Pathological voice recognition for vocal fold Disease

Pravena D
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham
Amrita School of Engineering,
Coimbatore-641112

Dhivya S
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham
Amrita School of Engineering,
Coimbatore-641112

Durga Devi A
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham
Amrita School of Engineering,
Coimbatore-641112

ABSTRACT

Pathology is the study and diagnosis of disease. Due to the nature of job, unhealthy habits and voice abuse, the people are subjected to the risk of voice problems. The diagnosis of vocal and voice disorders should be in the early stage otherwise it causes changes in the normal signal. It is well known that most of vocal fold pathologies cause changes in the acoustic voice signal. Therefore, the voice signal can be a useful tool to diagnose them. Acoustic voice analysis can be used to characterize the pathological voices. This paper presents the detection of vocal fold pathology with the aid of the speech signal recorded from the patients. We are going to recognize the disordered voice for vocal fold disease by focusing on the classification of pathological voice from healthy voice based on acoustic features. The method includes two steps. The first step is the extraction of feature vectors based on MFCC. The second is the classification of feature vectors using GMM. The extracted acoustic parameters from the voice signals are used as an input for the MFCC. The main advantage of this method is less computation time and possibility of real-time system development. This report introduces the design and implementation of the proposed system for recognizing pathological and normal voice. Also a description is given about the literature survey done and the implementation of different modules in the system. The result of the proposed system and the scope of improvements are also discussed in the report.

General Terms

Pathological voice, acoustic features.

Keywords

MFCC, GMM.

1. INTRODUCTION

Pathology is the study and diagnosis of disease. In this paper we are going to recognize the disordered voice for vocal fold disease. Vocal fold disease can affect the quality of the sound which is produced from vocal cord. The presence of pathologies in vocal folds causes significant changes in the normal vibratory patterns, which will results in the quality of voice production. The problems in the production of voice are due to the 1) functional disorder (due to the abuse or wrong use of the anatomical and physiologically intact voice system) or 2) Laryngeal pathologies (nodules of vocal folds. polyps, ulcers, carcinomas and paralysis of the laryngeal nerve. Some of the more common vocal cord disorders include laryngitis, vocal nodules, vocal polyps, and vocal cord paralysis. Diagnosis of pathological voice is one of the most important issues in biomedical applications of speech technology. In the past 20 years, a significant attention has been paid to the science of voice pathology diagnostic and monitoring. Normally physicians often use invasive technique like Endoscopy to diagnose the symptoms of vocal fold disorder [2]. Furthermore, the irregular vocal fold oscillations can be observed by means of a digital high-speed camera using image processing techniques in order to extract the vocal fold edges, estimate the minimum glottal area defined by the vocal fold positions, and compute the distance between the glottal midline and the vocal fold edges extracted at medial position in real-time. Voice pathologies may be assessed by either perceptual judgments or an objective assessment. The perceptual judgment resorts to qualifying and quantifying the vocal pathology by listening to patient's speech. Although this is the most commonly used method by clinicians, it suffers from several drawbacks. First of all, the perceptual judgment has to be performed by an expert jury in order to increase its reliability. Second, due to the lack of universal assessment scales and the dependence on expert's professional back-ground and experience or the knowledge of patients history, the perceptual judgment may involve large intra and inter-variability. Third, the perceptual analysis is very costly in time and human resources and cannot be planned regularly. Nowadays an increasing use of objective measurement-based analysis as a non-invasive technique for supporting diagnosis in laryngeal pathology has been observed [4]-[6]. Objective measurement-based analysis qualifies and quantifies the voice pathology by analyzing acoustical, aerodynamic, and physiological measurements. These measurements may be directly extracted from patient's speech utterance using a simple computer-based system or may require special instruments. The purpose of this work is to help patients with pathological problems for monitoring their progress over the course of voice therapy. Currently, patients are required to routinely visit a specialist to follow up their progress. Moreover, the traditional ways to diagnose voice pathology are subjective, and depending on the experience of the specialist, different evaluations can be resulted. Developing an automated technique saves time for both the patients and the specialist can improve the accuracy of the assessments. Through acoustic analysis, finding out which factors that affect the human voice production mechanism can lead to the noninvasive diagnosis of disease. Developing an automatic pathological voice classification is training a classification system which enables to automatically categorize any input voice as either normal or pathological. Once the signal features are extracted, if the extracted features are well defined, even simple classification methods will be good enough for classification of the data.

The objective of this work is the search for a technique that will allow the quantification of a speaker's voice quality by means of an audio sample. This technique will allow us not only to show the evolution of the patients voice quality throughout the treatment, but it could also be applied in the field of preventive medicine in order to achieve early detection of laryngeal pathologies. Furthermore MFCC based voice parameters are presented in this work in order to compare healthy and pathological voice samples. MFCC composed of 12 static features and energy plus delta features plus delta-delta features [7]. Cepstral coefficients may follow any statistical distribution on different speech segments; the well-known Gaussian Mixture Model (GMM) approach was chosen to fit a flexible parametric distribution to the statistical distribution of the selected speech segment. This technique increases the robustness of the models especially when sparse speech materials are available. We used GMMs to study nasalization in speech, comparing the voices of severe disease patients with those in a healthy control group [8].

2. SCOPE OF THE WORK

My work is to classify normal and pathological voice using acoustic feature MFCC and to classify the signal using GMM. In this project we are going to classify the cough, coughed speech, fan noise, white Gaussian noise and normal voice. This is done by extracting the features from the signal. Feature extraction is the first step in any speaker recognition system. MFCC is one of the most popular feature vectors. The cepstral representation of the signal allows us to characterize the vocal tract as a source-filter model and the Mel frequency characterizes the human auditory system which perceives the sound in a nonlinear frequency binning. GMM was used as classifier for speaker identification application.

3. MATERIALS AND METHODS

The patterns for training the MFFCs were obtained from recordings of people’s voices with normal voice and patients with pathologies of-the vocal system. Each signal is a recording of the sentence. In our implementation, two requirements were imposed. First, the features had to be efficient in terms of measurement cost and time. Second, both the vocal tract and excitation source information had to be included. The MFCC features were obtained by a standard short-term speech analysis, along with frame-level pitch, to form the feature vectors. Then, set of Gaussian mixture Model GMM classifiers were applied for the assessment of feature vectors. The architecture of the proposed system is given in figure 1

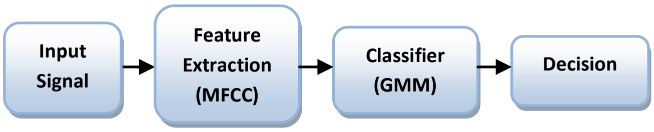

Fig 1: System Architecture

3.1 Feature Extractor

An important step in both training and classification stages of pattern recognition is the selection and extraction of the features. The features that can be used for speech recognition can be broadly divided as time domain features and frequency domain features or spectral features. The time domain features includes zero-crossing rate, average energy, maximum amplitude etc. The frequency domain features includes power spectral analysis, Linear Predictive Coding, Mel-Frequency Cepstral Coefficients. Even though time domain features are easy to extract, the most used feature is frequency domain feature, because of the insight they give into the relationship between the speech signals and the manner of articulation by the vocal organs. In our proposed system, MFCC features are used. For finding the best match in the knowledge base (speech models), for the incoming feature vectors, Gaussian Mixture Model based recognition component is used.

3.2 Mel Frequency Cepstral Coefficient

The most widely used feature vector is MFCC. The advantage of MFCC is that it takes into account the perceptual characteristics of the human ear. Psychophysical studies have shown that the frequencies perceived by the human ear are in a nonlinear logarithmic scale rather than in a linear scale and the frequencies are perceived in a nonlinear frequency binning (critical band filtering). The nonlinear scale is characterized by Mel scale and critical band filtering is characterized by Mel filter bank.

3.2.1 Mel Scale

Mel scale is a perceptual scale of pitches calculated by the judgment of listeners. The unit of 2 measurements in Mel scale is Mel. The equation relating frequency scale and Mel scale is

$$f_{mel} = 2595 * \log\left(1 + f/_{700}\right)$$

3.2.2 Critical bank filtering

The human ear is said to consist of bank of auditory filters that enhances certain frequencies and attenuates other. The bandwidths of these filters are known as critical bands. In order to model this characteristic of the human ear we design a bank of filters known as Mel filter bank. The most commonly used type is a bank of triangular filters which is shown in the figure 2.

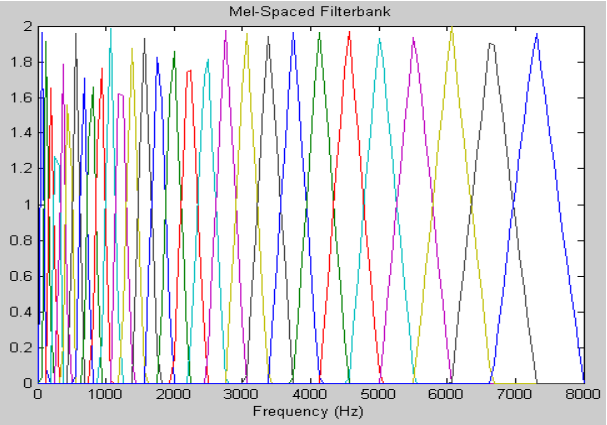

Fig 2: Mel Filter Bank

3.3 MFCC Extraction

The general procedure for generating MFCC is as follows:

1) Taking the Fourier Transform of the signal.

2) Mel filtering in the frequency domain to get the Mel filter bank coefficients.

3) Taking logarithm of the Mel filter bank coefficients.

4) Taking the Discrete Cosine Transform (DCT) of the log Mel filter bank coefficients.

Before extracting the coefficients, preprocessing has to be done for the signal. The preprocessing steps include: DC offset removal, pre-emphasis, normalization, framing and windowing. Figure3 gives the block diagram for the procedure.

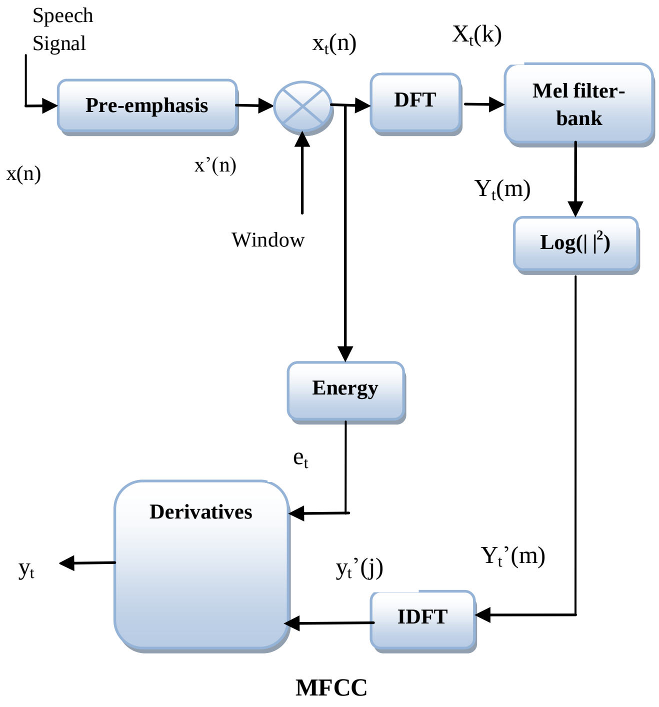

Fig 3: Block diagram of MFCC procedure

3.4 DC offset removal and pre-emphasis

DC offset is removed by subtracting the mean value from the signal. The following figure5 shows a speech signal before and after DC offset removal. To flatten the spectrum of the signal, pre-emphasis of approximately 20 dB per decade is done on the spectrum of the speech signal. Pre-emphasis filter is used to offset the negative spectral slope of the voiced speech signal. This is done to improve the efficiency of further analysis. Transfer function of a typical pre-emphasis filter is $H(Z)=1-kz^{-1}$

3.5 Framing

Speech signals are slowly time varying and can be treated as stationary when considered under a short time frame. Therefore, the speech signal is separated into small duration blocks, called frames, and further analysis is performed on these frames. The commonly used frame length and frame shift in speaker recognition are 20-30ms and 10ms respectively. This is because the vocal tract shape remains almost constant during a period of 30ms. Figure 4 shows framing.

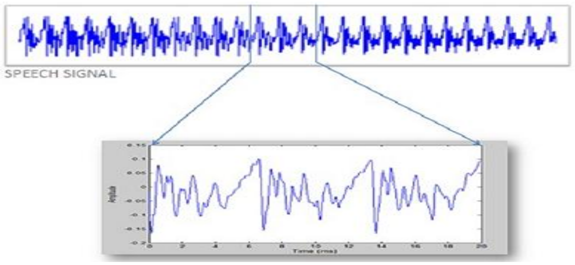

Fig 4: Framing

3.6 Windowing

After partitioning the speech signal into frames, each frame is multiplied by a window function. Windowing is done to reduce the effect of discontinuity introduced by the framing process by attenuating the values at the beginning and end of each frame. Hamming, Hanning and Blackman windows are used commonly. In the following figure5a is the most widely used Hamming window and figure 5b shows a single frame is multiplied by Hamming window and the resulting signal is shown.

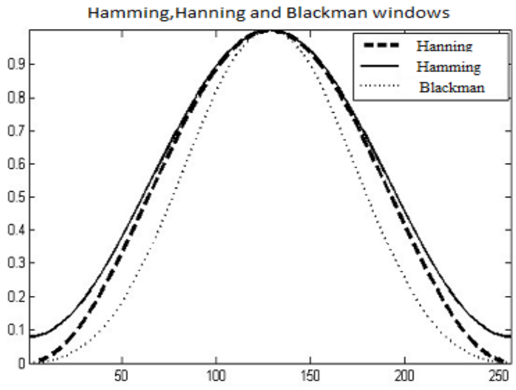

Fig 5a: Hamming window

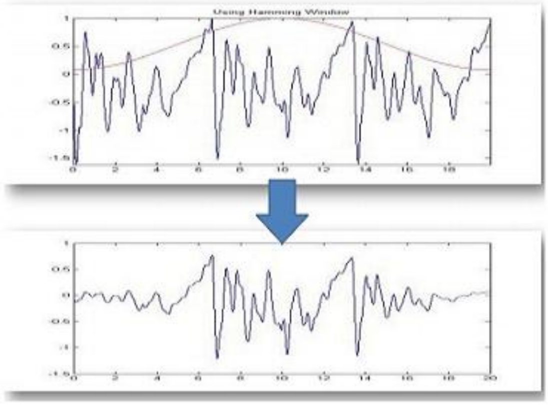

Fig 5b: Windowing

3.7 Discrete Fourier Transform

The spectral coefficients of the signal are calculated by taking Discrete Fourier Transform (DFT) of the signal using the formula:

$$x(k)=\sum_{n=0}^{N-1} x(n)e^{-j2\Pi kn/N}$$

Where k = 0, 1, 2....N-1 corresponds to frequency $f(k) = k\, f_s / N$ and x(n) is the windowed frame of length N samples.

The coefficients X (k) thus obtained are complex and contain both magnitude and phase information. As the phase component is not perceived by the ear, only the magnitude is retained for feature extraction. Taking DFT spectrum is more advantageous than taking linear predictive coding (LPC) spectrum. The LPC spectrum is a parametric estimate of the smoothed spectral envelope, while the DFT spectrum provides more details of the spectrum of the speech frame.

3.8 Mel Filtering

The magnitude spectrum coefficients $|x(k)|$ are then multiplied with Mel filter bank designed using the previously described method to get the Mel filter bank coefficients.

$x'(m) = \sum_{n=0}^{N-1} |x(k)|.H(k,m)$ for m = 1, 2, 3, ,M, where M is the number of filters in the filter banks.

3.9 Natural Logarithm

Natural logarithm is applied on Mel filter bank coefficients to get log-Mel filter bank coefficients. This characterizes the nonlinearity in the loudness and the sound intensity. Taking logarithm converts multiplication relation into addition relation. Besides this, it converts the multiplication relationship between parameters into addition relationship.

3.10 Discrete Cosine Transform

The DCT is applied on the log Mel filter bank coefficients to generate the cepstral coefficients. The coefficients after DCT becomes less correlated, therefore, it is possible to use diagonal matrix of the Gaussian in the Hidden Markov Model (HMM), and this significantly reduces the number of parameters in the acoustical model. The coefficients obtained after performing DCT are Mel frequency cepstral coefficients.

$c(l) = \sum_{m=1}^{M} \ln(x'(m)) \cos\left(l \frac{\Pi}{M} \left(m - \frac{1}{2} \right) \right)$ for l=1,2,…..M

3.11 Log energy calculation

In addition to the normal MFCC features, the energy of the speech frame is also used as a feature. The log energy, log E, is calculated directly from the time domain signal of a frame. Sometimes, it is replaced by, c_0 the 0th MFCC coefficient.

3.12 Derivatives and Acceleration Calculation

The trend of the speech signal in time is lost by frame by frame analysis. To recover this information, the first and second derivatives are calculated and concatenated to the MFCC coefficients to get a larger feature vector. Usually, we take 12 MFCC coefficients,1 energy coefficient,13 first and 13 second derivatives to get a 39 dimensional feature vector.

4. CLASSIFIER

4.1 Gaussian Mixture Model

Once the MFCC features from a given speech signal is extracted, we have to make model to recognize the speaker. A common way to model a speaker's voice in text-independent speaker recognition system is to build a Gaussian mixture model [21] with the features vectors from the training samples. After training, the feature vectors of test utterance speech signal is given to GMM model to compute the probability that the test sample belongs to a particular speaker.

4.1.1 Model description

A Gaussian mixture density is a weighted sum of M component densities, and given by the following equation $p(\vec{x}/\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x})$ where $\vec{x}$ is a D-dimensional random vector $b_i(\vec{x})$ i=1,…,M, are the component densities and $p_i(\vec{x})$ i=1,…,M, are the mixture weights.

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{ p_i, \vec{\mu}_i, \Sigma_i \}, i = 1,...M$$

For speaker identification each speaker is represented by a GMM and is referred to by his/her model λ .

4.1.2 Maximum Likelihood Parameter Estimation

The aim of Maximum Likelihood (ML) estimation is to find the model parameters which maximize the likelihood of GMM. For a sequence of T training vectors $X = \{ \vec{x}, ..., \vec{x}_T \}$ the GMM likelihood can be written as

$$p(x/\lambda) = \prod_{t=1}^{T} p(\vec{x}/\lambda)$$

This is a nonlinear function of λ and direct maximization is not possible. However ML parameters estimates can be obtained iteratively using a special case of the expectation-maximization (EM) algorithm and this is discussed in more detail in [24].

5. RESULTS AND DISCUSSION

The system records the human speech through a microphone and then the captured speech processed to recognize the uttered text. The system consists of two phases, Training Phase and Testing Phase.

5.1 Training Phase

In the training phase, many utterance of the same word is recorded from different people. From the recorded corpus, models for each of the words are created. The system is trained to learn the reference pattern that represents each of the words. Different samples of the same word are recorded from different speakers. The system cannot process the speech waveform directly. So the recorded analog speech signal is digitized. Then the words are forwarded to the Feature Extractor. The MFCC feature vectors of the each of the sample words are computed by the feature extractor. After the computation of feature vectors for all the words, the GMM based recognition component creates the speech model for each of the words. This is a supervised training procedure. Simple architecture is shown in the figure 6.

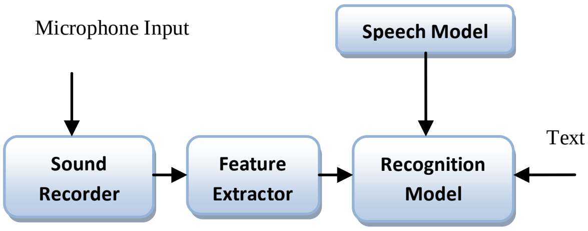

Fig 6: Training architecture

The recording was done using the sound recorder, which was implemented. The sound recorder was implemented such a way that it gets activated, when the speaker starts speaking. The idea behind the automation of the recorder is that, an energy threshold was set and the energy of the microphone input exceeds this, it starts recording and continues until the energy goes below the threshold. So when the speaker starts speaking, the energy component of the microphone input will increase than when it was in the silent case. And as the energy component of the microphone input exceeds the threshold value, sound recorder will get activated. In training phase, the voice of the speaker is recorded. After detecting the word from the input sound signal, it forwarded to the feature extraction module, where the MFCC feature-vectors are computed. The feature-vectors for different speech are stored separately. This process is continued for all the utterances of the words that need to be trained. The features-vectors are then forwarded to the GMM based recognition module, where speech models corresponding to feature vectors of each word in the training corpus are created. A supervised training is followed here, as the text for the corresponding sound signal is supplied to the system while training.

5.2 Testing Phase

After training, the system is ready to use. This stage of the system is known as the recognition phase, shown in Fig11, where system accepts input from the speaker through a microphone and gives the recognized output. Here, the user can speak any number of words with a small pause in between the words. As in training phase, the sound recorder forwards the input sound signal to the feature extractor. In this phase, the feature extractor performs the same functions as they were doing in the training phase. The basic flow once the training is done can be summarized as the sound input is taken from the sound recorder and after the words are identified, it is feed to the feature extraction module. The feature extraction module generates feature vectors out of it and then forwards it to the recognition component. The recognition component with the help of the speech model and comes up with the result, that is, after comparing with the models, the model which comes closer is declared as the recognized word. i.e. feature vectors are computed. The Feature-Vectors are then forwarded to the GMM based recognition module, where the input pattern is compared with the speech models created in the training stage, and the model coming closer to the input pattern is declared as the recognized result. Figure 7 shows the testing architecture

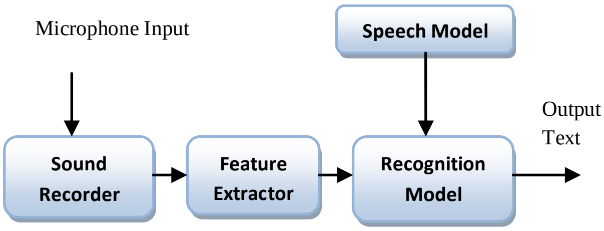

Fig 7: Testing architecture

5.3 Experiments on database

5.3.1 *Database*

The experiment was performed on a closed set database of 30 speakers. The database contained speech of 60 seconds for every speaker. The sampling frequency of the recording was 16 kHz.

5.3.2 *Preprocessing*

To improve the performance of the system and to reduce the effects of noise, preprocessing steps like pre-emphasis and DC offset removal was performed on the signal.

5.3.3 *Framing and windowing*

The speech signal was split into frames of length 20 ms (320 samples). The frame shift was taken to be 10 ms (160 samples). Hamming window was applied to each frame to avoid abrupt discontinuities.

5.3.4 *Feature Extraction*

The MFCC coefficients are computed for each frame separately. The functions melbankm and melcepst from voice box toolbox are used to design the mel filter bank and compute the MFCC coefficients respectively. Twelve filter banks are taken in the frequency range of 0 - 8 kHz. After taking DCT, 11 coefficients (except the 0th coefficient) are taken as MFCC coefficients and stored.

5.3.5 *Training the system*

The GMM system was trained with the first 30 seconds of the speech data available for each user. The training was done using the built-in MATLAB function gmdistribution.fit. The covariance matrix type 30 was set to be diagonal and the system was trained for GMM orders 16, 32 and 8. The function returns an object for the speakers.

5.3.6 *Testing*

The remaining 30 seconds of the speech of every speaker was used to test the system. The feature vectors were extracted for the test speech and the probability was calculated for each speaker using the MATLAB function posterior. The function returns the negative log of the probability. So, the speaker object for which the function returns the least value is taken as the correct speaker.

5.3.7 *Performance Measure*

This table shows the performance measures for different GMM order used during testing phase.

Table 1: Performance Measure

No. Of Mixtures	Training Speakers	Testing Speakers	Accuracy
8	20	10	83 %
16	20	10	98 %
32	20	10	95

6. CONCLUSION

This project work can be considered as an initiative to develop a automatic pathological voice recognition system. This is an isolated speech recognition system. Mel frequency cepstral coefficients have been used for generating feature vector for recognition. During training, speech models for each word in the training corpus are generated using Gaussian mixture models. This system can be viewed as a prototype, so that several applications can be developed using this system. Currently the system has been trained to recognize the cough, coughed speech, normal speech, fan noise, white Gaussian noise and pathological speech. Training data has been collected through many speakers and this makes the system speaker independent. The system gives good recognition accuracy for those speakers who were involved in the training set creation, and for other speakers the result was satisfactory.

7. REFERENCES

[1] Ce Peng., Wenxi Chen., Xin Zhu.,Daming Wei., and Baikun Wan., (2007). Pathological Voice Classification Based on a Single Vowels Acoustic Features. IEEE computer society. Seventh International Conference on Computer and Information Technology.,.

[2] lagadish Nayak ., Subhanna Bhat , P.,(2003). Identification of Voice Disorders using Speech Samples ," IEEE Trans. Speech processing.,

[3] Schwarz, R., Hoppe, U.,Schuster, M., Wurzbacher,U., and Eysholdt., Lohscheller, J. (2006). Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model. IEEE Trans. Biomed.Eng., vol. 53, no. 6, pp. 10991108.

[4] Constantine Kotropoulos ., and Gonzalo R. Arce. (2009). Linear Classifier with Reject Option for the Detection of Vocal Fold Paralysis and Vocal Fold Edema. EURASIP Journal on Advances in Signal Processing,

[5] Parsa, V., and Jamieson, D.G. (2003). Interactions between speech coders and disordered speech. Speech Communication.,vol.40, no. 7, pp. 365385,

[6] Hawley,M. S., Green ,P.,Enderby, P., Cunningham, S., and Moore, R.K., (2005). Speech technology for e-inclusion of people with physical disabilities and disordered speech. INTERSPEECH 05.,pp. 445448, Lisbon,Portugal

[7] Oscar Saz., Javier Simon.,Ricardo Rodriguez, W., Eduardo Lleida., and Carlos Vaquero., (2009). Analysis of Acoustic Features in Speakers with Cognitive Disorders and Speech Impairments EURASIP Journal on Advances in Signal Processing,

[8] Ruben Fernandez Pozo., Jose Luis BlancoMurillo.,Luis Hernandez Gomez., Eduardo Lopez Gonzalo.,JoseAlcazar Ramirez and Doroteo T. Toledano., (2009). Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker Recognition Techniques EURASIP Journal on Advances in Signal Processing,

[9] Alireza Afshordi Dibazar., Shikanth Narayanan.,A System for Automatic Detection of Pathological Speech (2002)

[10] Julian D. Arias-Londono., Juan I. Godino-Llorente.,Nicolas Saenz-Lechon.,Victor Osma-Ruiz., and German Castellanos-Dominguez., Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients (2011) IEEE Trans. on biomedical engineering vol. 58, no. 2.,

[11] Darcio G. Silva., Luis C. Oliveira.,and Mario Andrea., Jitter Estimation Algorithms for Detection of Pathological Voices(2009) EURASIP Journal on Advances in Signal Processing

[12] Jianglin Wang., cheolwoo Jo.,Vocal Folds Disorder Detection using Pattern Recognition Methods (2007) IEEE EMBS ,23-26(8).

[13] Karthikeyan Umapathy., Sridhar Krishnan.,Donald G.Jamieson.,Discrimination of Pathological Voices Using a Time-Frequency Approach (2005) IEEE Trans. on biomedical engineering vol. 52, no. 3.,

[14] Maria Markaki., Yannis Stylianou.,Voice Pathology Detection and Discrimination Based on Modulation Spectral Features (2011) IEEE Trans. on audio, speech, and language processing vol. 19, no. 7.,

[15]Fetisova,O.G., Lamtyugin,D.V.,Makukha,V.K.,Voronin,E.M.,Spectrum analysis of vocalization application for voice pathol-ogy detection (2007) IEEE Trans. The International Conference on Computer as a Tool .,

[16] Alireza A. Dibazar., Theodore W. Berger.,Shrikanth S. Narayanan.,Pathological Voice Assessment (2006) IEEE Trans.EMBS Annual International Conference .,

[17] Ghazaleh vaziri ., Farshad Almasganj .,PATHOLOGICAL ASSESSMENT OF VOCAL FOLD NODULES AND POLYP VIA FRACTAL DIMENSION OF PATIENTS VOICES (2008) IEEE Trans. Iran National Science Foundation .,

[18] Paulo Rogerio Scalassara., Maria Eugenia Dajer., Jamille Lays Marrara., Carlos Dias Maciel., Jose Carlos Pereira.,Analysis of Voice Pathology Evolution Using Entropy Rate (2008) IEEE Trans. International Symposium on Multimedia .,

[19] Patricia Henriquez., Jesus B. Alonso., Miguel A. Ferrer., Carlos M. Travieso., Juan I. Godino-Llorente., and Fernando Diaz-de-Maria.,Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics (20009) IEEE Trans. AUDIO, SPEECH, AND LANGUAGE PROCESSING VOL. 17, NO. 6.,

[20] Mark Gales., Steve Young., The Applications of Hidden Markov Models in Speech Recognition., Foundations and Trends in Signal Processing, 1,3(2007),195-304.

[21] Reynolds, D. A., and Rose,R. C., Robust text-independent Speaker Identification using Gaussian mixture speaker models, 1(2005) IEEE Trans. On Speech and Audio Processing, vol. 3, pp. 7283,

[22] Douglas A. Reynolds., Speaker Identification and verification using Gaussian mixture speaker models, Speech Communi-cation 17 (1995) 91-108, Elsevier.

[23] G. McLachlan, Mixture Models. New York: Marcel Dekker, 1988.

[24] A. Dempster, N. Laird, and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Royal Stat. vol. 39,pp. 1-38,1977.