(6)	(6)	(2)	98.056	66.332	81.805
2	(4,5)	(3,4)	(6,7)	99.155	75.512	84.824
3	(5,6,7)	(5,6,7)	(5,6,7)	99.466	82.619	89.261
4	(5,6,7,8)	(4,5,6,7)	(4,5,6,7)	99.442	84.550	91.043
5	(4,5,6,7,8)	(4,5,6,7,8)	(3,4,5,6,7)	99.554	86.784	89.728
6	(3,4,5,6,7,8)	(3,4,5,6,7,8)	(2,3,4,5,6,7)	99.570	86.409	90.742
7	(2,3,4,5,6,7,8)	(2,3,4,5,6,7,8)	(1,2,3,4,5,6,7)	99.554	87.242	**91.690**
8	(1~8)	(1~8)	(1~8)	**99.673**	**88.696**	90.295

Table 4: Best accuracies on different filter number from the three used database using cross correlation function between two successive filters to extract the features

Number of filter(s)	Best Filter performance			Best Accuracy %		
	MEEI	SVD	AVPD	MEEI	SVD	AVPD
1	(5)	(6)	(3)	97.124	67.738	74.552
2	(5,6)	(5,6)	(5,6)	99.394	82.001	83.290
3	(5,6,7)	(4,5,6)	(5,6,7)	99.753	85.692	88.483
4	(4,5,6,7)	(4,5,6,7)	(3,4,5,6)	**99.809**	90.025	90.490
5	(3,4,5,6,7)	(3,4,5,6,7)	(3,4,5,6,7)	99.777	**90.979**	**91.168**
6	(1,2,3,4,5,6)	(2,3,4,5,6,7)	(2,3,4,5,6,7)	99.562	90.856	90.914
7	(1~7)	(1~7)	(1~7)	99.753	89.519	91.038

In addition, Table 5 also shows the best obtained results from the experiments that use cross correlation between two successive frames in the same filter to extract the features from these databases. As we can see from that table, the highest attained accuracies are 99.74%, 93.85%, and 86.41% in the MEEI, AVPD, and SVD databases, respectively. These accuracies were obtained in case of the combined filters too.

Table 5: Best accuracies on different filter numbers from the three used databases using cross correlation between the two successive frames to extract the features

Number of filter(s)	Best Filter Performance			Best Accuracy (%)		
	MEEI	**SVD**	**AVPD**	**MEEI**	**SVD**	**AVPD**
1	(6)	(7)	(1)	98.932	70.460	76.234
2	(6,7)	(5,6)	(1,2)	99.704	77.447	86.825
3	(5,6,7)	(4,5,6)	(5,6,7)	99.598	81.597	88.807
4	(2,3,4,5)	(5,6,7,8)	(4,5,6,7)	99.704	85.108	91.205
5	(2,3,4,5,6)	(4,5,6,7,8)	(3,4,5,6,7)	99.715	85.329	91.726
6	(2,3,4,5,6,7)	(2,3,4,5,6,7)	(2,3,4,5,6,7)	99.693	85.535	92.288
7	(1,2,3,4,5,6,7)	(2,3,4,5,6,7,8)	(1,2,3,4,5,6,7)	99.567	**86.413**	**93.850**
8	(1~8)	(1~8)	(1~8)	**99.736**	83.394	93.446

To see the effect of each feature (peak *or* lag) on the detection process, seven additional experiments were performed using one type of feature separately. These experiments were only performed using cross filter correlation to extract features. Table 6 shows the obtained accuracies for each individual feature, and for them together. As we can see from this table, each feature contributes differently than the other one. The highest obtained detection accuracy for peak are 75.79%, 98.717%, and 73.131% in SVD, MEEI, and AVPD respectively while the highest obtained detection accuracy in case of lag are 78.78%, 92.996%, and 75.863% in the same mentioned databases.

Table 6: Best detection accuracies for different number of filters using cross correlation between two successive filters

Number of Filters	SVD Database Accuracies			MEEI Database Accuracies			AVPD Database Accuracies		
	Both	**Peak**	**Lag**	**Both**	**Peak**	**Lag**	**Both**	**Peak**	**Lag**
1-(6)	67.738	56.564	60.104	89.960	89.203	73.052	68.683	64.396	70.195
2-(5,6)	82.001	59.959	70.481	97.602	95.211	80.757	83.290	68.628	72.036
3-(4,5,6)	85.692	67.804	72.319	99.602	98.279	84.940	87.844	69.093	72.507
4-(4,5,6,7)	90.025	**75.793**	77.152	99.737	**98.717**	92.112	90.417	71.068	74.496
5-(3,4,5,6,7)	90.979	75.787	77.269	99.793	98.669	**92.996**	91.168	72.954	74.700
6-(2,3,4,5,6,7)	90.856	74.581	79.106	99.817	98.598	92.829	90.914	72.807	74.929
7-(1,2,3,4,5,6,7)	89.519	74.531	**78.782**	99.825	98.534	92.940	91.038	**73.131**	**75.863**

In case of the classification process, we chose the best acquired accuracies in case of the detection process shown in Table 4 that belongs to the cross filter correlation function between two successive filters' frames (features contain both peak and lag), then we performed three different experiments on each database individually depending on the classification type. In this case, 21 experiments were performed on each database with different types of classification. Table 7 shows the achieved accuracies of classification on these databases by using the cross correlation function to extract the features. The obtained accuracies in this case vary from one classification type to another in the same database. The highest attained accuracies are 99.255%, 98.941, and 95.188% in the MEEI, SVD, and AVPD databases, respectively, for cyst vs others.

Table 7: Best accuracies for classification on the three used database

Classification type	**Number of filter(s)**	**Databases Accuracy (%)**		
		MEEI	**SVD**	**AVPD**
cyst vs (Paralysis & Polyp)	1	92.836	98.379	81.768
	2	98.737	98.379	89.266
	3	**99.255**	98.379	94.344
	4	98.716	98.379	93.495
	5	98.323	98.818	**95.188**
	6	98.737	98.721	93.720
	7	98.716	**98.941**	94.610
Paralysis vs (Cyst & Polyp)	1	78.820	81.770	66.667
	2	94.658	87.306	86.083
	3	**97.288**	91.447	88.642
	4	96.605	91.867	89.376
	5	95.424	92.222	89.936
	6	96.915	92.138	89.977
	7	96.170	**92.255**	**90.139**
Polyp vs (Cyst & Paralysis)	1	84.265	83.269	68.764
	2	95.528	87.371	87.343
	3	96.998	92.158	90.121
	4	96.211	**92.339**	**90.786**
	5	96.542	91.156	90.445
	6	96.874	90.233	90.497
	7	**97.081**	92.080	89.440

Finally, we performed a *t-test* between two classes of normal and pathological samples on each database separately and computed the *p-values* of the two extracted features (peak and lag) for each class. In our

work, the *p-values* probability are computed for the extracted features where three correlation functions are used for each database to extract them. For example, Table 8 shows the *p-value* for the autocorrelation function which was performed on the three databases for each individual filter.

Table 8: p-values for the extracted peak and lag using autocorrelation function from the three different databases.

Filter No.	Databases					
	MEEI		SVD		AVPD	
	Peaks	Lags	Peaks	Lags	Peaks	Lags
1	0	1.667E-304	4.1688E-39	0.36516378	8.9921E-14	0
2	0	1.445E-304	1.159E-39	0.29408271	2.0607E-13	0
3	0	6.618E-303	4.9559E-42	0.30095416	4.355E-12	0
4	0	1.119E-302	8.3193E-52	0.26644587	2.1886E-07	0
5	0	2.017E-292	4.267E-104	0.20139113	0.00824899	0
6	0	2.452E-219	0	0.0537987	2.255E-250	0
7	0	2.386E-29	3.559E-101	6.8811E-55	0.52072373	0
8	1.9034E-77	0.15735378	4.699E-142	9.7787E-31	9.4469E-88	4.0865E-69

However, Table 9 shows the *p-value* for the cross correlation function for each filter, where filter one represents the cross correlation between filters one and two, filter two represents filters two and three, and so on until filter seven represents the cross correlation between filters seven and eight.

Table 9: p-values for the extracted peak and lag using cross correlation function between two successive filter from the three different databases.

Filter No.	Databases					
	MEEI		SVD		AVPD	
	Peaks	Lags	Peaks	Lags	Peaks	Lags
1,2	0	3.196E-304	7.8997E-37	0.0038409	0.09947441	8.853E-152
2,3	0	2.545E-304	4.4936E-38	0.00561598	0.05795023	3.158E-190
3,4	0	5.359E-303	1.791E-43	0.00607769	0.00375946	9.69E-196
4,5	0	1.373E-297	6.4922E-69	0.00983982	9.7901E-13	9.948E-196
5,6	0	1.085E-252	8.098E-221	0.01998128	1.2E-141	2.213E-198
6,7	0	6.276E-102	1.98E-244	2.295E-159	3.934E-123	0.01530406
7,8	0.85627226	3.0194E-27	2.285E-123	1.8356E-81	5.501E-93	1.0664E-29

In addition, Table 10 shows the *p-values* for the cross correlation between two successive frames in the same filter.

Table 10: p-values for the extracted peak and lag using cross correlation function between two successive frames from the three different databases.

Filter No.	Databases					
	MEEI		SVD		AVPD	
	Peaks	Lags	Peaks	Lags	Peaks	Lags
1	0	0	8.8838E-44	1.6175E-12	7.4443E-29	0
2	0	0	2.3235E-44	2.5858E-53	2.4358E-28	0
3	0	0	1.0088E-46	1.6688E-53	2.0378E-26	0
4	0	0	5.3195E-57	1.8255E-55	3.3386E-19	1.119E-303
5	0	0	1.871E-111	2.954E-64	0.18796485	1.355E-234
6	0	0	0	2.232E-121	1.324E-203	1.217E-247
7	0	0	2.66E-29	0	5.1327E-25	5.431E-09
8	3.046E-99	0	1.74E-157	1.164E-252	1.332E-99	3.409E-255

We performed multiple *t-test* on the three databases with the three different correlation functions that used to extract features to see the individual performance of these correlation functions in discriminating between normal and pathological samples on each database. As we can see from tables 8, 9, and 10 the ability of discriminating between normal and pathological subjects are varies for each database from one correlation function to another. For example, in case of MEEI database the contribution of peak and lag in the three methods are much closer if we compare their contribution to the contribution of peak and lag on the other two databases with the three correlation functions. The lowest contribution for peak and lag in the discrimination between normal and pathological subjects is occurred in case of using SVD database with the three correlation functions. From these multiple t-test we can infer that whether these correlation function will perform well or not.

In this study, we performed an additional experiment of pathology detection using the MEEI subset consisting of 53 normal samples and 173 pathological samples. By comparing the results, we conclude that the proposed method is robust against the sample size. Table 11 shows the obtained accuracies for the three correlation function with different samples size. For example the difference between the two obtained accuracies in case of autocorrelation is 0.048 which represents very small values. Which indicate that the sample size does not affect the robustness of the proposed method.

Table 11: The obtained accuracies from the three different correlation function on MEEI database using different samples.

Correlation Functions	Samples (53 N - 173 P)	Samples (53 N - 101 P)
Autocorrelation	99.625	99.673
Cross Filter correlation	99.750	99.753
Cross Frames Correlation	99.075	99.736

Finally, we performed different experiments using Mel filter bank for the eight filters individually and for the eight combined filters. Table 12 shows the obtained accuracies for each filter and for the combined filters for the octave and Mel scales. As we can see from this table, generally, the performance of octave scale is better than Mel scales in term of the obtained accuracies. In addition, it can be inferred that the proposed method is robust and reliable because of the use of different type of scales did not affect the obtained accuracies.

Table 12: The obtained accuracies for each filter with two different scales

Scale	Filters								
	1	2	3	4	5	6	7	8	1 ~ 8
Octave	95.649	95.625	95.721	94.861	95.960	98.056	90.127	78.685	99.673
Mel	95.697	96.112	96.143	96.677	97.355	97.418	90.191	82.319	99.474

Figure 4-(a) shows the ROC curve of the highest achieved detection accuracy in case of using autocorrelation to extract the feature from the three databases. It demonstrates that the best performance is obtained with the features extracted from the MEEI. In addition, Figure 4-(b) shows the ROC curve of the highest achieved detection accuracy in case of using cross frame correlation to extract the feature from the three databases. As we can see from that figure, the best performance was obtained with the feature extracted from the MEEI.

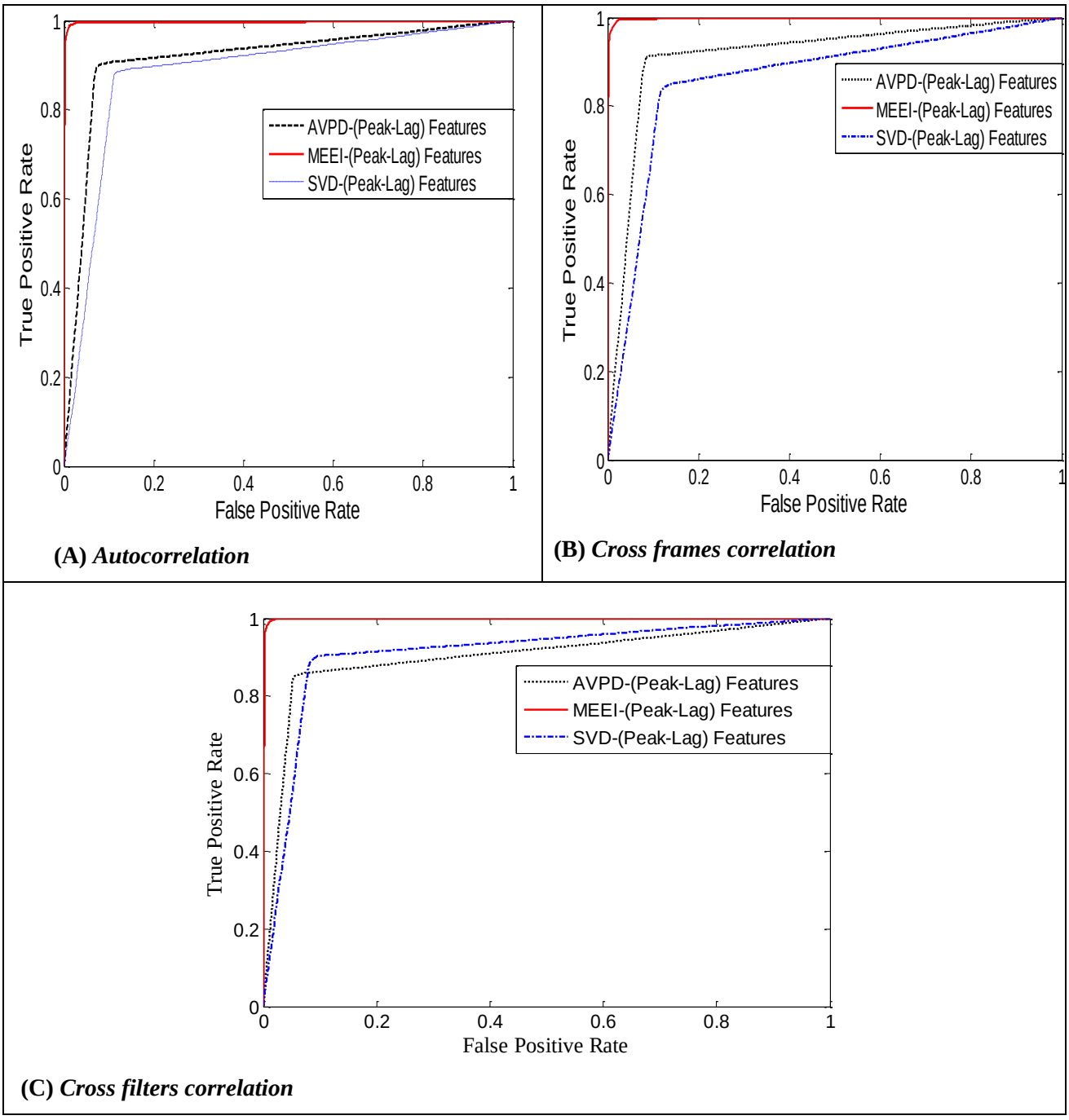

Figure 4: Best accuracies for features extracted from the three databases

Finally, Figure 4-(c) shows the ROC curve of the highest achieved detection accuracy in case of using the cross filters correlation to extract the feature from the three databases. Also, the best performance was obtained with the feature extracted from the MEEI. In all ROC curves mentioned, the 95% confidence interval is 0.9449-0.9870, and the 1-tail *p-value* is zero (<0.05) describing the significance of the data in the two classes.

Discussion

We investigated the extracted features using different correlation functions on different frequency bands for voice pathology detection and classification. Based on the obtained result, mentioned above, we can infer that the variation in the achieved accuracies in the same database is referred to the type of correlation function that was used to extract the features. The reason behind this variation is that every method has different values in performing the operation of the correlation. For example, in case of autocorrelation it is performed frame by frame in the same filter, while in case of cross frame correlation it is performed on two successive frames in the same filter. In addition, in case of cross filter correlation, it is performed in frame one from filter one with frame one from filter two and so on (two successive filters). This variation of the accuracies in the three different databases may be caused by different reasons: (1) the severity of voice disorders, which are not the same between the three databases, as shown, for instance, in Table 2, where sensitivity (to pathological samples) varies from one database to another; (2) the recording environment and the regulation of the recording are not the same between the three databases; (3) in the case of the MEEI database, the recording environments for pathological and normal samples were not the same; and (4) the number of samples taken from each database in this study are not the same. Besides that, the achieved accuracies vary within the same database depending on the correlation function used to extract the features due to the performed calculations, which are different from one method to another. Moreover, the variation in the accuracies in the same database also were different from one filter to another as a result of the fact that the frequency bands of each filter were not the same, which indicates that every frequency band has a different contribution to the detection and classification of pathologies. Figure 5 reflects this variation in the accuracies in case of the extracted feature from the SVD database using the three forms of the correlation functions. As it is seen from that figure, the greatest contribution for detection and classification is achieved in case of filters 4, 5, 6, and 7. This also confirms the findings of Fraile *et al.* [36], which state that "power in bands between 2000 and 6400 Hz is significantly less stable in dysphonic voices." Besides that, as we notice from the experimental result, the highest achieved accuracies occurred in the combined filters in all experiments, because each filter has the ability to detect some components in the specified range of frequencies components than the other filter, and when we combine more than one filter together their frequency range is expanded, which leads to higher accuracy of detection and classification.

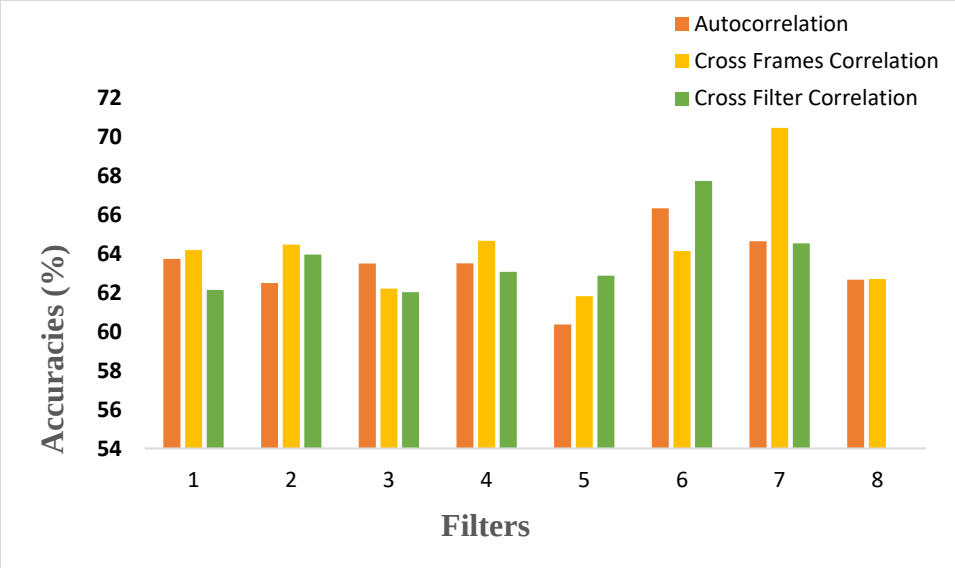

Figure 5: Three correlation functions applied to the SVD database

In addition, the calculated *p-values* for peaks and lags shown in Tables 7, 8, and 9 indicate the contribution of each feature to discriminating between normality and pathology. The contribution of peaks and lags separately for detection varies between the databases. Peaks have a more positive contribution in the MEEI database and the SVD in most cases, whilst lags have a more positive contribution in the AVPD. They performed very well in all the databases in case of their combination. For instance, Table 10 shows the best achieved accuracies from a different number of filters in the SVD database. It can be seen that the individual performance for peaks and lags is less when compared with them together. As we previously mentioned, each individual feature has a contribution to the detection and classification processes, but this varies from low to high and when we found it to be very low we stopped performing the experiment on this feature and continued doing the experiments with the other feature that had a high contribution. We made this decision depending on the obtained result shown in Tables 8, 9, and 10. From the results of all experiments, we found the following:

- Every extracted feature has a contribution to the classification and detection processes, but the two features together performed better than the individual feature.
- The performance of each frequency band varied from one to another and the best performance was in the bands of frequency range 1000 ~ 8000 Hz.
- The combined filters performed better than the individual filter in both processes (detection and classification).

- There is a need for a mixed database experiment between the three databases to ensure the independency of features from the database that were used to extract features. This is a point to investigate as a future work.
- The severity of voice disorder did not addressed in this study and it was left for future work.
- The proposed method is not dependent on the recording environment, because it achieved high accuracies in the MEEI database (where the recording environments are different), and the SVD and the AVPD (where the recording environments were the same per database).
- The proposed method is robust against the sample size.
- Both the peaks and the lags are used (to get the highest accuracy), the final system is independent of choosing a particular one.

In general, our results are better than, or comparable with, other reported accuracies using the MEEI and SVD databases in different studies. For example, the reported results in [21] were 94.26% for detection and 91.55% for classification, where Arjmandi et al. used the MEEI database and the same classifier that we used, but different pathological samples. Moreover, in [14] Godino et al. used the MEEI database and the achieved accuracy in this study was 94.07%, while in [34] Martinez et al. used the SVD database and the same classifier that we used and the attained accuracy in this study was 81%. They also used the MEEI database and the achieved accuracy was 94.80%. In addition, Markaki et al. in [35] used the MEEI database and the achieved classification accuracy was 94.10%. Further, in [26] Muhammad et al. used MPEG features to detect and classify voice pathology using the MEEI database and the SVM classifier, obtaining 99.994% with 45 features, and 99.412% accuracy with only seven features. In this study the authors used the same database and classifier that we used, but with a different number of samples and more features. In our study the highest achieved accuracies in detection were 99.809%, 90.979%, and 91.168% in the MEEI, SVD, and APVD databases, respectively. However, in classification, the highest acquired accuracies were 99.255%, 98.941%, and 95.188% in the three databases, respectively. Figure 6 shows the comparison between our method and the other methods used in different studies in detection using the MEEI database.

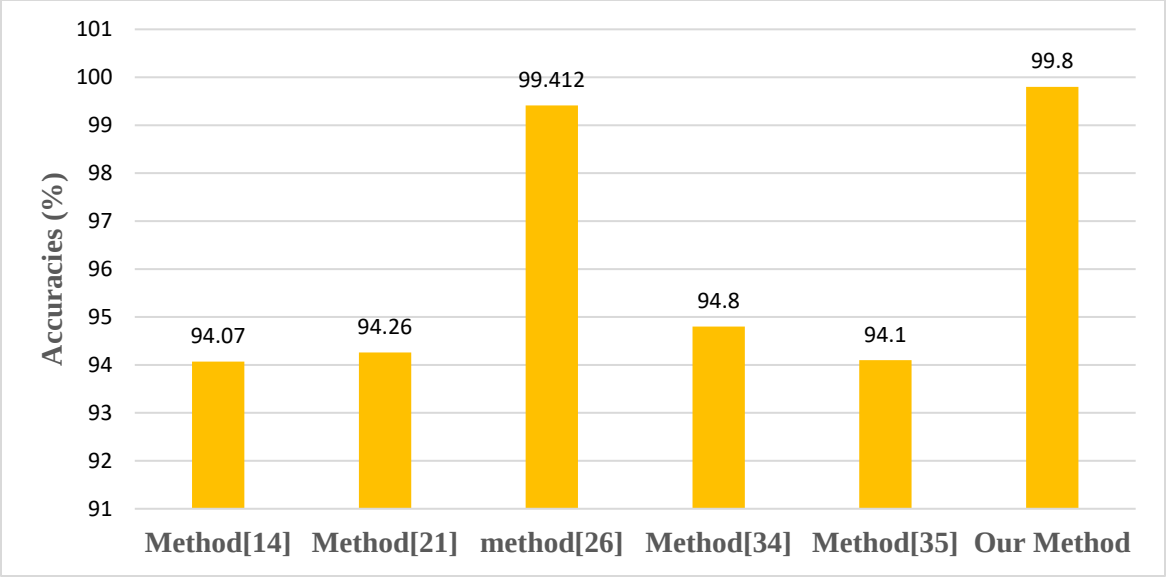

Figure 6: Different achieved accuracies from various studies

Conclusion

In this study, we evaluated the features (peak and lag) on three different databases (MEEI, SVD, and AVPD) with three different correlation functions used to extract these features. In addition, we investigated the performance of these features on eight frequency bands to see the effects of each band on the detection and classification processes. The accuracies of detection and classification varied from one database to another with the same correlation function used in extracting the features. The best accuracies we obtained in case of detection were 99.809%, 90.979%, and 91.168% in the MEEI, SVD, and APVD databases, respectively, while the best accuracies of classification were 99.255%, 98.941%, and 95.188% in the three databases, respectively.

Some of the frequency bands performed better in comparison with others. The best performance was in the bands of frequency range 1000 ~ 8000 Hz. In a future study, we will perform experiments on mixed databases samples to verify the independency of the proposed features across databases.

Acknowledgment

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-MED-2474-02).

References

[1] G. Muhammad, Z. Ali, M. Alsulaiman, and K. Almutib, "Vocal Fold Disorder Detection by applying LBP Operator on Dysphonic Speech Signal", RAICMS, 222-228, 2014.

[2] National Institute on Deafness and Other Communication Disorders: Voice, Speech, and Language: Quick Statistics, 2014.Availableat http://www.nidcd.nih.gov/health/statistics/vsl/Pages/stats.aspx. Accessed on March, 2015.

[3] Research Chair of Voicing and Swallowing Disorders. Available at http://c.ksu.edu.sa/vas/en/vsb. Accessed on October, 2015.

[4] N. Roy, R.M. Merrill, S. Thibeault, R.A. Parsa, S.D. Gray, and E.M. Smith, "Prevalence of voice disorders in teachers and the general population," J Speech Lang Hear Res., vol.47, no. 2, pp. 281-93, Apr 2004.

[5] K.H. Malki, "Voice Disorders Among Saudi Teachers in Riyadh City", Saudi Journal of Oto-Rhinolaryngology Head and Neck Surgery, 2010.

[6] B. Boyanov, and S. Hadjitodorov, "Acoustic analysis of pathological voices. a voice analysis system for the screening of laryngeal diseases", Proceedings of IEEE International Conference on Engineering in Medicine and Biology Society, vol.16, pp.74-82, 1997.

[7] C. E. Martinez, and L H. Rufiner, "Acoustic analysis of speech for detection of laryngeal pathologies", Proceedings of 22nd Annual IEEE International Conference on Engineering in Medicine and Biology Society, vol. 3, pp.2369-2372, 2000.

[8] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and recognition" J. Acoustic. Soc. Amer., vol. 54, no. 6, pp. 1304-1312, 1974.

[9] L. Xugang, and D. Jianwu, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification", Speech Communication' 07, vol. 50, no. 4, pp. 312-322, Oct 2007.

[10] M. A. Anusuya, S. K. Katti, "Front end analysis of speech recognition: a review", International Journal of Speech Technology, vol. 14, pp. 99-145, Dec. 2010.

[11] L. Rabiner and B.H. Juang, Fundamentals of speech recognition. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[12] Z. Ali, M. Aslam., and M.E. Ana María, "A speaker identification system using MFCC features with VQ technique", Proceedings of 3rd IEEE International Symposium on Intelligent Information Technology Application, pp. 115-119, 2009.

[13] W.J.J. Roberts, and J.P. Willmore, "Automatic speaker recognition using Gaussian mixture models", proceedings of Information, Decision and Control, IDC'99, pp. 465 – 470, 1999.

[14] J.I. Godino-Llorente, P. Gomes-Vilda and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters", IEEE Transactions on Biomedical Engineering, vol. 53, no. 10, pp. 1943-1953. Oct. 2006.

[15] L.E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov Chains", Ann. Math. Stat., vol. 37, pp. 1554-1563, 1966.

[16] S. Abe, Support Vector Machines for Pattern Classification. Springer-Verlag, Berlin Heidelberg New York, 2005

[17] T. Ritchings, M. McGillion, and C. Moore, "Pathological voice quality assessment using artificial neural networks," Med. Eng. Phys., vol. 24, no. 8, pp. 561–564, Sept 2002.

[18] M. Brockmann, M.J. Drinnan, C. Storck, and P.N. Carding, "Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task, Journal of voice, vol. 25, no. 1, pp. 44-53, 2011.

[19] M. Rosa, J.C. Pereira, and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," IEEE Trans. Biomed. Eng., vol. 47, no. 1, pp. 96–104, Jan 2000.

[20] Kay Elemetrics, Multi-Dimensional Voice Program (MDVP) [Computer Program], 2012.

[21] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, and A. Moqarehzadeh, " Identification of voice disorders using long-time features and support vector machine with different feature reduction methods", Journal of Voice, vol. 25, no. 6, pp. 275-289, Nov 2011.

[22] J. Wang and C. Jo, "Vocal folds disorder detection using pattern recognition method", Proceedings of 29th Annual International Conference of the IEEE EMBS, pp. 3253-3256, Lyon, France, 2007.

[23] T. Li, C. Jo, and S. Wang, "Classification of pathological voice including severely noisy cases", Proceedings of 8th International Conference on Spoken Language Processing, I, Jeju, Korea, pp. 77-80, 2004.

[24] N. Sáenz-Lechón, J.I. Godino-Llorente, Ví. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," Biomedical Signal Processing and Control, vol. 1, no. 2, pp. 120-128, April 2006.

[25] H. Oğuz, M. A. Kiliç, and M. A. Şafak, "Comparison of results in two acoustic analysis programs: PRAAT and MDVP," Turkish Journal of Medical Sciences 41.5, pp. 835-841, 2011.

[26] G. Muhammad and M. Melhem, “Pathological Voice Detection and Binary Classification Using MPEG-7 Audio Features," Biomedical Signal Processing and Controls, 11, pp. 1 – 9, 2014.

[27] G. Muhammad, T. Mesallam, K. Almalki, M. Farahat, A. Mahmood, and M. Alsulaiman, “Multi Directional Regression (MDR) Based Features for Automatic Voice Disorder Detection,” Journal of Voice, Elsevier, Vol. 26, No. 6, pp. 817.e19-817.e27, 2012.

[28] A. A-Nasheri, Z. Ali, G. Muhammad, and M. Alsulaiman, “Voice Pathology Detection Using Auto-Correlation of Different Filters Bank,” 11th ACS/IEEE International Conference on Computer Systems and Applications, Doha, Qatar, November 10-13, 2014.

[29] V. Leden, Hans, P. Moore, and R. Timcke. "Laryngeal vibrations: Measurements of the glottic wave: Part III. The pathologic larynx." AMA Archives of Otolaryngology 71.1 , 16-35, 1960.

[30] P. Lieberman, "Perturbations in vocal pitch." The Journal of the Acoustical Society of America 33, no. 5, 597-603. 1961.

[31] Mansour, David, and Biing Hwang Juang. "The short-time modified coherence representation and noisy speech recognition." IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, no. 6, pp. 795-804, 1989.

[32] Kay Elemetrics Corp., Disordered Voice Database, Version 1.03 (CD-ROM), MEEI, Voice and Speech Lab, Boston, MA (October 1994).

[33] W.J. Barry, M. P¨utzer, Saarbrucken Voice Database, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/

[34] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba "Voice Pathology Detection on the Saarbruecken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit," Advances in Speech and Language Technologies for Iberian Languages. Springer Berlin Heidelberg, 99-109, 2012.

[35] Markaki, M. and stylianou, Y., “Voice pathology detection and discrimination based on modulation spectral features", IEEE Trans. Audio, Speech, and Language processing, 19(7): 1938-1948, 2011.

[36] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and J. M. Gutiérrez-Arriola. "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: sustained vowels and running speech", Journal of Voice, vol. 27, no. 1, pp. 11-23, 2013.

[37] G. Pouchoulin, C. Fredouille, J-F. Bonastre, A. Ghio, and J. Révis. "Characterization of the Pathological Voices (Dysphonia) in the frequency space." In International Congress of Phonetic Sciences (ICPhS), pp. 1993-1996, Saarland University conference unit, August 2007.

[38] Kent, R. D. and Kim, Y. (2008) Acoustic Analysis of Speech, in The Handbook of Clinical Linguistics (eds M. J. Ball, M. R. Perkins, N. Müller and S. Howard), Blackwell Publishing Ltd., Oxford, UK. doi: 10.1002/9781444301007.ch22, p. 364-365.

[39] Lee, J.W., Kang, H.G., Choi, J.Y., & Son, Y.I. (2013). An Investigation of Vocal Tract Characteristics for Acoustic Discrimination of Pathological Voices. BioMed Research International, Vol. 2013, Article ID 758731. DOI: 10.1155/2013/758731.