VOICE DISORDER DETECTION BASED ON AUTOMATIC SPEAKER IDENTIFICATION TECHNIQUES

Mohamed FEZARI, Fethi AMARA
Badji Mokhtar Annaba University,BP:12, 23000 Algeria
Laboratory of Automatic and signal Annaba,Faculty of Engineering,
mohamed.fezari@uwe.ac.uk, Amarafethi13@gmail.com

Abstract

In this paper, we investigate the proprieties of automatic speaker identification (ASI) to develop a system for voice pathologies detection, where the models do not correspond to different speakers but it corresponds to classes of patients who share the same diagnostic. One 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. A supervised algorithm is used to accomplish this task, Mel frequency cepstral coefficients (MFCCs) with first and second derivations are proposed as features, and modeled by weighted Gaussian mixture model (GMM) as it is used in ASI. The work is simulated using MATLAB, thanks to the toolbox voicebox for features extraction and dcpr for training and testing steps. The results are encouraging for further investigation on better classifiers.

Keywords- Voice disorders, Speech patholodies detection, classifiction techniques, machine learning, laryngeal deseases.

1. INTRODUCTION

it is an exciting time for clinicians to tune in to speech science, because a number of recent technological breakthroughs are having a great impact on the field. 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] [17], the first one is subjective because of the variability between listeners, although the second is objective it is invasive for one hand , on the other hand it has a limited reliability. It is well known that vocal fold pathologies alter the mechanisms of speech production; such disturbance is reflected in voice quality deterioration. Human speech production under both healthy and vocal fold pathology conditions suggests that alternative production models other than traditional may be more accurate [18].

Some related works are described here, Marek Wisniewski and all. In 2010 [19] developed a work on improving approach to automatic detection of speech disorders based on the HMMM technique, they apply it in Polish language. It is worth emphasizing that this method enables detection of a category of speech disturbance ie: fricative, nasal, vowels, etc… prolongation, but also provides the information about specific phoneme being disturbed.

In [20] Lotfi Salhi et All. Presented a new method for voice disorders classification based on multilayer network. The processing algorithm is based on hybrid technique witch uses wavelets energy coefficients as input of the multilayer neural network. The training step uses a speech database of several pathological and normal voices collected from the national hospital of “Tunis” and was conducted in a supervised mode for discrimination of normal and pathology voices. However, the database used in the tests was very short and the tests were used off line, thus the results (100%) of classification do not reflect the reality if the classifier is used on large database and in real-time.

Dean R. Hess [21], studies the effect of tracheotomy tube on voice production, the tube decreases the ability of the patient to communicate effectively, in mechanically ventilated patients, speech can be provided by the use of a talking tracheotomy tube, using a cuff-down technique with a speaking valve and using cuff-down technique without valve. They concluded that team work between the patient and the patient care team can result in effective restoration of speech in many patients with long-term tracheotomy.

This is why the development of automatic system for classification is proposed; in voice processing we distinguish three principal approaches: acoustic, parametric and non-parametric approach and statistical methods. The first approach consist to compare acoustics 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 for features selection [8-9].

The classification of voice pathology can be seen as pattern recognition so statistical methods are an important approach. We try to mimic the brain comportment where we can recognize persons from their voice. Many researches are realized for this task, Support vector machine (SVM) is applied to test the effectiveness and reliability of the short term cepstral and noise parameters [10], the same features are used with Hidden Markov Model (HMM) [11]. In [12] the MFCCs are proposed to be the input of multi-layer perceptron (MLP)

In this paper, the conception of our detector is inspired from a system of ASI [13]. 12 MFCCs, energy, dynamic parameters (first derivate and second derivate) are extracted to be the input of GMM. This classifier is trained with the algorithm of expectation maximization (EM) to get maximum likelihood (ML). The clustering algorithm K-mean is used for the initialization. Figure 1.a illustrates the main parts of the voice production system that can be affected by a pathology: throat, tongue, mouth and nasal cavity.

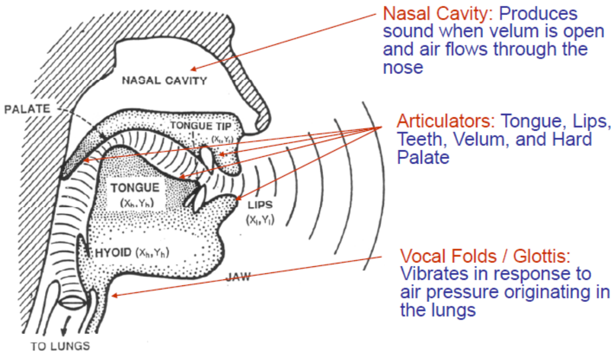
Nasal Cavity: Produces sound when velum is open and air flows through the nose
NASAL CAVITY
PALATE
TONGUE TIP
(Xt,Yt)
Articulators: Tongue, Lips, Teeth, Velum, and Hard Palate
TONGUE
(Xh,Yh)
LIPS
(Xl,Yl)
HYOID (Xh,Yh)
JAW
Vocal Folds / Glottis: Vibrates in response to air pressure originating in the lungs
TO LUNGS

Fig 1.a Voice production system

The number of Gaussians those make up the model is chosen as power of 2 in order to test its influence on the classification rate.

- *Description of ASR system:*

We present in figure 1.b the principals step to develop a system for speaker recognition:

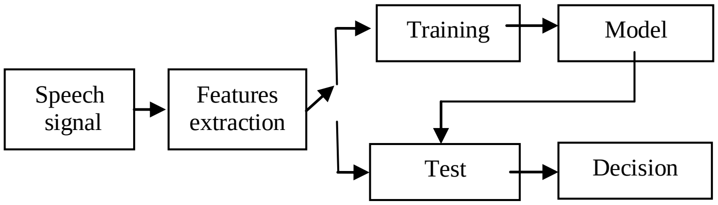

Fig 1.b: Block diagram for speaker recognition [14]

The difference between a system for ASR and a system for voce pathology detection is in two essential key points:

- In ASR the model corresponds to a speaker while the model in second system corresponds to 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 system, the experiments are in section 3. The results are presented in section 4 and the last section is reserved for the conclusion and future work.

2. 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 “fig2” show different steps adapted to our system.

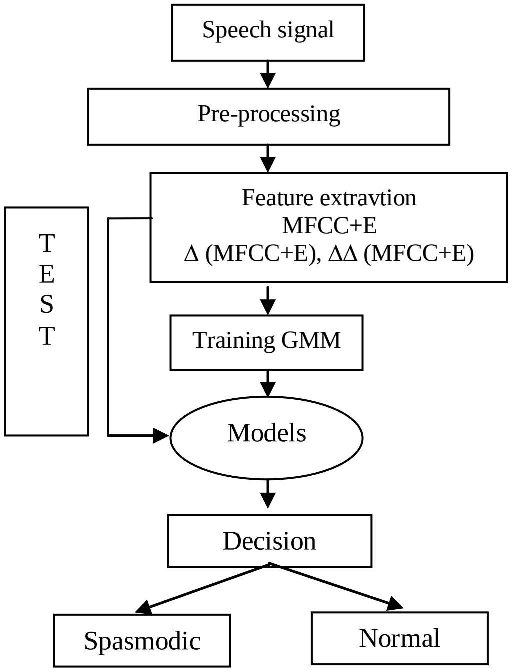

Figure 2 block diagram adapted to the detector.

2.1 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. [12]

In this work we have choose a German database for voice disorder developed by PUTZER [15] which contain healthy and pathological voice, where each one pronounce vowels [i, a, u] /1-2 s in wav format at different pitch (low, normal, high) it contain also phrase and electroglottograph signal (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 particularly 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.

2.2Pre- 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 study we are interesting to remove silence knowing that the efficient features are included in speech portion [16].

2.3 Features extraction:

Features extraction means finding good data that hepls 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. Mel frequency cepstral coefficients are given by:

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

These parameters are extracted by 32 filter bank applied on 10 ms (256 points) Hamming windowed frames at 50% of overlap.

The MFCCs are calculated thanks to the toolbox voice box with *melcepst* function.

D. training

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

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

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

λ is the model.

Each component has the general form:

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

Σ is the d-by–d covariance matrix and |∑| is its determinant it characterizes 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 the non-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 of the univariate probability densities for the elements of x.

In order to get optimal model the GMMs one way to get this is the use of Maximum likelihood estimation (MLE) In practice it is often more convenient to work with the logarithm of the likelihood function, called the log-likelihood and the MLE are given in the following :

$$LE = \ln(P(\lambda / x_1, x_2, .., x_n)) = \sum_{i=1}^{n} \ln(P(x_i / \lambda)$$

$$A_v = \frac{1}{n} \ln(LE) \qquad (5)$$

$$LME\theta = aeg \max(A_v)$$

Maximizing the likelihood of observing x as 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 the algorithm of clustering K-means.

E. Test steps:

Once models are created and that we have managed to train the GMM, we can proceed to the classification test. A new feature vector Xt is said to belong to an appropriate model if it maximizes p (Xt | λ) for every possible class.

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

Table 2 typical aspect of a confusion matrix

System's decision	**Actual diagnosis**	
	Pathological	**Normal**
Pathological	True positif (TP)	False positive (FP)
Normal	False negative (FN)	True negative (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.

3. EXPERIMENTAL PROTOCOLS:

As mentioned above the sample voice (normal and spasmodic) is divided in two set one for the training and one for test so we will create two model.

Some experiments are realized in order to evaluate the effect of different factors in our system, these experiments are described briefly:

- Change the length of segment.
- Use Mel frequency cepstral coefficients MFCCs, their first and second derivate
- Use of different number of Gaussian (power of 2).
- Change number of iteration for the EM algorithm.

4. 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: hamming window of 256 points, 64 centers (Gaussian), 39 MFCCs and 1000 iterations.

The results are represented in confusion matrix in table 3.

Table 3 confusion matrix

System's decision	Actual diagnosis	
	Pathological	**Normal**
Pathological	79.92%	18.10 %
Normal	20.08%	81.90%

This recognition rate presents the percentage of the recognized frames among the total number of frames of the test set witch contain all a files of the class and then averaged.

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. This result is based on off-line tests. 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.

-The classification rate depend to the number of Gaussian and the number of parameters MFCCs as mentioned in "figures 3"

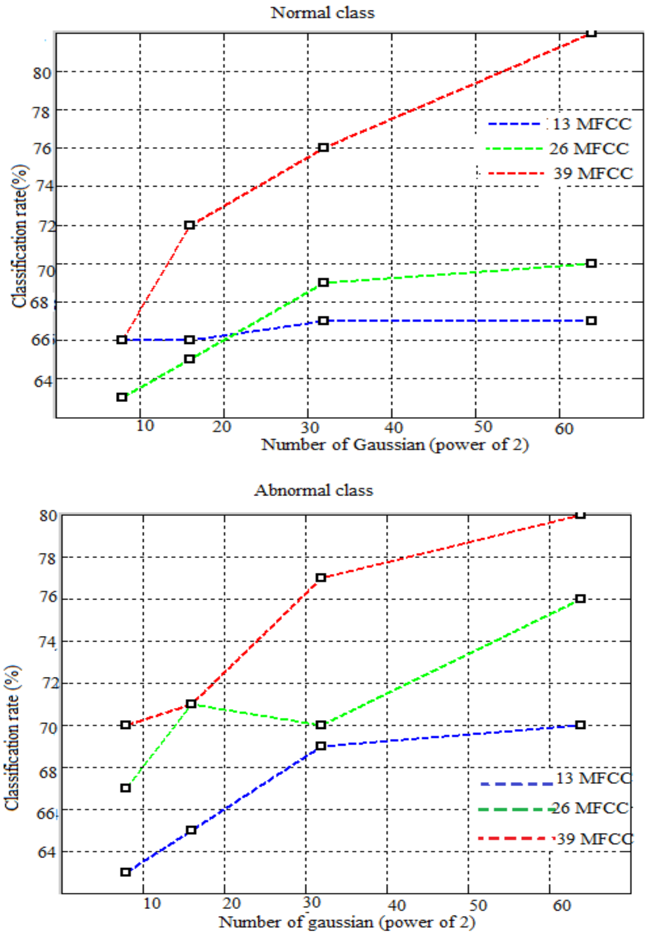

Figure3.Classification rate for different mixtures and parameters for normal and abnormal class.

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

-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 Gaussian particularly with the abnormal class which contains a small number of file.

5. CONCLUSION

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

A good classification rate needs efficient features to characterize each class, in this work, on one hand the accuracy of system increase with the number of parameters (best accuracy with 39 coefficients) that means that the difference between normal and abnormal become noticeable with second derivate of MFCC and energy more than the others, on the other hand the 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.

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, it must give a similar results. We will also validate this work with other pathologies for example organic pathologies. In order to improve the obtained classification rate we will be interested in improving the classification phase by combining a hybrid system GMM-SVM, and defining more classes such as throat, nasal cavity of mouth diesis.

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 '' Folia 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] Maria Markaki and Yannis Stylianou ''Using Modulation Spectra for Voice Pathology detection and Classification''..

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

[12] 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.

[13] 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.

[14] Charles pelltier "classification des son respiratoires en vue d'une detection automatique des sibllants" these november 2006.

[15] 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.

[16] 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.

[17] Sanchez I, Avital A, Wong I, Tal A, Pasterkamp H. “Acoustic vs. spirometric assessment of bronchial responsiveness to methacholine
in children” . Pediatr Pulmonol 1993;15(1):28-35.

[18] Hansen J. H.L. et All., « A non-linear Operator based speechFeature Analysis Method with Application to Vocal fold pathology Assessment », IEEE Transaction Biomedical Eng. 29 pgs. 1995.

[19] Marek Wisniewski, W. Kuniszyk, E. Smolka , W Suszynski,” Improved approach to automatic detection of speech disorders basedon the Hidden Markov Models Approach”, in Journal of med. Informatics & technologies, Vol. 15/2010 ISSN 1642-6037.

[20] Lotfi Salhi, Talbi Mourad and Adnene Cherif, “ Voice Disorders Identification using Multilayer NeuralNetwork”, in the International Arab journal of Information Technology, Vol. 7, no 2, april 2010.

[21] Dean R. Hess, “ Facilitating Speech in the patient With a Tracheostomy”, in Journal of Respiration care , April 2005, Vol. 50 No.4, pp. 519-525.