Voice Pathologies Classification Using GMM And SVM Classifiers.

Amara Fethi
Department of electronic
University of Badji Mokhtar
Annaba, Algeria
amarafethi13@gmail.com

Fezari Mohamed
Department of electronic
University of Badji Mokhtar
Annaba, Algeria
mouradfezari@yahoo.fr

Received: July 15, 2019. Revised: August 13, 2021. Accepted: September 22, 2021. Published: November 27, 2021.

Abstract—**In this paper we investigate the proprieties of automatic speaker recognition (ASR) to develop a system for voice pathologies detection, where the model does not correspond to a speaker but it corresponds to group of patients who shares the same diagnostic. One of essential part in this topic is the database (described later), the samples voices (healthy and pathological) are chosen from a German database which contains many diseases, spasmodic dysphonia is proposed for this study. This problematic can be solved by statistical pattern recognition techniques where we have proposed the mel frequency cepstral coefficients (MFCC) to be modeled first, with gaussian mixture model (GMM) massively used in ASR then, they are modeled with support vector machine (SVM). The obtained results are compared in order to evaluate the more preferment classifier. The performance of each method is evaluated in a term of the accuracy, sensitivity, specificity. The best performance is obtained with 12 coefficientsMFCC, energy and second derivate along SVM with a polynomial kernel function, the classification rate is 90% for normal class and 93% for pathological class.This work is developed under MATLAB**

Keywords-Speech pathologies detection, voice disorders, classifiction, machine learning, laryngeal diseases.

I.INTRODUCTION

Assessment voice quality is an important tool for dysphonia evaluation; it is based on perceptual analysis [1] and instrumental evaluation which comprise acoustic and aerodynamic measure [2], the first one is subjective because of the variability between listeners, although the second is objective it is invasivefor one hand , on the other hand it is has a limited reliability.

This is why the development of automatic system for classification is proposed as a complementary tool with the other mentioned technics, we distinguish three principal approaches: acoustic, parametric and non-parametric approach and statistical methods. The first approach consist to compare acoustic parameters between normal and abnormal voices such as fundamental frequency, jitter, shimmer, harmonic to noise ratio, intensity [3-6]. The evaluation of acoustic parameters depends on the fundamental frequency; the evaluation of the latter is difficult particularly in the presence of Pathology. MDVP and PRAAT are two available software to calculate these parameters [7].

The second approach is the parametric and non-parametric features extraction [8-9].

The classification of voice pathology can be seen as pattern recognition so statistical methods are an important tool to discriminate between normal and pathological voice or to know the disease from a speech signal. The statistical methods are used to mimic the brain comportment where we can recognize persons from their voice. Many researches are realized for this task, the conception of these systems has the same principal steps starting by feature selection then training and at the last testing. Support vector machine (SVM) is applied to test the effectiveness and reliability of the short term cepstral and noise parameters [10] and it is applied on discrete wavelet transform it gave a very promising results[11], GMM is used as classifier with MFCC [12], in [13] the training is supported byHidden Markov Model (HMM). The neural network is massively used for this topic in [14] the MFCCs are proposed to be the input of multi-layerperceptron (MLP).

In this paper the conception of our detector is inspired from a system of ASR [15]. 12 MFCCs, energy, dynamic parameters (first derivate and second derivate) are extracted to be the input of GMM and SVM. The main idea behind this work is to test the efficiency of the cepstral analysis to characterize pathological voices and to compare the performances of the two classifiers.

Although, the developed system is inspired from ASR system there are a principals differences betweenthe two systems which we cannot ignored, we can limited theme in two essential key point

- In ASR the model corresponds to a speaker while the model in second system c orresponds to the group of patients with the same diagnostic.
- In voice pathologies detection samples used for train are different from samples used for test unlike in ASR where the two set is similar.

This paper is organized as a follow: in second section is dedicated to describe different steps to develop the systemand how we designed the two classifiers based on GMM and SVM, theexperiments are presented in section 3. The results are presented in section 4 and the last section is reserved for the conclusion and future work.

II.METHODOLOGY

Our system will pass by the same steps to concept a system for ASR, we will describe theme step by step, the block diagram in “fig1” show different steps adapted to our system.

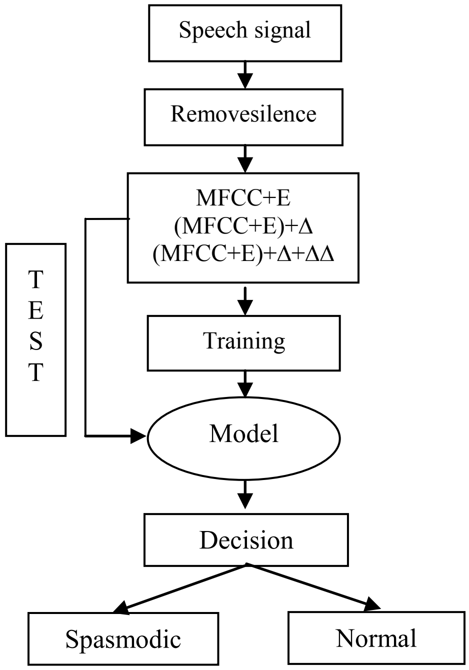

Figure 1 Block diagram adapted to the detector.

A. Speech signal:

In this work the creation of the data base is not our goal so we will not discuss the speech acquisition but we will describe the database which the results are built around it.

The database presents an essential factor to develop a detector where the use of standard one helps to compare the obtained results in order to test the effectiveness and the reliability of methods [14].

In this work we have choose a German database for voice disorder developed by PUTZER [16] which contain healthy and pathological voices, where each one pronounce vowels [i, a, u] /1-2 s in wav format at different pitch (low, normal, high), it contain alsophrase and electroglottographsignal (EGG). All files are sampled at 50 KHz.

From this large database we have select patients suffer from neurological pathology (spasmodic dysphonia), this disease affects women than men that is why we have choose a female voice for training and testing step, Table.1 show the selected samples. As mentioned above the recording files contain phrase, this study is built around the phrase “good morning how are you” pronounced in Germany. The goal to use phrase in one hand is to get more data for training where GMM need an important quantity of data pa rticularly when use a high number of mixture (Gaussian), in other hand the diversity of data enhance the accuracy of a system.

Table1. Description of dataset

	Training set		**Test set**	
	Number	Age	Number	Age
Normal	52	20-60	11	20-60
Pathological	29	30-82	9	30-82

Those files are down sampled to 25 KHz in order to get optimal analysis where the speech signal is considered stationary by frame of 10 to 30s so the use of a very high frequency oblige the use of a large window to get a stationary frame which minimize the size of the extracting features.

B. Pre- processing:

Pre-processing of Speech Signal serves various purposes in any speech processing application. It includes Noise Removal, Endpoint Detection, Pre-emphasis, Framing, Windowing and silence remove. In this this study we are interesting to remove silence knowing that the efficient features are included in speech portion [17].

C. Features extraction:

Features extraction means finding good data allows to categorize the healthy status of patient, features selection make a boundary between each class.

Spasmodic dysphonia is a disorder of vocal function, characterized by spasms of the muscles of the larynx that disrupt or impede the regular flow of voice this leads us to choose the MFCCs parameters in order to split the glottal source from the effect of cavities or filter in order to have a parameters with significant difference between pathological and healthy voices.

MFCC parameters are obtained calculating the Discrete Cosine Transform (DCT) over the logarithm of the energy in several frequency band given by:

$$C_m = \sum_{k=1}^{M} \log(S_k) \cos[m(k-\tfrac{1}{2})]\frac{\Pi}{M} \qquad (1)$$

Temporal derivatives

In order to use the proprieties of the dynamic behavior of speech signal the analysis can be extended to compute the temporal derivate of the MFCC parameters, first derivate (Δ) is given by:

$$\Delta c_m\,[p] \approx \mu \sum_{k=-K}^{K} k c_m\,[p+k] \qquad (2)$$

The second derivate (ΔΔ) are calculated with the same equation.These parameters are calculated thanks to the toolbox voice box with *melcepst* function.

D. Training

This study use two well-known classifier in statistical pattern recognition, GMM and SVM, for one hand, the main idea behind the use of the SVM is that this classifier is performed with organics pathologies (10). In other hand, the comparison with the GMM is recommended where theprevious work was based in GMM classifier. We describe in the follow subsection how to design the two classifiers.

GMM

In pattern recognition (machine learning) the learning is supported by the statistical classifier, Gaussian mixture model (GMMs) consist to represent the data (features) obtained at last step by a simple Gaussian curve described by:

$$P\,(x_i \backslash \lambda) = \sum_{j=1}^{M} p(\,x_i \backslash N_j\,) W_j \qquad (2)$$

$$\sum_{j=1}^{m} W_j = 1 (3)$$

λ is the model.

Each component has the general form:

$$p(x) = \frac{1}{(2\Pi)^{\frac{d}{2}} |\Sigma|^{2}} e^{\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]} \qquad (4)$$

Each cluster is represented by a Gaussian as in “fig 3”

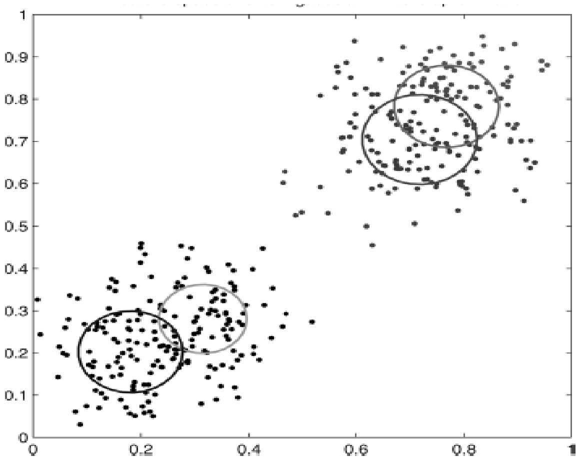

Figure 3. Scatterplot of a two-dimensional (2-D) cepstral vector and its approximation by means of a 2-D Gaussian mixture.[10]

∑ is the d-by–d covariance matrix and |∑| is its determinant itcharacterizes the dispersion of the data on the d-dimensions of the feature vector. The diagonal element σ_{ii} is the variance of x_i, and thenon-diagonal elements are the covariances between features. Often, the assumption is made that the features are independent. Thus, ∑ is diagonal and p(x) can actually be written as the product ofthe univariate probability densities for the elements of x. the proposed model must be optimal, one ideal way to get optimal model this is the use of Maximum likelihood estimation (MLE) given by:

$$p(X \backslash \lambda) = \prod_{i=1}^{M} p(x_i \backslash \lambda) \quad (5)$$

Maximizing the likelihood of observing xas being produced by the patient. Nevertheless, in the case where all the parameters are unknown, the maximum likelihood yields useless singular solutions. Thus there is a need for an alternate method. In literature the use of Expectation Maximization (EM) is the most used solution for this problem. EM is an iterative algorithm starts from initial model calculated here with a K-means algorithm for clustering.

SVM

SVM is a two-class classifier that maximizes the distancebetween nearest points of the two classes. Our task is to predict whether a test sample belongs to one of two classes. We receive training examples of the form :{x_i, y_i}, i = 1,…,n and $x_i \in R^d$, $y_i \in \{1; +1\}$. We call {x_i} the co-variates or input vectors and {y_i} the response variables or labels. We consider a very simple example where the data are in fact linearly separable we can draw a straight line$f(x) = w^T x - b$such that all cases with $y_i = -1$ fall on one side and have$f(x_i) < 0$ and cases with$y_i = +1$ fall on the other and have $f(x_i) > 0$

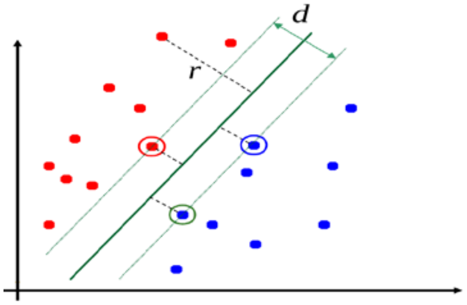

Figure 4. Support vector machine with linear separation

When a data is not linearly separable a kernel function is proposed for better separation as mentioned in “fig5”

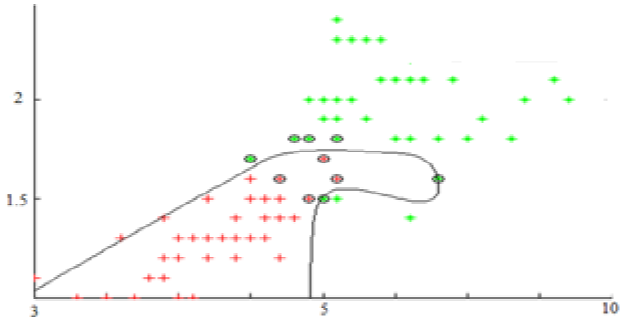

Figure5. SVM with polynomial kernel function.

E. Test step:

Once models are created and that we have managed to train the classifier, we can proceed to the classification test.

For a GMM: anew feature vector Xt is said to belong to an appropriate model if it maximizes p (Xt | λ) for every possible class. For SVM we could classify new test cases according to the rule$y_{test} = \text{sign}(x_{test})$.

In order to evaluate the performance of the system the results are presented by a confusion matrix represented in “Table 2”

Table2. Typical aspect of a confusion matrix

System’s decision	Actual diagnosis	
	Pathological	**Normal**
Pathological	True positive (TP)	False positive (FP)
Normal	False negative (FN)	Truenegative (TN)

True positive (TP) or sensitivity, is the ratio between pathological files correctly classified and the total number of pathological voices. False negative rate (FN) is the ratio between pathological files wrongly classified and the total number of pathological files. True negative rate (TN), sometimes called specificity, is the ratio between normal files correctly classified and the total number of normal files. False positive rate (FP) is the ratio between normal files wrongly classified and the total number of normal files.

The final accuracy of the system is the ratio between all the hits obtained by the system and the total number of files.

III.EXPERIMENTALPROTOCOLS:

As mentioned above the sample voice (normal and spasmodic) is divided in two set one for the training and one for test.Some experiments are realizedin order to evaluate the effect of different factors in our system; two groups of experiments are compared. One is based on the GMM classifier, whereas the other is using SVM classifier.

GMM classifier

-Use 12 MFCCs, energy, their derivate (Δ)and(ΔΔ).

-Use of different number of Gaussian (power of 2).

SVM classifier

-Use 12 MFCCs, energy, their derivate (Δ)and(ΔΔ).

-Use different kernel function.

In this study, the K-mean algorithm for clustering is used before training SVM so we will not separate features but we will separate their centers or cluster in order to assure convergence of SVM training and to reduce the cost of computation.

IV.RESULT AND DISCUSSION

In our experiment we need to know the optimal model which give best classification rate, this is obtained by a model with proprieties: 64 centers (Gaussian) for GMM and with a SVM with a polynomial kernel function with 39 MFCCs. The results are represented in confusion matrix in table 3.

Table3 Confusion matrix.

System's decision	Actual diagnosis			
	GMM		*SVM*	
	Pathological	**Normal**	**Pathological**	**Normal**
Pathological	79.92%	18.10 %	93 %	10 %
Normal	20.08%	81.90%	7 %	90%

The results of the classification given in a frame that means the rate of classification represent the number of known frame by the total of the frame.

If we test each file (normal and pathological) separately, we get an accuracy of **100%** for the two classes, by setting up a threshold to the number of classified frames. If more than **70%** of the frames of a file are assigned to a certain class, then the whole file is assumed to belong to that class.

•*Discussion:*

In this subsection, we discuss some experimental results obtained from the proposed analysis methods.

GMM

-The classification rate depend to the number of Gaussian and the number of parameters MFCCs as mentioned in "fig6"

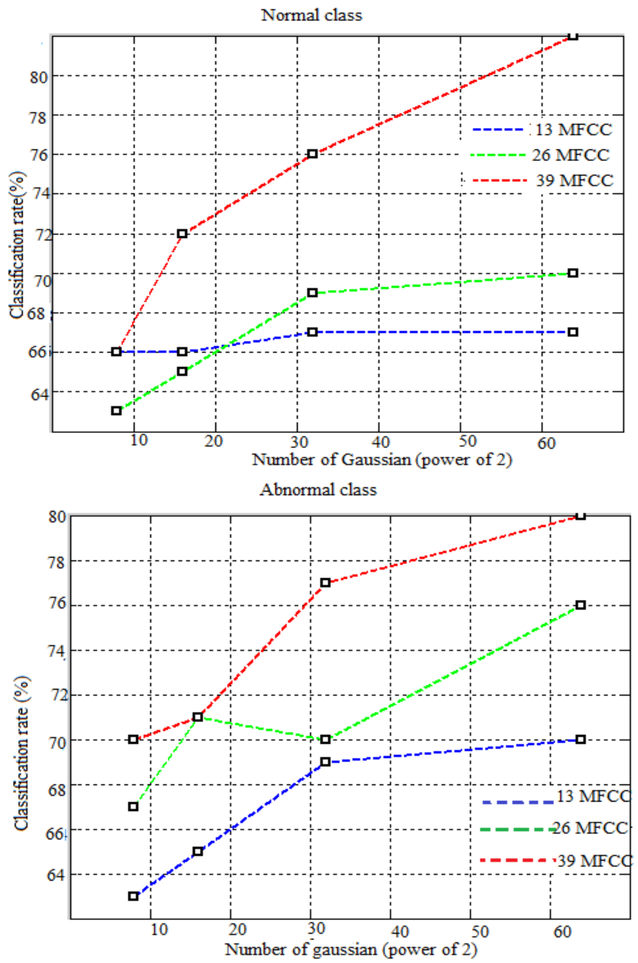

Figure 6. Classification rate for different mixtures and parameters for normal and abnormal class.

-From the two curve we note that when we increase the number of Gaussian with the increase of the MFCCs coefficients the classification rateimproves

-Modeling by GMM requires a large number of data for the training, particularly when we use a high number of Gaussian to create a model, this prevents us to use more than 64 Gaussians particularly with the abnormal class which contains a small number of file.

SVM

As with GMM, classification rate improve with the increase of MFCC coefficient the results are represented in figure 7

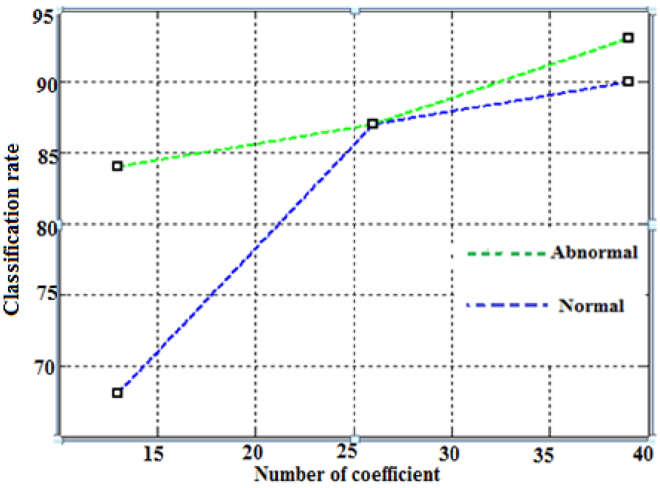

Figure7 Classification with SVM

-The precedent figure shows that SVM is more preferment than GMM.

V.CONCLUSION

This work is focused on pathological voices detection (spasmodic dysphonia) and it is built around a system for automatic speaker recognition based onGMM and SVM as classifier.

A good classification rate needs efficient features to characterize each class, in this work, on one handthe accuracy of system increase with the of the number of parameters (best accuracy with 39 coefficients) that means that the difference between normal and abnormal become noticeable with second derivate (ΔΔ) of MFCC and energy more than the others, on the other handthe effect of the number of Gaussian which makes up the model is important where a sufficient number of mixtures allows to represent data (features) optimally.We can deduce also that the quantity of data used for training a system is very important.Both GMM and SVM the best accuracy is obtained with (ΔΔ) dynamics parameters while SVM is more preferment than GMM where the accuracy for an abnormal class is 93% and 87% for the normal class.

The very promising result motivates us to improve this work;the future work will be concerned on the use of another database to assess the independence of the method used for the database. We will also validate this work with other pathologies for example organic pathologies.We will interest to the hybrid approach.

REFERENCES

[1] Ghio A. Dufour S. Rouaze M. Bokanowski V. Pouchoulin G. Révis J. Giovanni A. ‘’ Mise au point et évaluation d’un protocole d’apprentissage de jugement perceptif de la sévérité de dysphonies sur de la parole naturelle’’. REV LARYNGOL OTOL RHINOL.2011;132,1:1-9.

[2] Antoine Giovanni1, Pirng Yu2, Joana Révis1, Marie-Dominique Guarella1, Bernard Teston3, Maurice Ouaknine1 ’’Analyse objective des dysphonies avec l’appareillage EVA’’. Fr ORL - 2006 ; 90 : 183

[3] Darcio G. Silva, Luıs C. Oliveira and Mario Andrea ‘’Jitter Estimation Algorithms for Detection of Pathological Voices’’ Hindawi Publishing Corporation, EURASIP Journal on Advances in Signal Processing Volume 2009, Article ID 567875, 9 pages.

[4] Miltiadis Vasilakis, Yannis Stylianou ‘’Voice Pathology Detection Basedeon Short-Term Jitter Estimations in Running Speech’’ Folia Phoniatr Logop 2009;61:153–170.

[5] Sonu, R. K. Sharma ‘’ Disease Detection Using Analysis of Voice Parameters’’ International Journal of Computing Science and Communication Technologies, VOL.4 NO. 2, January 2012.

[6] Jacques Koremana, Manfred Pützer, Manfred Just ‘’Correlates of Varying Vocal Fold Adduction Deficiencies in Perception and Production: Methodological and Practical Considerations ‘’ F olia Phoniatr Logop 2004;56:305–320

[7] Miltiadis Vasilakis, Yannis Stylianou ‘’Voice Pathology Detection Based eon Short-Term Jitter Estimations in Running Speech’’ Folia Phoniatr Logop 2009;61:153–170.

[8] Raissa Tavares , Nathália Monteiro , Suzete Correia , Silvana C. Costa , Benedito G. Aguiar Neto (2) and Joseana Macêdo Fechine ‘’Optimizing laryngeal pathology detection by using combined cepstral features’’ Proceedings of 20th International Congress on Acoustics, ICA 2010 23-27 August 2010, Sydney, Australia ICA 2010.

[9] Julián D. Arias-Londoño, Juan I. Godino-Llorente, Germán Castellanos-Domínguez, Nicolás Sáenz-Lechón, Víctor Osma-Ruiz ‘’Complexity Analysis of Pathological Voices by means Markov Entropy measurements’’31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, September 2-6, 2009

[10] Juan Ignacio Godino-Llorente, Pedro Gómez-Vilda,Nicolás Sáenz-Lechón1, Manuel Blanco-Velasco, Fernando Cruz-Roldán, and Miguel Angel Ferrer-Ballester” Support Vector Machines Applied to the Detection of Voice Disorders” Springer-Verlag Berlin Heidelberg pp. 219 – 230, 2005.

[11] Nafise ErfanianSaeedi, FarshadAlmasganj, FarhadTorabinejad ‘’Support vector wavelet adaptation for pathological voice assessment’’ Computers inBiologyandMedicine41(2011)822–828

[12] Juan Ignacio Godino-Llorente*, Member, IEEE, Pedro Gómez-Vilda, Member, IEEE, andManuel Blanco-Velasco, Member, IEEE “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, OCTOBER 2006

[13] Alireza A. Dibazar, Theodore W. Berger, and Shrikanth S. Narayanan” Pathological Voice Assessment”IEEE EMBS 2006 NEW YORK.

[14] Nicolas Saenz-Lechon, Juan I. Godino-Llorente, Vıctor Osma-Ruiz, Pedro Gomez-Vilda ‘’Methodological issues in the development of automatic systems for voice pathology detection’’ Biomedical Signal Processing and Control 1 (2006) 120–128.

[15] G. Pouchoulin, C. Fredouille1, J.-F. Bonastre, A. Ghio, M. Azzarello, A. Giovanni ‘’Modélisation Statistique et Informations Pertinentes pour la Caractérisation des Voix Dysphonies’’ Actes des XXVIes journ´ees d’´etudes sur la parole Dinard, juin 2006.

[16] Manfred Putzer & Jacques Koreman ‘’A german databse for a pattern for vacal fold vibration ‘’ Phonus 3, Institute of Phonetics, University of the Saarland, 1997, 143-153.

[17] Ayaz Keerio, Bhargav Kumar Mitra, Philip Birch, Rupert Young, and Chris Chatwin“On Preprocessing of Speech Signals”On Preprocessing of Speech Signals” World Academy of Science, Engineering and Technology 47 2008