Acoustic Analysis for Detection of Voice Disorders Using Adaptive Features and Classifiers

Mohamed FEZARI, Fethi AMARA and Ibrahim M. M. EL-EMARY

Abstract—Voice diseases are increasing dramatically, due mainly to unhealthy social habits and voice abuse. In this paper, we investigate the methods of acoustic voice analysis (AVA) with adaptive features to develop a system for voice pathologies detection, where the models correspond 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, non-neurological pathologies (such as chronic laryngitis and Vocal fold nodules), is proposed for this study. A supervised algorithm is used to accomplish this task, Mel frequency cepstral coefficients (MFCCs with variation of Jitter & shimmer), and modeled by weighted Gaussian mixture model (GMM) as it is used in AVA. The work is simulated using MATLAB, for features extraction, for training and testing steps. The results are encouraging for further improvement of combining classifiers and investigation multi-pathologies classifier.

Keywords— Voice disorders, Acoustic voice analysis, classification techniques, Jitter and Shimmer, laryngeal deseases.

I. INTRODUCTION

Acoustic analysis may provide a useful means to quantitatively characterize the tremulous voice. Assessment voice quality is an important tool for dysphonia evaluation, vocal fold polyp, voice Trimor and other voice pathologies; 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].

Pitch and Formants Analysis gives many information concerning speaker identification, speaker emotions and voice disorders of speaker. Figure 1.a illustrates the pitch variation by application of the cestrum method analysis of normal and pathological female sounds (32 years).

The high distortion and the variation of the pitch around the expected value (250 Hz) demonstrate a state of the glottic signal anomaly, resulting of a laryngeal pathology.

Fig 1.a: Sound of vowel /a/ normal voice and creaky voice Wav, spectrogram and pitch form [23].

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.

N. Saenz-Lechon et All [9], studied the effect of audio compression in automatic detection of voice disorders, they investigated the detection of pathologies in voice when the voice samples have been compressed in MP3 format and different binary rates (160, 96, 64, 48, 24 and 8 kb/s). their detector employs cepstral and noise measurements with their derivatives, to characterize the voice signals. The classification is performed using GMM and SVM. The results between the different proposed detectors are compared by means of detector
error tradeoff (DET), they concluded that there is no significant
differences in the performance of the detector when the binary
rates of the compressed data are above 64 kb/s.

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.

Saenz-Lechon et All in [10], presented an overview of
previous classification schemes applied to voice disorders on
Massachusetts Eye & ear infirmary (MEEI) Database [11b],
they described some methodological paradigms to be
considered when designing an automatic pathological voice
detection system. They insisted on the use of a commercially
well-known databases, a cross-validation strategy based on
several partitions to obtain averaged classification performances with confidence intervals, a report of the means
of a detection error trade-off (DET), and an investigation of the
area under receiver operating characteristic (ROC) curves.

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
tracheostomy 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 tracheostomy.

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 the fundamental frequency can be
calculated by methods presented in [7].

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.

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 based on
AVA and inspired from a system of Automatic speech
recognition [22]. 12 MFCCs, energy, dynamic parameters (first
derivative and second derivate) with Jitters and Shimmers
parameters 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.

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.

• Algorithm :

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

![Fig 1.b Voice production system](image)

**Figure 1.b Voice production system**

![Fig 1.c: Block diagram for speaker recognition](image)

**Figure 1.c: Block diagram for speaker recognition [14]**

The difference between a system for ASR and a system for
voice 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 sets are used in test.

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.
II. METHODOLOGIE

Our system will pass by the same steps to concept a system for ASR, we will describe them step by step, the block diagram in “fig2” show different steps adapted to our system.

![Block diagram for the voice disorder detector.](image)

**A. Speech signal:**

In this work the creation of the database 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 in [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**

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>Age</td>
<td>Number</td>
</tr>
<tr>
<td>Normal</td>
<td>52</td>
<td>20-60</td>
</tr>
<tr>
<td>Pathological</td>
<td>29</td>
<td>30-82</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>29-82</td>
</tr>
</tbody>
</table>

Those files are down sampled to 25 KHz in order to get optimal analysis.

**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 study we are interesting to remove silence knowing that the efficient features are included in speech portion, then we selected only vowels /a/ and /u/ for the training and tests. [16].

**C. Features extraction:**

Features extraction means finding good parameters that helps to classify between the healthy and abnormal patients, 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. More over we added the as features variations of Jitter and Shimmer (which represent variation in frequencies, and in amplitudes of pitch) as formulated in (2) and (3).

**C.1 MFCC features**

Mel frequency cepstral coefficients are given by:

$$C = \sum_{k=1}^{K} \log (S(k)) \cos (\frac{\pi}{2} - \frac{\pi}{2}) \frac{n}{K}$$

(1)

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

**C.2 Jitter & shimmer Features**

Jitter may occur during voice production, especially in vowel phonation, and it is defined as small fluctuations in glottal cycle lengths [3,4] and [7]. Jitter and shimmer
(amplitude perturbations) over successive speech cycles help give the vowel its naturalness in contrast to constant pitch and amplitude that can result in a machinelike sound. Moreover, jitter (and shimmer) contributes to the voice quality of a speaker. In terms of signal processing, jitter is a form of modulation noise. Specifically, jitter is a modulation of the periodicity of the voice signal. A high degree of jitter results in a voice with roughness that is usually perceived in recordings of pathological voices.

Therefore, a reliable estimation of jitter can be used to discriminate between healthy and dysphonic speakers.

Which are defined as:
Jitter : % change in cycle duration between cycles
Shimmer : % change in speech amplitude between cycles.

The equations (2) and (3) yield to determine the percentage of Jitter and Shimmer in speech signal.

\[
\text{Jitter} = \frac{1}{N-1} \sum_{i=1}^{N} |T_i - T_{i+1}| \quad (2)
\]

\[
\text{Shimmer} = \frac{1}{N-1} \sum_{i=1}^{N} |A_i - A_{i+1}| \quad (3)
\]

\(T_i\), time and \(A_i\): Amplitude, \(N\): number of cycles.

D. training using GMM

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:

\[
P(x|\lambda) = \sum_{j=1}^{M} p(x_i | \lambda_j) W_j \quad (4)
\]

\[\sum_{j=1}^{M} W_j = 1 \quad (5)\]

\(\lambda\) is the model.

Each component has the general form:

\[
p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (6)
\]

\(\Sigma\) is the d-by-d covariance matrix and \(|\Sigma|\) is its determinant it characterizes the dispersion of the data on the d-dimensions of the feature vector. The diagonal element \(\sigma_{ii}\) 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, \(\Sigma\) 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) given by:

\[
p(X|\lambda) = \prod_{i=1}^{M} p(x_i | \lambda) \quad (5)
\]

\[X = (x_1, x_2, ..., x_M) \quad (6)\]

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 step:

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

A new feature vector \(X_t\) is said to belong to an appropriate model if it maximizes \(p(X_t|\lambda)\) 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>
<thead>
<tr>
<th>System’s decision</th>
<th>Actual diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathological</td>
<td>Pathological</td>
</tr>
<tr>
<td>Normal</td>
<td>Normal</td>
</tr>
</tbody>
</table>

True positive (TP) or sensitivity, is the ratio between pathalogical files correctly classified and the total number of pathological voices. False negative rate (FN) is the ratio between pathalogical files wrongly classified and the total number of pathalogical 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. EXPERIMENTAL PROTOCOLS

A. Database details

The used database [16] was collected in a collaboration project of the department of phonetics and ENT at the Caritas clinic St. Theresia in Saarbrucken and the Institute of Phonetics of university of Saarland in Germany.
The collection of the database has combined research methodologies from speech science with phonetic methods. Methods from speech research which were used are Electro-glottography (EGG) and recording of the sound pressure waveform (microphone signal). Both signals were recorded onto DAT tape simultaneously in a quiet room. The signals were recorded for a read text and for the vowels /i:/, /a:/ and /u:/ at normal, high and low pitch.

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.

B. Test Scenarios

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 plus Energy.
- Include Jitter and Shimmer as Features.
- Use of different number of Gaussian (power of 2).
- Change number of iteration for the EM algorithm.

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

The results are represented in confusion matrix in table 3.

<table>
<thead>
<tr>
<th>System’s decision</th>
<th>Actual diagnosis (MFCCs and Energy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pathological</td>
</tr>
<tr>
<td>Pathological</td>
<td>79.92%</td>
</tr>
<tr>
<td>Normal</td>
<td>20.08%</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
</tr>
<tr>
<td>Pathological</td>
<td>18.10%</td>
</tr>
<tr>
<td>Normal</td>
<td>81.90%</td>
</tr>
</tbody>
</table>

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.

A. 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”

Table 3. Confusion matrix with MFCC and Energy coefficients

<table>
<thead>
<tr>
<th>System’s decision</th>
<th>Actual diagnosis (MFCCs and Energy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pathological</td>
</tr>
<tr>
<td>Pathological</td>
<td>79.92%</td>
</tr>
<tr>
<td>Normal</td>
<td>20.08%</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
</tr>
<tr>
<td>Pathological</td>
<td>18.10%</td>
</tr>
<tr>
<td>Normal</td>
<td>81.90%</td>
</tr>
</tbody>
</table>

Fig3.a Classification rate for different mixtures and parameters for normal class.

<table>
<thead>
<tr>
<th>System’s decision</th>
<th>Actual diagnosis (MFCCs + Jitter &amp; Shimmer)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pathological</td>
</tr>
<tr>
<td>Pathological</td>
<td>82.14%</td>
</tr>
<tr>
<td>Normal</td>
<td>17.86%</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
</tr>
<tr>
<td>Pathological</td>
<td>17.4 %</td>
</tr>
<tr>
<td>Normal</td>
<td>82.6%</td>
</tr>
</tbody>
</table>

Fig3.b. Classification rate for different mixtures and parameters for 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.
-Including the Jitters and Shimmers as parameters in features then applying GMM has increased so how the rate of classification in both normal and pathological sets as shown in table 4.

V. CONCLUSION

This work is focused on pathological voices detection (spasmodic dysphonia) and it is built around a system for acoustic voice analysis, with automatic speaker recognition techniques, based on MFCC and variation in frequencies and amplitudes as features (Jitter & Shimmer) and GMM with multiple numbers of Gaussians as classifier.

A good classification rate needs efficient features to characterize each class, in this work, on one hand the accuracy of system increases with the number of parameters (best accuracy with 39 coefficients including Jitter & Shimmer) 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 directed to the use of another database to assess the independence of the method used for the database, it must give a similar or better 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. Features selection will be investigated more to get correlated features with specific pathologies and defining more classes such as throat, nasal cavity of mouth pathologies.

REFERENCES

[2] Antoine Giovanni1, Pirmg Yu2, Joana Revis1, Marie-Dominique Gurel1, Bernard Teston3, Maurice Ouaknine1 "Analyse objective des dysphonies avec l’appareillage EVA". Fr ORL - 2006 ; 90 : 183