Automatic Voice Pathology Detection With Running Speech by Using Estimation of Auditory Spectrum and Cepstral Coefficients Based on the All-Pole Model

Automatic Voice Pathology Detection with Running Speech by Using Estimation of Auditory Spectrum and Cepstral Coefficients Based on the All-Pole Model

Zulfiqar Ali[1,2], Irraivan Elamvazuthi[1], Mansour Alsulaiman[2], Ghulam Muhammad[2]

[1]Centre for Intelligent Signal and Imaging Research, Department of Electrical and Electronic Engineering
Universiti Tekhnologi PETRONAS
Tronoh 31750, Perak, Malaysia.
zulfiqar_g02579@utp.edy.my, irraivan_elamvazuthi@petronas.com.my

[2]Digital Speech Processing Group, Department of Computer Engineering
King Saud University
Riyadh 11543, Saudi Arabia.
msuliman@ksu.edu.sa, ghulam@ksu.edu.sa

Corresponding Author:

Zulfiqar Ali

Centre for Intelligent Signal and Imaging Research
Department of Electrical and Electronic Engineering
Universiti Tekhnologi PETRONAS
Tronoh 31750, Perak, Malaysia.

Email: zulfiqar_g02579@utp.edy.my, zulfiqarbutt2000@gmail.com

Abstract

Background and Objective: Automatic voice pathology detection using sustained vowels has been widely explored. Because of the stationary nature of the speech waveform, pathology detection with a sustained vowel is a comparatively easier task than that using a running speech. Some disorder detection systems with running speech have also been developed, although the majority of them are based on a voice activity detector (VAD) that is itself a challenging task. Pathology detection with running speech needs more investigation, and systems with good accuracy are required. Furthermore, pathology classification systems with running speech have not received any attention from the research community. In this paper, automatic pathology detection and classification systems are developed using text-dependent running speech without adding a VAD module.

***Method*:** A set of three psychophysics conditions of hearing (critical band spectral estimation, equal loudness hearing curve, and the intensity loudness power law of hearing) is used to estimate the auditory spectrum. The auditory spectrum and all-pole models of the auditory spectrums are computed and analyzed, and used in a Gaussian mixture model for an automatic decision.

***Results*:** In the experiments using the Massachusetts Eye & Ear Infirmary (MEEI) database, an accuracy of 99.56% is obtained for pathology detection, and an accuracy of 93.33% is obtained for the pathology classification system. The results of the proposed systems outperform the existing running-speech based systems.

***Discussion*:** The developed system can effectively be used in voice pathology detection and classification systems, and the proposed features can visually differentiate between normal and pathological samples.

Keywords: Running speech, voice pathology detection, voice pathology classification, auditory spectrum, all-pole model, GMM

1 Introduction

Because of its noninvasive nature, the automatic assessment of voice pathology is strongly being considered as a primary screening tool or helping tool for the clinician. It will be of great help to an ENT specialist if an automatic assessment system can discriminate between normal and pathological samples, as well as classify the voice pathologies. The process of differentiating between normal and pathological subjects is a two-class problem referred to as pathology detection. By contrast, discriminating between different types of pathologies is a multiclass problem referred to as pathology classification. Automatic pathology detection is a widely explored area by the research community, while pathology classification is considered to be a more difficult task as compared with pathology detection and receives less attention.

Most of the automatic pathology assessment systems, pathology detection, and pathology classification present in the literature are developed by using a sustained vowel /ah/ [1-4]. This is a comparatively easier task than developing assessment systems with continuous (running) speech. A speech signal remains stationary in the case of a sustained vowel, but it varies over time in the case of continuous speech. This is the reason why pathology assessment systems that use continuous speech are challenging and require more investigation. Moreover, these systems are more realistic because people use continuous speech in their conversations in daily life. Running speech contains fluctuations of vocal characteristics in relation to voice onset, voice termination, and voice breaks, which are considered to be crucial in quality voice evaluation. These characteristics are not fully represented in short signals of phonation such as a sustained vowel [5].

Running-speech-based systems sometimes perform voice activity detection (VAD) before extracting the speech features. The VAD module is responsible for automatic segmentation of the voiced, unvoiced and silence parts of a speech signal. In a study [6], Umapathy et al. acknowledged that voice activity detection is itself a challenging task. On the other hand, in the existing pathology detection system that uses running speech, there are no visual cues for the features for medical doctors to use in their decisions. If a detection decision is verified by clear visual evidence, the system will be more useful to doctors. Moreover, present features based on human auditory mechanisms using a few psychophysics conditions of hearing, showing good performance for pathology detection but not for pathology classification. The proposed system will use the features incorporated by more conditions of hearing so that it may provide good results for both types of tasks (pathology detection and pathology classification).

VAD-free pathology detection and classification systems based on text-dependent running speech are developed in this study. The proposed features are referred as auditory processed spectrum (APS) and are estimated by using three principles of hearing, namely, critical band spectral estimation, the equal loudness hearing curve, and the intensity loudness power law of hearing. The idea is to simulate the human perception of voice. The APS features are further analyzed by using an all-pole model to obtain all-pole model based cepstral coefficients (APCC). The all-pole models are obtained by linear predictive (LP) analysis [7, 8], which is widely used in speech-based applications, is well known to approximate the high-energy regions of a speech spectrum, and provides a fine harmonic structure. The proposed system has two important characteristics: (i) it is VAD-free, so there is no extra complexity to calculate VAD; and (ii) the features have the capability to detect pathologies by visual inspection, so doctors can have an extra level of screening apart from the automatic decision from the system.

The rest of the paper is organized as follows: Section 2 describes the related work on text-dependent running-speech-based pathology systems, Section 3 provides information on the developed systems, Section 4 describes the experimental setup and results for pathology detection and classification, Section 5 contains a discussion of the proposed systems, and finally Section 6 draws some conclusions.

2 Related Work

Speech features are broadly divided into following two categories based on simulated human hearing mechanisms or model human voice production systems. One of the first types of speech features is Mel-frequency cepstral coefficients (MFCC), and an example of another type of feature is linear prediction coefficients (LPC), which are based on all-pole model. MFCC used triangular band-pass filters (BPFs) to divide the spectrum into certain frequency bands. The

center frequencies of the BPFs are spaced on a Mel scale, and the bandwidths correspond to well-known auditory perception phenomena called critical bandwidths. MFFC mimic the behavior of the human auditory system, and have shown great success in pathology detection: 97.46% with sustained vowels [9] and 96% with running speech [10]. The performance of MFCC for pathology classification was not good: it provided an average accuracy of 70% in [11]. The proposed features also use critical bandwidth phenomena. Additionally, they are incorporated by other conditions of hearing, explained in Section 3, so that the developed system may provide good results for pathology detection as well as for pathology classification.

Another MFCC-based system for disordered detection by using text-dependent running speech was developed in [12]. The running speech of a limited number of normal and disordered subjects (12 and 26, respectively) was used to evaluate the developed system. An accuracy of 91.66% was reported. The Gaussian mixture model (GMM) was implemented as a classification technique with a varying number of Gaussian mixtures. A limited number of samples were used for experiments; therefore, no reliable conclusions could be made.

In [10], Godino et al. proposed a text-dependent running-speech-based pathology detection by using MFCC, and a VAD module was implemented to extract the voiced part of the continuous speech. The obtained accuracy (ACC) was 96%. A receiver operating characteristic (ROC) curve was also plotted; the area under the ROC curve was 99.8%. The results with running speech were better than those with the sustained vowels. To perform the experiments, a subset of the Massachusetts Eye & Ear Infirmary (MEEI) database [13] containing 117 pathological and 23 normal subjects was selected. The "Rainbow" passage was used as a running speech; its text is given in Appendix A. To avoid biased results from the classifier, in the VAD module it was better to reject a voice segment than to accept an unvoiced segment. It is difficult to classify voiced and unvoiced (V-UV) speech segments accurately. In a study [6], Umapathy and his coauthors acknowledged that voice activity detection is itself a challenging task. An accuracy of 93.4%, without a voice activity detector, was obtained in the study. In a study by Lowell et al. [14], cepstral analysis was performed by measuring two features: cepstral peak prominence (CPP) and smoothed CPP (CPPS) [15]. CPP and CPPS can be calculated only in voice segments; therefore, VAD was necessary for these features.

Long-term acoustic features, such as shimmer and jitter, were also used for running-speech-based pathology detection systems. Such types of measurements normally involve the accurate estimation of the pitch period, which is a very difficult task, especially in pathological samples. Vicsi et al. [16] used acoustic features, shimmer, jitter, harmonic-to-noise ratio (HNR), and MFCC to distinguish between normal persons and dysphonic patients. The best obtained accuracy was 84% with the sustained vowel /i/ by using shimmer and jitter. The accuracy of the system was increased to 86% when tested with running speech: a folk tale, "The North Wind and the Sun." The study was conducted with a private database containing 33 pathological and 26 healthy subjects. In a study by Parsa and Jameison in 2001, voiced/unvoiced segments of running speech were extracted by using contours of fundamental frequency [17]. The authors used nine acoustics parameters including shimmer, jitter, fundamental frequency, and linear predictive (LP) modeling-based measures for the discrimination of normal and pathological subjects. Fifty-three normal and 175 pathological samples of the MEEI database for both sustained vowel and running speech were considered in the study. The LP-based measures provided the highest accuracy of 96.5%.

Zhang and Jiang [18] differentiated normal and pathological subjects with the help of perturbation measures including shimmer and jitter, signal-to-noise ratio (SNR), and nonlinear dynamic (NLD) measures including correlation dimension and second-order entropy by using sustained vowels and running speech. As the results suggested, shimmer and jitter did not exhibit a significant difference between normal and pathological signals for the running speech, although nonlinear dynamic measures and SNR were statistically significant for running speech. The results in the form of ACC or AUC were not provided; therefore, a comparison with other studies is not possible.

In [19], Watts and Awan reported that an accuracy of 91% is obtained for a running-speech-based pathology detection system. The performance of the system is not as good as that of other studies.

A pathology classification with running speech has been reported in a recent study [20]. Three different types of vocal fold pathologies (edema, paralysis, and nodules) are considered for the experiments. The maximum achieved accuracy was 76.2% when edema and nodules were combined and detected from paralysis and healthy samples. The authors

mentioned that there are few works in voice-pathology classification, and all use the sustained vowel /a/ as a speech signal. No significant work is present for pathology classification with running speech.

3 Method

Automatic voice pathology detection and classification systems are developed in this system. Two sets of features are used in the developed systems and are described in the following subsections.

3.1 Auditory Processed Spectrum (APS)

The auditory processed spectrum is estimated from running speech by using a set of psychophysics conditions of hearing. The perception can be modeled as a sequence of the critical band spectral estimation, equal loudness hearing curve, and intensity loudness power law of hearing. A set of auditory transformations based on the human hearing system modifies the spectrum. The steps to compute the APS features are depicted in Fig. 1. The spectrum of the normal/pathological speech sample is generated by applying a Fourier transformation (FT). Then the spectrum is passed through a set of band-pass filters to produce a critical band spectrum. The Fourier transformation provides the information of energy at each frequency component. Before applying the FT, the speech signal is divided into overlapping frames and multiplied with a hamming window [12] to remove the spectral leakage at the ends of the divided frames.

The inner ear of a human plays a very important role in separating different frequencies. The inner ear transfers the energy from different frequencies to the basilar membrane. The higher frequencies are localized at the basal turn and the lower frequencies towards the apex of the cochlea, as shown in Fig 2.

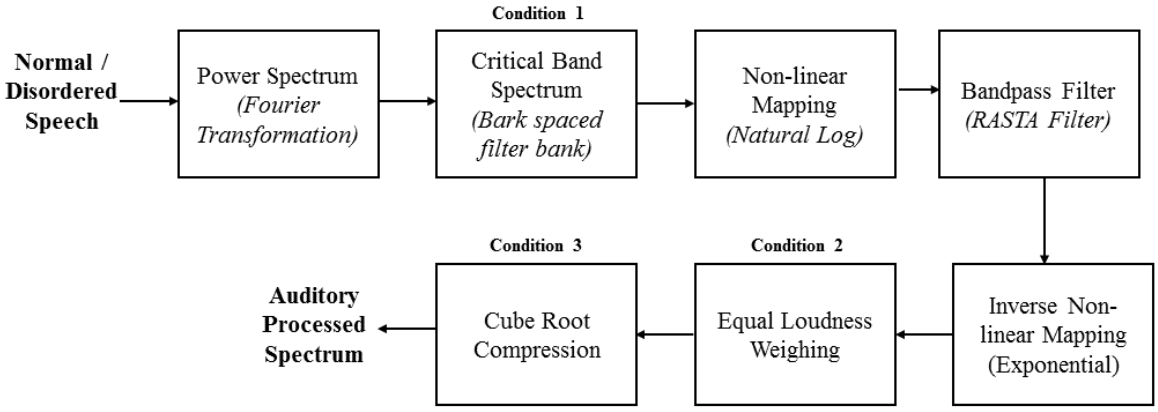


Figure 1. Computational steps to estimate the auditory spectrum.

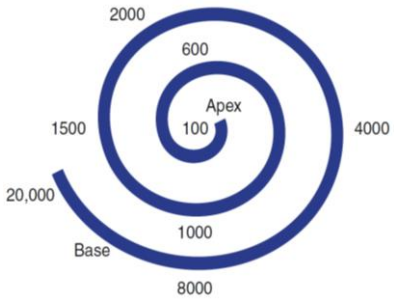


Figure 2. Frequency distribution in Hz along cochlea [21].

Each point on the basilar membrane can be considered as a band-pass filter. The bandwidth of the cochlear filter is nonlinear and increases with an increase in frequency, as shown in Fig 3. Critical bandwidth—one critical band is referred to one bark—is linear up to 500 Hz and increases by 20% of the center frequency above 500 Hz. To approximate critical bands, two different scales are proposed: one by Zwicker [22] and the other by Schroeder [23]. The scale by Zwicker is used in this study and is given by Eq. (1):

$$Bark_{\text{Zwicker}} = 13\arctan(0.00076\,f) + 3.5\arctan\left(\frac{f}{7500}\right)^2 \quad (1)$$

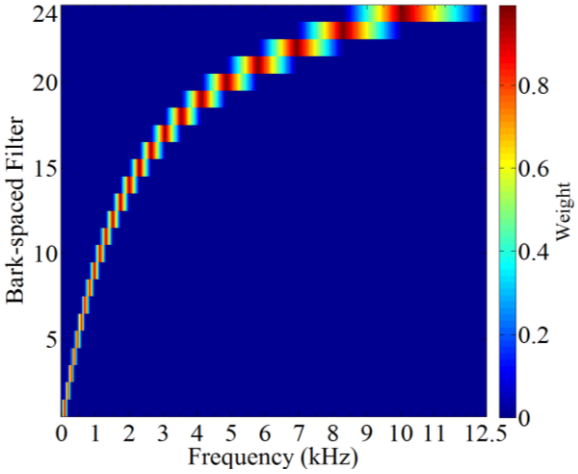


Figure 3. Twenty-four bark spaced band-pass filters.

The spectrum of the speech signal is wrapped with the Bark scale by filtering the spectrum through band-pass filters. The center frequencies of the band-pass filters are provided in Table 1. Twenty-four filters are used in this study.

Table 1: Center Frequencies of Bark spaced band-pass filter bank

Filter No.	Frequency (HZ)	*Filter No.*	Frequency (HZ)
	94.6	*13*	1737.4
2	186.1	*14*	1995.8
3	276.3	*15*	2293.0
4	368.5	*16*	2638.5
5	468.4	*17*	3045.1
6	577.3	*18*	3530.5
7	696.3	*19*	4120.3
8	827.1	*20*	4852.0
9	971.5	*21*	5783.2
10	1131.6	*22*	6923.0
11	1310.2	*23*	8298.8
12	1510.7	*24*	10046.4

After remapping the frequency axis to the bark scale, a Bark-warped critical band spectrum is obtained. Then, Log is applied on the obtained spectrum to dynamically compress the spectral amplitude. Moreover, it converts multiplicative distortions to additive distortions, which can be filtered. The log spectrum is now passed through a band-pass filter—relative spectra, known as RASTA—to remove the effect of constant and slowly varying parts in each component of the estimated critical band spectrum. This filtering emphasizes the spectral changes occurring in the range of 1–10 Hz. The human auditory system is relatively insensitive to those slowly varying stimuli. The RASTA filter is given by Eq. (2):

$$H(z) = z^4 \times \frac{\left(0.2 + 0.1z^{-1} - 0.1z^{-3} - 0.2z^{-4}\right)}{1 - 0.94z^{-1}} \tag{2}$$

The inverse of Log is applied to the output of RASTA filter. To incorporate the phenomenon that human hearing is more sensitive to the middle frequency range of the audible spectrum, each critical band spectrum is multiplied by an equal loudness curve. The curve suppressed the low and high frequencies relative to midrange from 400 Hz to 1200 Hz. According to the power law of hearing, a nonlinear relationship exists between the intensity of sound and perceived loudness. The phenomenon is incorporated after taking the cube root of the spectrum, which compresses the spectrum, and the obtained output is referred to as the auditory spectrum of the input signal. The obtained auditory processed spectrum is given to the GMM for classification of normal and pathological subjects.

3.2 All-Pole Model Based Cepstral Coefficients (APCC)

Voice pathologies affect the vocal folds, and these disorders produce irregular vibrations in the vocal folds owing to the malfunctioning of the voice box. Vocal fold pathologies exhibit variations in the vibratory cycle of the vocal folds because of their incomplete closure. A voice disorder also changes the shape of the vocal tract and produces irregularities in spectral properties. Vocal tract properties can be modeled using the all-pole model with the help of linear predictive (LP) analysis.

The steps to perform an LP analysis are presented in Fig. 4. Autocorrelation is implemented with an inverse Fourier transformation, and LP coefficients (LPC) are obtained after applying the Levinson-Durbin algorithm. Then the LPC are given to a recursive routine, determined by Eq. (3), to calculate the cepstral coefficients. LP analysis provides fine a harmonic structure, and is well known to approximate the high-energy regions of a speech spectrum. The computed cepstral coefficients are inputted to the GMM for the differentiation of normal and pathological patients.

In Eq. (3), σ^2 is the gain term in the LPC, p is the order of the LP analysis, a_n are LPC, and c_n are obtained cepstral coefficients.

$$\left.\begin{aligned} c_1 &= \ln \sigma^2 \\ c_n &= a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c_k \; a_{n-k} \;, \quad 1 \le n \le p \\ c_n &= \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c_k \; a_{n-k} \;, \quad n > p \end{aligned}\right\} \tag{3}$$

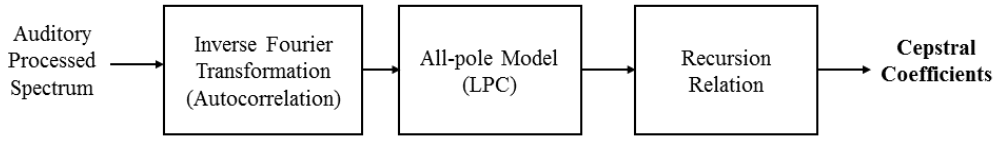


Figure 4. Steps to compute all-pole model of estimated spectrum.

3.3 Gaussian Mixture Model

GMM [24] is a state-of-the-art modeling technique that has been widely used in different scientific areas including voice pathology [12, 25-27]. GMM copes more with the space of the features rather than the time sequence of their appearance. The basis for using GMM is that the distribution of feature vectors extracted from an individual's speech data can be modeled by a mixture of Gaussian densities. A Gaussian mixture density is a weighted sum of *M* component densities given by

$$p(X \mid \Theta) = \sum_{i=1}^{M} w_i \cdot g(X \mid \mu_i, \Sigma_i), \qquad i = 1,2,3,...,M \tag{4}$$

where μ_i, Σ_i, and w_i are the mean vector, covariance matrix, and weight (prior probability) of the i^{th} Gaussian component, respectively. A K-means algorithm is used to initialize the parameters. These parameters are estimated and tuned by the well-known Expectation-Maximization (EM) algorithm [28] to converge to a model giving a maximum log-likelihood value. The calculated features are represented by D-dimensional data vector $X = \{x_1, x_2, x_3, ..., x_D\}$, and the density of each component is given by a D-dimensional Gaussian function of the form

$$g(X \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu_i)^T \Sigma^{-1} (X-\mu_i)\right) \tag{5}$$

The weights of Gaussian components satisfy the following constraint:

$$\sum_{i=1}^{M} w_i = 1 \quad \text{and} \quad 0 \le w_i \le 1 \tag{6}$$

By using the GMM model, denoted by $\Theta = (\mu_i, \Sigma_i)$, two types of experiments are performed. The first type of experiment is for pathology detection to differentiate between normal and pathological subjects. The other type of experiment is for pathology classification to detect the type of voice pathology.

In pathology detection, all normal subjects generate one GMM model, and all disordered subjects generate a second GMM model. The GMM models are generated by using different numbers of Gaussian component/mixtures. Once models are generated, a parametric test utterance (extracted features) will be compared with both models, and the log-likelihood of both models will be computed to make a decision. If the log-likelihood of the test utterance is greater for the GMM model of normal subjects, then the test utterance belongs to the normal class; otherwise, it belongs to the pathological class.

During the feature extraction process, the speech signal was divided into short frames. A final score of the log-likelihood to a parametric representation of the test utterance will be assigned by adding the log-likelihood score for each frame. In this process, independence between the frames is assumed. A block diagram showing pathology detection by using GMM is depicted in Fig. 5.

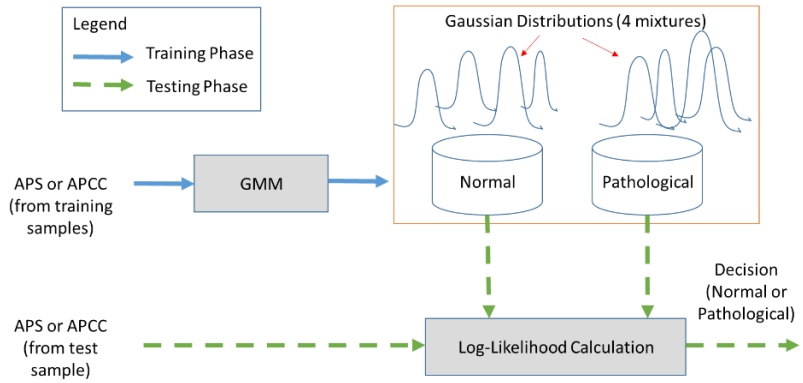


Figure 5. Pathology detection by using GMM.

For pathology classification, GMM models for all disorders (adductor, keratosis, nodules, polyp, and paralysis) are generated by using a different number of GMM components. Then, to find the type of disorder, a parametric test utterance will be compared with all the generated models one by one, and the log-likelihood values of the test utterance will be computed with each model. The GMM model with the maximum log-likelihood will decide the type of disorder for the test utterance. The process for pathology classification is shown in Fig. 6.

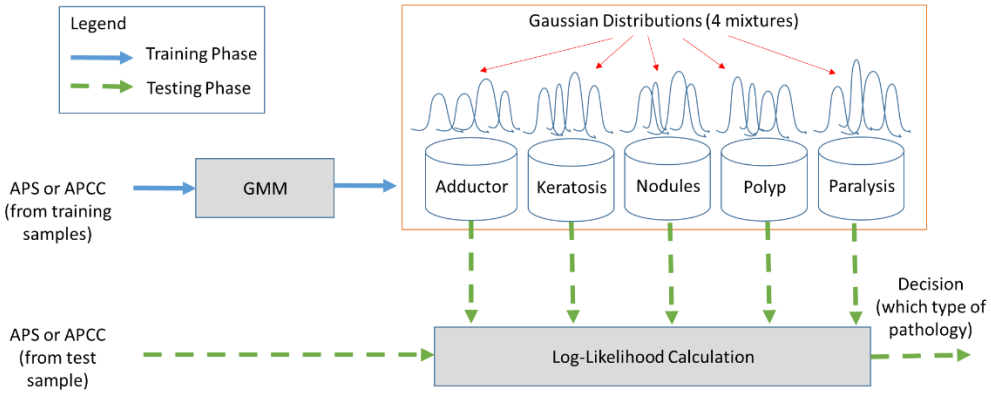


Figure 6. Pathology classification by using GMM.

4 Material and Experimental Results

The APS and APCC features are computed for normal and pathological speech signals. Speech signals are taken from the MEEI database. The database was recorded with a condenser microphone in a controlled environment at the Massachusetts Eye & Ear Infirmary Voice & Speech Laboratory. The database recorded a large number of voice disorder patients having different kinds of voice complaints after a clinical evaluation. The MEEI database contains two types of signals: sustained vowel /ah/ and running speech that corresponds to the Rainbow passage.

The MEEI database was recorded at two sampling frequencies (25 KHz and 50 KHz). Therefore, samples having a sampling frequency of 50 KHz were downsampled to 25 KHz. A subset of the MEEI database as mentioned in [1, 4] is used for the evaluation of the disorder detection and classification. The subset contains 173 pathological and 53 normal subjects. The pathological subjects are suffering from adductor spasmodic dysphonia, vocal fold nodules, keratosis, vocal fold polyp, and paralysis. A distribution of the normal and pathological samples having different types of disorders is shown in Table 2.

Table 2: Distribution of normal and pathological samples in the MEEI subset

Subjects	Disorders Type	No. of Samples	Total
Pathological	Adductor spasmodic dysphonia	22	173
	Vocal fold nodules	20	
	Keratosis	26	
	Vocal fold Polyp	20	
	Paralysis	85	
Normal	---	53	53

The group of normal subjects contained 21 male and 32 female, while the number of males and females for pathological subjects were 70 and 103, respectively. For normal subjects, the age range for male speakers is 26–59 years, and for female speakers the range is 22–52 years. The average age for males and females in the group of normal subjects is 38.81 and 34.1 years, respectively. For pathological subjects, the age of the male patients is within the

range of 26–58 years, and the range for females is 21–51 years. The average age for males is 41.71 years, and for females the average age is 37.58 years.

All subjects who recorded the sustained vowel also recorded the Rainbow passage. There were two instances when a pathological subject recorded the sustained vowel but did not record the Rainbow passage; the missing subjects are FXC12AN and MCA07AN. Therefore, in this study, 171 pathological and 53 normal speech samples are used.

To compute the APS features, the speech signals were divided into windows of 10 ms, with 50% overlapping, and multiplied with the hamming window. A 256-point FT and 24 band-pass filters in a bark-spaced filter bank are used to calculate the APS features. All-pole models are calculated by using 11th-order LPC analysis to obtain the APCC features. The first- and second-order derivatives are also calculated on these cepstral components by using a linear regression calculated by Eq. (7):

$$\Delta t = \frac{\sum_{i=1}^{B} i\left(c_{t-i,m} - c_{t+i,m}\right)}{2\sum_{i=1}^{B} i^2} \tag{7}$$

where Δt corresponds to the first-order derivative at the t^{th} frame, $c_{t,m}$ represents mth coefficients of the t^{th} frame, and B is a length of the regression window, which is 5 in this study.

The results of the developed system based on the proposed features, are expressed in terms of sensitivity, specificity, accuracy, and the area under the ROC curve. These features are represented by SEN, SPE, ACC, and AUC, respectively. The terms are define as follows: accuracy is the ratio between correctly detected samples and the total number of samples, sensitivity is the ratio between truly identified pathological samples and the total number of pathological samples, and specificity is the ratio of truly classified normal samples and the total number of normal samples. The performance parameters SEN, SPE, and ACC are calculated by using following relationships:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{8}$$

$$SN = \frac{TP}{TP + FN} \times 100 \tag{9}$$

$$SP = \frac{TN}{TN + FP} \times 100 \tag{10}$$

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 the pathological subject as a normal subject, and false positive (FP) means that the system detects the normal subject as a pathological subject.

4.1 Pathology Detection

A five-fold approach is used to perform the experiments for pathology detection. The database is divided into five equal disjoint sets where each time one set is used in the system evaluation while the remaining four sets performed training. In this way, each sample from the database is used as a testing utterance, and a log-likelihood value is computed for it. The classifier makes the decision on the basis of the computed log-likelihood. By using the five-fold approach, the results of the developed system become robust against training and testing sets of the samples. The reported performance parameters SEN, SPE, and ACC (in Tables 3, 4, and 5) are averaged over five folds, and the standard deviation (STD) of the performance parameter over the five folds is also provided.

Gaussian models for normal and pathological subjects are estimated by using a different number of Gaussian mixtures, i.e., 8, 16, 32, 48, 64, and 80. Experiment results showed that the GMMs of the two classes are estimated accurately when the number of mixtures is increased. Pathological subjects are represented as a positive class, and normal subjects are considered as a negative class.

Detection with Auditory Processed Spectrum (APS)

The performance of the APS features for voice pathology detection is shown in Table 3. The highest obtained accuracy is 98.22% and the AUC is 98.57%. The attained SEN and SPE are 97.66% and 100%, respectively. The STD is zero for SPE which shows that the SPE is independent of the training and testing data. For each fold in the five-fold approach, it was 100%. The ROC curves for the APS with 64 and 80 Gaussian mixtures are shown in Fig. 7 (a). The decision values of the classifier, which are the log-likelihoods, are used to plot the ROC curve. A high value of the AUC indicates that classifier is reliable in differentiating normal and pathological samples.

Table 3: Performance measures (%) for pathology detection with APS

Gaussians	SEN ± STD	SPE ± STD	ACC ± STD	AUC
8	92.42 ± 3.2	100 ± 0	94.19 ± 2.5	96.35
16	96.49 ± 2.4	100 ± 0	97.32 ± 1.8	97.72
32	97.06 ± 2.4	100 ± 0	97.77 ± 3.8	97.75
48	97.06 ± 2.9	100 ± 0	97.77 ± 2.2	97.76
64	97.06 ± 3.6	100 ± 0	97.77 ± 2.7	97.70
80	**97.66 ± 2.4**	**100 ± 0**	**98.22 ± 1.8**	**98.57**

Detection with All-Pole Model Based Cepstral Coefficients (APCC)

The results of pathology detection with 12 coefficients are provided in Table 4, and that with 36 coefficients (12 static + 12 first derivative +12 second derivative) are presented in Table 5. The results with 24 coefficients (12 static + 12 first derivative) are not provided as they did not show any improvement compared with the results of 12 coefficients. An accuracy of 98.22 % is obtained with 12 coefficients, and the AUC is 98.50%. The overall maximum accuracy is achieved with 36 coefficients. That maximum accuracy is 99.56% with an STD equal to 0.9. The other performance measures (SEN and SPE) are 99.41% and 100%, respectively; and the AUC is 99.99%. ROC curves are plotted only for 80 Gaussian mixtures to avoid overlapping with curves of other Gaussians. This is shown in Fig. 7 (b).

Table 4: Performance measures for pathology detection with APCC by using 12 features

Gaussians	SEN ± STD	SPE ± STD	ACC ± STD	AUC
8	94.74 ± 4.3	92.55 ± 7.6	94.20 ± 1.9	95.53
16	97.08 ± 5.0	98.18 ± 4.0	97.33 ± 3.6	97.67
32	97.66 ± 3.8	98.18 ± 4.0	97.78 ± 2.7	98.27
48	97.08 ± 5.0	98.18 ± 4.0	97.33 ± 3.6	97.80
64	97.65 ± 5.2	98.18 ± 4.0	97.78 ± 3.8	97.97
80	**98.24 ± 3.9**	**98.18 ± 4.0**	**98.22 ± 2.8**	**98.50**

Table 5: Performance measurements for pathology detection with APCC by using 36 features

Gaussians	SEN ± STD	SPE ± STD	ACC ± STD	AUC
8	96.50 ± 2.4	98.18 ± 4.0	96.88 ± 1.2	97.56
16	98.27 ± 2.5	98.18 ± 4.0	98.22 ± 1.8	99.30
32	98.25 ± 2.6	100 ± 4.0	98.67 ± 1.9	99.04
48	99.41 ± 1.3	98.18 ± 4.0	99.11 ± 1.2	99.74
64	99.41 ± 1.3	98.18 ± 4.0	99.11 ± 1.9	99.91
80	**99.41 ± 1.3**	**100 ± 0**	**99.56 ± 0.9**	**99.99**

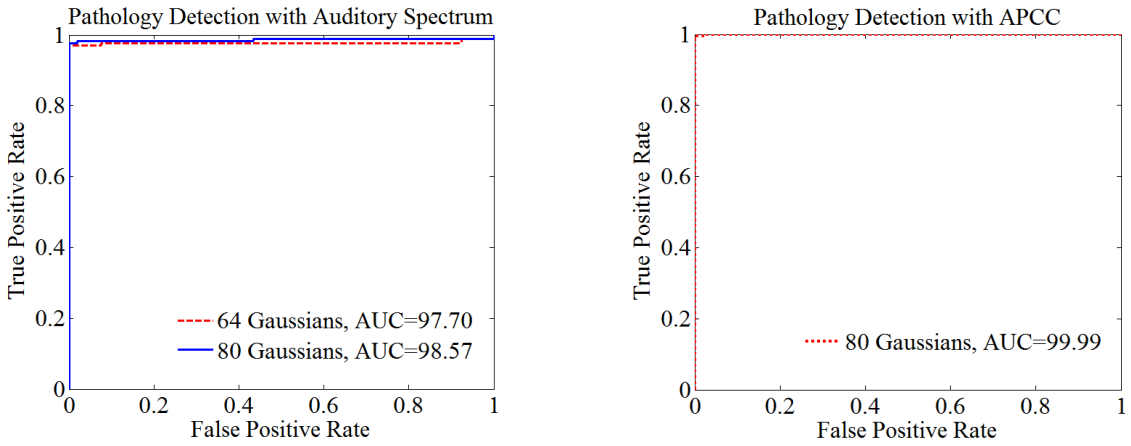


Figure 7. ROC curves for pathology detection with (a) APS by using 64 and 80 Gaussian mixtures and (b) 36 APCC by using 80 Gaussian mixtures.

The log-likelihood values are used as the discriminative measure to differentiate between two types of subjects. The Mann-Whitney U-test is performed to check the discriminant power of the log-likelihood at the 5% significant level. The obtained two-sided *p*-values are 0.0001E-22 and 0.0004E-24. For APS and APCC features, respectively, both *p*-values are less than 0.05, which rejected the null hypothesis that the log-likelihood values of normal and pathological classes are from continuous distributions with equal medians. For APS features, the mean and standard deviation of the log-likelihood for pathological and normal classes are (1.48, 0.14) and (3.73, 0.48), respectively. For APCC, the results are (1.07, 0.23) and (4.73, 0.28), respectively. Different means and a small standard deviation of the log-likelihood show that both types of features can detect normal and pathological subjects.

For visual indication, energy distributions for the APS features for different pathological and normal subjects are depicted in Fig. 8. The APS features were computed by following the steps mentioned in Fig. 1. The spectrums were calculated for the entire speech signal, but for the sake of clarity, energy contours are plotted by considering the same part of the Rainbow passage, i.e., "when the sunlight." It can be observed that high energies (circled in the figures) for the normal subjects belong to the region of lower frequencies, while those for voice-disordered subjects belongs to a high-frequency region. This suggests that normal and pathological subjects can be classified based on the high-energy regions that have significantly different representations in the auditory spectrum for both types of subjects. For a pathological subject, higher bands in the APS features model the noisy components owing to the lack of closure of vocal folds.

The energy distribution for the APCC features is also plotted in Fig. 8. The all-pole model provides a fine harmonic structure and highlights the high-energy regions of the spectrums. It can be observed from Fig. 8 that the high-energy for a normal subject is concentrated, and the high-energy for a pathological subject is sparse. The different representations of the energy contour, concentrated for normal, subjects and scattered for pathological subject, leads to the hypothesis that these coefficients can discriminate between normal and pathological voices. The cepstral coefficients obtained after Fig. 4 are also referred to as RASTA-PLP coefficients. The spread of energy contours in pathological voices can be supported by [29], where the authors showed that the energy is concentrated in a normal voice, while that for a pathological voice is distributed.

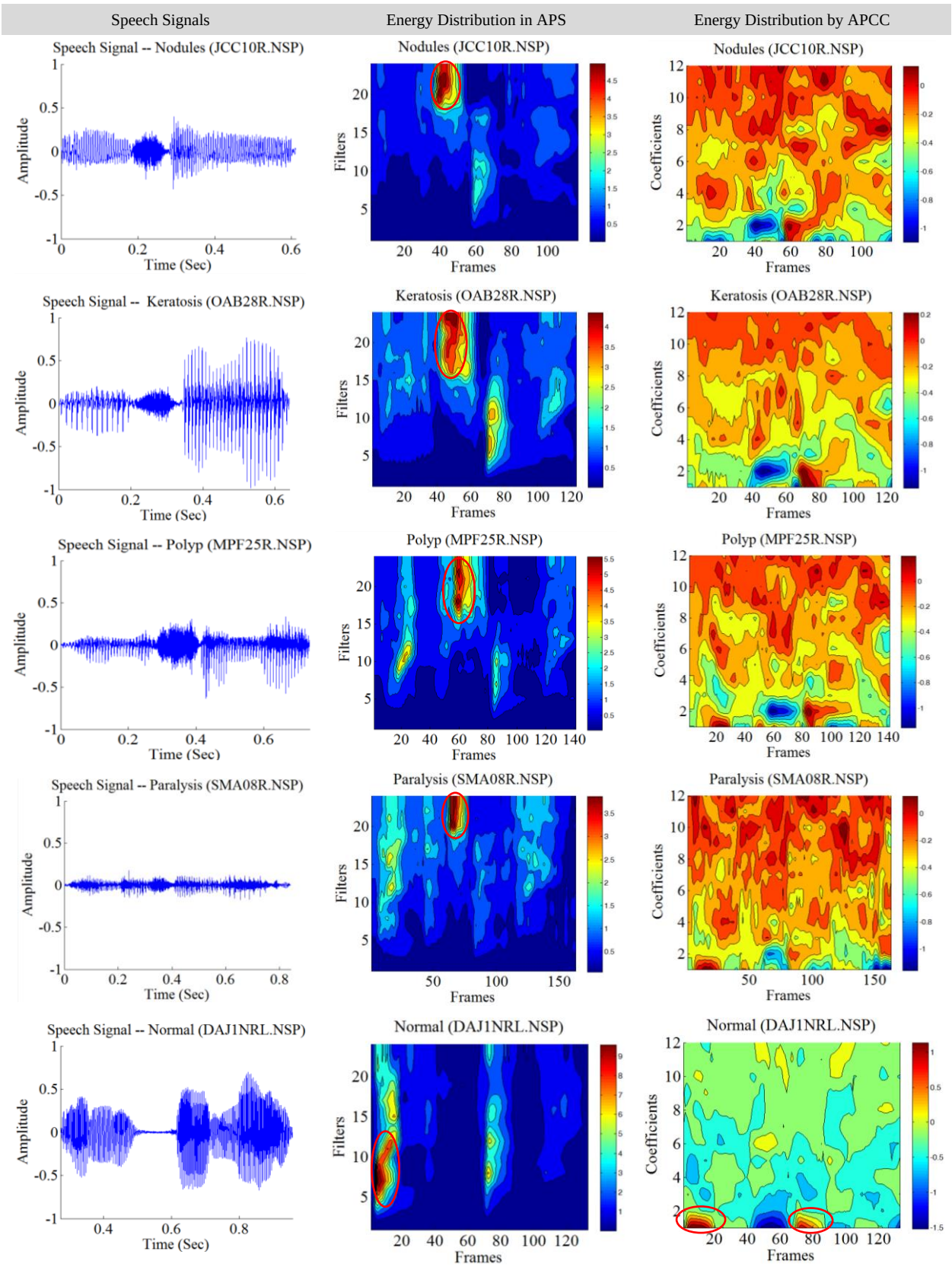


Figure 8. Energy distribution by APS and APCC for different normal and disordered subjects. All plotted speech signals correspond to the same part of the Rainbow passage "when the sunlight."

4.2 Pathology Classification

For pathology detection, the five-fold approach was used, while for pathology classification all experiments are performed by using two-fold cross-validation owing to the limited number of samples for the disorders. The samples that are labeled with multiple disorders are removed from the list. Different numbers of Gaussian mixtures (4, 8, 16, 32, and 48) are used to train each model.

Classification with Auditory Processed Spectrum (APS)

The classification results of each disorder (adductor, keratosis, nodules, polyp, and paralysis) with APS features are mentioned in Table 6, and the best-obtained accuracy for each disorder is represented by a bold value. The maximum accuracy for adductor is 93.33%, which was obtained with 32 Gaussian mixtures. The highest accuracy for keratosis is 85.53% with 48 Gaussians. Similarly, the best accuracy for nodules, polyp, and paralysis are 85.53%, 89.33%, and 78.42%, respectively. The AUC was 94.01%, 86.92%, 88.36%, 91.99%, and 82.66% for adductor, keratosis, nodules, polyp, and paralysis, respectively. The ROC curve for each disorder is depicted in Fig. 9(a).

Table 6: Accuracies (%) of all disorders with APS for different number of mixtures

Disorder Type	No. of Gaussian Mixtures				
	4	*8*	*16*	*32*	*48*
Adductor	84	78.67	84	**93.33**	89.33
Keratosis	60.53	71.05	72	84.21	**85.53**
Nodules	81.33	**85.53**	76	80.26	81.58
Polyp	44	**89.33**	84.21	85.53	88.16
Paralysis	64	69.33	77.11	75.33	**78.42**

* Bold values represent best-obtained accuracy for each disorder

Classification with All-Pole Model Based Cepstral Coefficients (APCC)

The results of all disorders by using 12 APCC are presented in Table 7. The best-obtained accuracy for adductor is 89.33% with 16 Gaussians; for keratosis, 81.58 with 16 Gaussians; for nodules, 85.33% with 48 Gaussians; for polyp, 89.47%; and for paralysis, 78.42%. The AUC for adductor, keratosis, nodules, polyp, and paralysis are 90.20%, 83.02%, 87.60, 91.85%, and 82.62%, respectively. The ROC curves for each disorder are shown in Fig. 9(b) for APCC.

Table 7: Accuracies (%) for each disorder with 12 APCC by using different number of mixtures

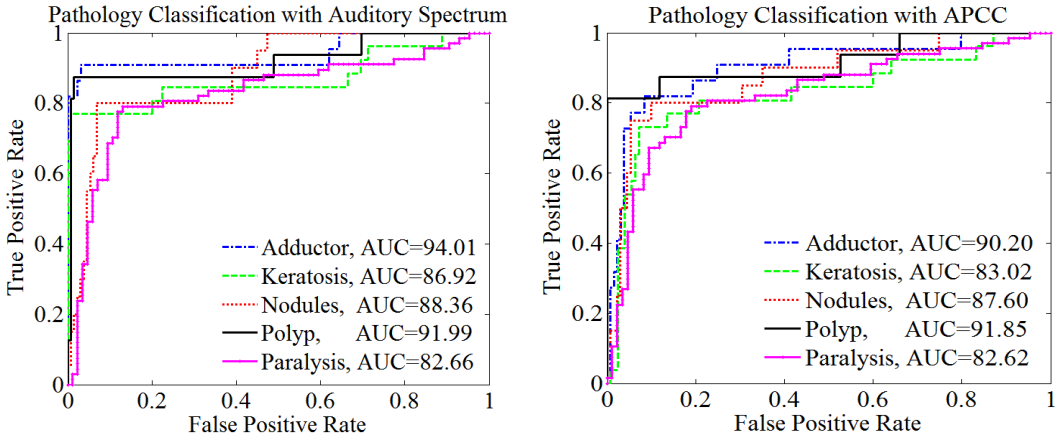
Disorder Type	No. of Gaussian Mixtures				
	4	*8*	*16*	*32*	*48*
Adductor	81.33	82.89	**89.33**	89.33	88
Keratosis	68.42	72.37	**81.58**	77.63	80.26
Vocal Nodules	78.67	82.67	78.67	81.33	**85.33**
Vocal Fold Polyp	81.33	75	81.58	**89.47**	89.47
Paralysis	73.33	76.67	78.00	76.67	**78.42**

* Bold values represent best-obtained accuracy for each disorder

To plot the ROC curve for the disorder adductor, log-likelihood values of the disorder adductor are considered as a positive class, while the log-likelihood for all other disorders are considered as a negative class. Similarly, the ROC curve for each disorder is plotted in Fig. 9. The log-likelihood are used to decide the type of voice disorder. For the significance of log-likelihood values, the Mann-Whitney U-test is performed at a 5% significance level. The two-sided *p*-value obtained for all disorders are provided in Table 8. The obtained *p*-values are less than 0.05, which rejects the null hypothesis that log-likelihood values of both classes are from continuous distributions with equal medians. The test shows that the proposed features can detect the type of a disorder. The AUC and *p*-values are calculated by using the log-likelihood values that correspond to the best accuracy.

Table 8: *p*-values for each disorder obtained by performing Mann-Whitney U-test at 5% significant level (α = 0.05)

Features	Adductor vs. Rest	Keratosis vs. Rest	Nodules vs. Rest	Polyp vs. Rest	Paralysis vs. Rest
APS	0.0004E-7	0.0003E-5	0.0003E-4	0.0004E-4	0.0005E-8
APPC	0.0001E-5	0.0001E-3	0.0006E-4	0.0004E-4	0.0006E-8



(a) (b)

Figure 9. ROC curves for each disorder with (a) APS and (b) APCC.

A summary of the results for pathology classification with both types of features is shown in Fig. 10. ACC-APS and ACC-APP represent the accuracy of APS and APCC features, respectively, while AUC-APS and AUC-APCC represent AUC for APS and APCC, respectively. It can be observed that overall highest obtained accuracy is achieved by adductor, i.e., 93.33%. The performance of AS features is better than that of APCC features in the case of adductor and keratosis, whereas for the rest of the disorders the performance of both types of feature is the same.

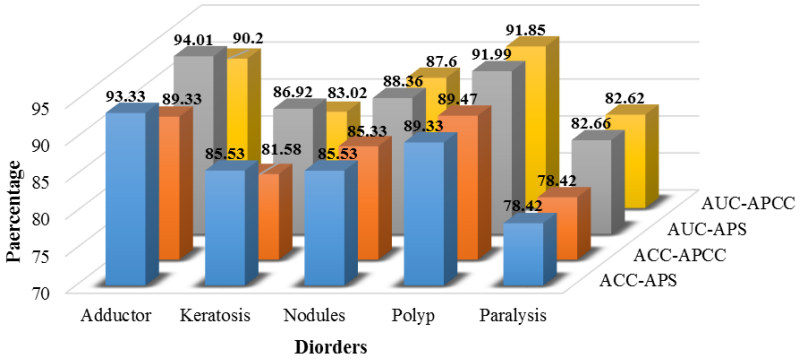


Figure 10. Summary of results for pathology classification.

5 Discussion

Automatic voice pathology detection and classification systems are developed in this study by using the proposed features APS and APCC. The proposed features are implemented with a set of three psychophysics conditions of hearing: critical band spectral estimation, equal loudness hearing curve, and the intensity loudness power law of hearing. The features provided good results for pathology detection as well as for classifications and they are also visually appealing to differentiate between normal and pathological voices. This characteristic of the proposed features

can thus convince medical doctors to make a decision even without a classifier. The higher bands in the APS features model the noisy components owing to lack of closure of vocal folds in pathological subjects, whereas APPC features provide a fine harmonic structure for normal subjects and are sparse for pathological subjects.

By observing different studies in the literature, it can be concluded that MFCC behaves like a clinician because for a clinician it is easier to differentiate between normal and pathological subjects by auditory perception than to discriminate between different disorders. In [30], MFCC did not provide good results to differentiate between the voice disorders compared with other features used in the study, while the performance of the MFCCs were much better when all normal subjects were combined in one class and all disorders were grouped in second class. Studies [9] and [11] also strengthen the conclusion that MFCC perform better for disorder detection rather than disorder classification. MFCC provided a detection rate of 97.46% in [9] with the Rainbow passage, and an average accuracy of 70% was obtained in [11] when MFCC were used for the classification of disorders. MFCC were implemented by using one of the hearing conditions, referred to as a critical band. The developed system based on the proposed system also compared other types of features in Table 8 to ensure that it achieved best accuracy.

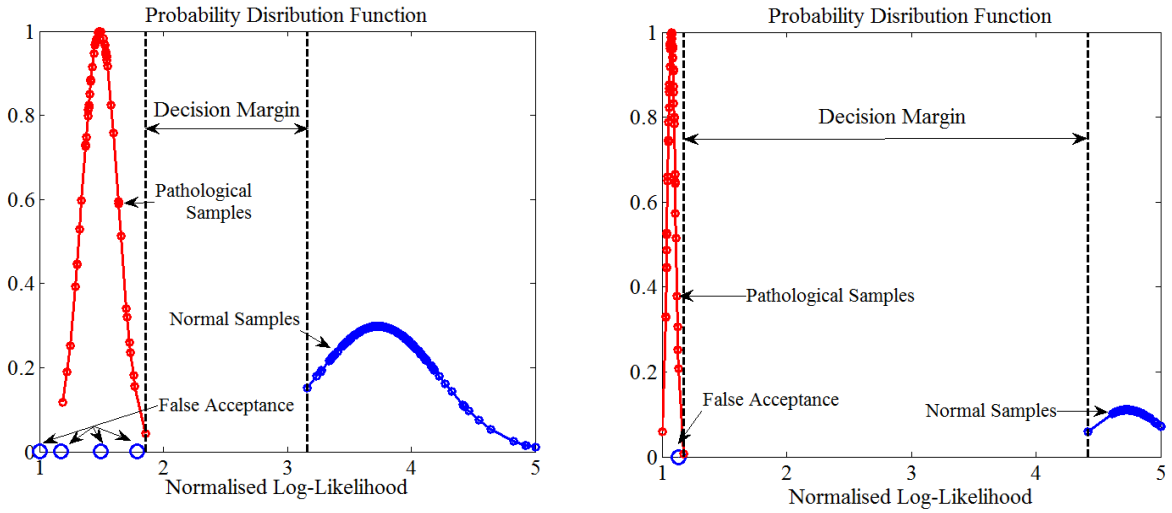
In this study, the obtained detection rate of 99.56% with an STD equal to 0.9 and AUC of 99.99% shows that the proposed features performed well in discriminating between normal and pathological subjects. The proposed features also provided good accuracy in differentiating between different voice disorders. The best-obtained accuracy is 93.33% for the adductor, with an AUC equal to 94.01%.

Both features, APS and APCC, have many dimensions, and they required a multidimensional analysis whose interpretation for the human mind is not easy. Studies based on these types of multidimensional features need a machine-learning stage to make a decision for a test utterance [10]. In the proposed study, the output of the GMM classifier, which is log-likelihood, might be considered as a discriminant measurement to differentiate between the samples of different classes. The significance of the results depends on the discriminative power of the computed log-likelihood.

To practice the developed systems in a clinic, let us consider that Θ_n denotes the GMM model of the normal subjects, and Θ_p represents the GMM model of the disordered subjects. The parametric representation X of the test utterance will get two log-likelihood values, $\log p\left(X \mid \Theta_n\right)$ when compared with a GMM model of the normal subjects, and $\log p\left(X \mid \Theta_p\right)$ when compared with the model of disordered subjects. A threshold ϕ can be used to decide the class. If $\log p\left(X \mid \Theta_p\right) < \phi$, the utterance is a pathological sample, otherwise, the utterance is a normal subject. The probability distribution function (pdf) is plotted for the log-likelihood values of normal and pathological samples to observe the significance of the computed log-likelihood. The pdf for the APS and APCC features are depicted in Fig. 11(a) and 11(b), respectively, where log-likelihood values are normalized over a scale 1 to 5.

The log-likelihood values used in the figures correspond to the best-achieved accuracies mentioned in Table 3 for APS, and in Table 5 for APCC. By analyzing Fig. 11, it can be observed that the log-likelihoods for normal and pathological subjects for both types of features are significantly different. The decision margins for differentiation of normal and pathological samples, for both types of features, are large and exhibit the significance of the proposed features. The decision interval for the APS features is I_S = [1.85 3.16], and for APCC features it is I_A= [1.17 4.41]. For clinical practice, the threshold ϕ can be any value that lies within the decision intervals I_S and I_A for APS and APCC, respectively.

A common threshold can also be adjusted for both features by selecting it from the intersection of the decision intervals I_S and I_A. A common decision margin for both features is I_S. Moreover, two sided *p*-values for the log-likelihoods 0.0001E-22 and 0.0004E-24 are obtained by performing a Mann-Whitney U-test with features at the 5% significant level (α = 0.05). The *p*-values are approximately zero, which rejects the null hypothesis and concludes that log-likelihood can be used reliably to differentiate between normal and pathological classes.



(a) (b)

Figure 11. Probability density functions of log-likelihood values for features (a) APS and (b) APCC. Fig. 11(a) represents that log-likelihood values of pathological subjects for APS features are less than 2.0. It shows that if for a sample log-likelihood value is 2.0 or less, then the subject is suffering from a voice disorder. Fig. 11(b) shows that log-likelihood values of pathological subjects are less than 1.5 for APCC features. It represents that if log-likelihood value of a sample is less than 1.5, then it belongs to pathological subjects. A common threshold value, say 2.0, can also be used for both types of features to decide the presence of a voice disorder.

The significance of the results for pathology classification is presented in Table 8. The p-values <0.05 show that both types of features can differentiate between types of disorders. As mentioned in a recent study [20], little research has been conducted on pathology classification, and all of it used the sustained vowel. The highest result reported in [20] to detect the type of disorder was 76.2%. In our proposed work, the best-obtained result is 93.33%.

The results of the proposed detection system are compared with the results of existing systems in the literature. The accuracies of the existing systems are taken from the respective studies mentioned in the first column of Table 9. A system that uses a VAD is mentioned by "Yes," otherwise "No." From Table 9, we can find that the proposed system outperformed all of the mentioned works. The proposed system works without a VAD but still manages to perform accurately.

The best accuracy in Table 9 without a VAD is 93.4% [6], and by using a VAD the accuracy is 96.5% [17]. The proposed system achieved better accuracy than both of these systems.

Table 9: Comparison of proposed detection system with existing systems

Reference	Database	VAD	Features	Accuracy
[17]	MEEI	Yes	9 acoustic parameters (shimmer, jitter, etc.)	96.5%
[6]	MEEI	No	Frequency ratio, energy ratio, length ratio, and octave mean	93.4%
[10]	MEEI	Yes	MFCC	96%
[31]	MEEI	Yes	Nonlinear dynamic measures	95%
[16]	Private	No	MFCC, shimmer, jitter, and HNR	86%
[12]	Private	No	MFCC	91.66%
Proposed System	MEEI	No	APS and APCC	99.56%

6 Conclusion

Automatic voice pathology detection and classification are developed by using the proposed features APS and APCC. The developed system does not contain a VAD module because extraction of the voiced and unvoiced parts of a speech signal is itself a challenging task and increases the computational cost of a system. The systems are developed by means of running speech owing to its dynamic aspects (onset, offset, etc.). Moreover, running speech is more suitable for screening purposes in the context of daily communication.

The proposed features showed a good performance for pathology detection as well as for classification. The features are also visually appealing to differentiate between normal and pathological voices; this characteristic of the proposed features can thus convince medical doctors to make a decision even without a classifier. The APS features are calculated by using a set of hearing principles which show that high energy for normal subjects belongs to the low-frequency region. This is significantly different from pathological subjects, where high energy appeared in the high-frequency region. Moreover, APCC features provided a fine harmonic structure for normal persons but a sparse structure for vocal-fold patients. An accuracy of 99.56% is achieved for pathology detection and 93.33% for classification. According to the best of our knowledge, this is the highest obtained accuracy for any running-speech-based pathology detection and classification system.

In a future work, we will investigate the proposed system using other databases and try to find the energy differences for different phonemes in a running speech.

Acknowledgment

This project was funded by the National Plan for Sciences, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-MED2474-02).

Appendix A

“When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above…”

References

[1] M. Markaki and Y. Stylianou, "Voice Pathology Detection and Discrimination Based on Modulation Spectral Features," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1938-1948, 2011.

[2] J. W. Lee, H. G. Kang, J. Y. Choi, and Y. I. Son, "An Investigation of Vocal Tract Characteristics for Acoustic Discrimination of Pathological Voices," *Biomed Research International,* vol. 2013, pp. 1-11, 2013.

[3] L. Jung-Won, S. Kim, and K. Hong-Goo, "Detecting pathological speech using contour modeling of harmonic-to-noise ratio," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5969-5973, 2014.

[4] G. Muhammad and M. Melhem, "Pathological voice detection and binary classification using MPEG-7 audio features," *Biomedical Signal Processing and Control,* vol. 11, pp. 1-9, 2014.

[5] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Otolaryngol,* vol. 90, pp. 441-51, 1980.

[6] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *IEEE Transactions on Biomedical Engineering, ,* vol. 52, pp. 421-430, 2005.

[7] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Communication,* vol. 50, pp. 312-322, 2008.

[8] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J Acoust Soc Am,* vol. 55, pp. 1304-22, 1974.

[9] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," *Proc. of the Second Joint EMBS/BMES Conference,* vol. 1, pp. 182-183 vol.1, 2002.

[10] J. I. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, and P. Gómez-Vilda, "Automatic detection of voice impairments from text-dependent running speech," *Biomedical Signal Processing and Control,* vol. 4, pp. 176-182, 2009.

[11] A. A. Dibazar, T. W. Berger, and S. S. Narayanan, "Pathological Voice Assessment," *Proc. of 28th Annual International Conference of the IEEE* on *Engineering in Medicine and Biology Society, EMBS '06,* pp. 1669-1673, 2006.

[12] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, and T. A. Mesallam, "Vocal fold disorder detection based on continuous speech by using MFCC and GMM," *Proc. of 7th IEEE i*n *GCC Conference and Exhibition (GCC), 2013*, 2013, pp. 292-297.

[13] Massachusette Eye & Ear Infirmry Voice & Speech LAB, "Disordered Voice Database Model 4337 (Ver. 1.03) ", ed. Boston, MA: Kay Elemetrics Corp, , 1994.

[14] S. Y. Lowell, R. H. Colton, R. T. Kelley, and Y. C. Hahn, "Spectral- and cepstral-based measures during continuous speech: capacity to distinguish dysphonia and consistency within a speaker," *J Voice,* vol. 25, pp. e223-32, 2011.

[15] Y. D. Heman-Ackah, R. J. Heuer, D. D. Michael, R. Ostrowski, M. Horman, M. M. Baroody, *et al.*, "Cepstral peak prominence: a more reliable measure of dysphonia," *Ann Otol Rhinol Laryngol,* vol. 112, pp. 324-33, 2003.

[16] V. Klára, I. Viktor, and M. Krisztina, "Voice Disorder Detection on the Basis of Continuous Speech," *Proc. of 5th European Conference of the International Federation for Medical and Biological Engineering*. vol. 37, pp. 86-89, 2012.

[17] V. Parsa and D. G. Jamieson, "Acoustic Discrimination of Pathological Voice: Sustained Vowels Versus Continuous Speech," *J Speech Lang Hear Res,* vol. 44, pp. 327-339, 2001.

[18] Y. Zhang and J. J. Jiang, "Acoustic Analyses of Sustained and Running Voices From Patients With Laryngeal Pathologies," *Journal of Voice,* vol. 22, pp. 1-9, 2008.

[19] C. R. Watts and S. N. Awan, "Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts," *Journal of Speech, Language, and Hearing Research,* vol. 54, pp. 1525-1537, 2011.

[20] H. Cordeiro, C. Meneses, and J. Fonseca, "Continuous Speech Classification Systems for Voice Pathologies Identification," *Proc. of Technological Innovation for Cloud-Based Engineering Systems*. vol. 450, pp. 217-224, 2015.

[21] P. L. Dhingra and S. Dhingra, *Diseases of ear, nose and throat*, 6 ed.: Elsevier, 2014.

[22] E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America,* vol. 33, pp. 248-248, 1961.

[23] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America,* vol. 66, pp. 1647-1652, 1979.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*: Springer-Verlag New York, 2006.

[25] J. Yang, X. Yuan, X. Liao, P. Llull, D. J. Brady, G. Sapiro, *et al.*, "Video Compressive Sensing Using Gaussian Mixture Models," *Image Processing, IEEE Transactions on,* vol. 23, pp. 4863-4878, 2014.

[26] J. I. Godino-Llorente, P. Gómez-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, pp. 1943-1953, 2006.

[27] T. H. Falk and C. Wai-Yip, "Nonintrusive speech quality estimation using Gaussian mixture models," *IEEE Signal Processing Letters*, vol. 13, pp. 108-111, 2006.

[28] R. A. Redner and H. F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, vol. 26, pp. 195-239, 1984.

[29] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, A. Mahmood, and M. Alsulaiman, "Multidirectional regression (MDR)-based features for automatic voice disorder detection," *J Voice*, vol. 26, pp. 817 e19-27, 2012.

[30] B. G. Aguiar Neto, J. M. Fechine, S. C. Costa, and M. Muppa, "Feature Estimation for Vocal Fold Edema Detection Using Short-Term Cepstral Analysis," *Proc. of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2007*, pp. 1158-1162, 2007.

[31] J. R. Orozco, J. F. Vargas, J. B. Alonso, M. A. Ferrer, C. M. Travieso, and P. Henriquez, "Voice pathology detection in continuous speech using nonlinear dynamics," *Proc. of 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pp. 1030-1033, 2012.