Experimental Speech recognition from pathological voices

HADJI SALAH
Laboratory of nano-materials (LANSER) of Energy Center (CRTEn)
Technopole of Borj-cedria, Hamm-lif, Tunis, 2050 TUNISIA
msahli.salah@yahoo.fr

***Abstract* :** Speech recognition has been the subject of quite a few research subjects as it is the adequate means for dynamic, efficient and interaction Human communication simultaneously using the two phenomena of phonation and hearing between speakers, the applications of its searches are enormous for example one can quote: the dictation, the speech synthesis within the Windows software, the speech recognition of the Google search engine at the Smartphone level …… ..etc. all its applications depend on the conditions of use in which they are implemented, to be done and to overcome the puzzles of imperfection it is necessary to be sure to properly characterize the speech signal by extracting the most relevant characters such as: the fundamental frequency (pitch in English), timbre, tonality, to extract them many techniques are possible, the most used of which are acoustic such as: MFCC, PLP, LPC, RASTA and other in the form of combination (hybridization) namely: PLP RASTA, MFCC PLP …… etc. They are used in data transmission, speaker recognition and even in speech synthesis.

Keywords : speech signal, Parametrisation, SVM, pathological voices, classifier, MFCC, PLP, RASTA, LPC
Received: June 17, 2021. Revised: March 19, 2022. Accepted: June 13, 2022. Published: July 29, 2022.

1. Introduction

The extraction of acoustic parameters or characteristics, such as fundamental frequency, formants, etc. is done by applying signal processing methods which are for example: time-frequency analysis, spectral analysis, Cepstral analysis..etc Parameterization constitutes the initial block (fig.2) for any recognition of a speech signal, its role is to extract from a speech signal the most relevant information possible in order to be able to make a separation between the sounds [8]. The extracted information is presented as a sequence of acoustic vectors. In order to be able to extract these parameters, several methods exist, taking into account the superposition of the noises of the sounds, we will make a comparison of the different methods (MFCC, PLP, PLP RASTA, and the combination of several other parameters such as LPC, pitch, forming, energy). Given the redundancy of the speech signal and its complexity, to process it, different methods are admitted to have a better parameterization. In this paper we will give a brief overview on the signal processing tools such as short-term energy and weighting windows,then see the different speech signal parameterization methods which are: LPC (Linear Predictive Coding) analysis, Homomorphic or Cepstral analysis on which the MFCC (Mel Frequency Cepstral Coefficient) is based, PLP (Predictive Linear Perceptual) and PLP-RASTA (Real Ative SpecTrA).this involves using an SVM classification to distinguish between speech signals from people with speech pathology (Nodule or Oedeme) and normal signals (no pathology). In this paper, two types of classifications have been used:

- A classification in two classes: in which we used samples of corpus from pathological signals (Nodule and Oedema) and another from normal signals to make this classification, (see Fig.) Shows us the principle of classification.

- A multi-class classification: in which we took samples of each type of pathology among the two that we have (Nodule and Oedme) to constitute the first and second classes and samples from normal signals, which is for the third class. To perform a multi-class classification, we used a One VS all type of algorithm, that is to say "one against all", this algorithm consists of taking a signal and comparing it with all the classes.

2. Parameterization methods

There are several methods of parametrization; there are those, which based on the perception of the human ear like the MFCC, PLP and others, which are interested in the model of speech production such as the Cepstral method and the LPC.

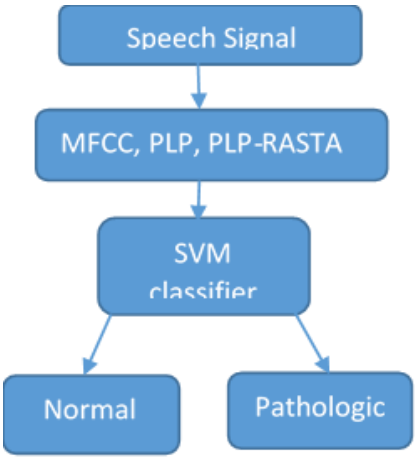


Fig.2. Classification algorithm

2.1 Cepsral Frequency Coefficients on the Mel scale (MFCC):

Obtaining the Cepstral Coefficients at the Mel scale was developed in 1980 by Davis and Mermelstein, to do this it is necessary to apply a Hamming window to each frame of the signal, we then obtain a Cepstral characteristic vector per frame, then we apply the Discrete Fourier Transform (DFT). let us then keep the Logarithm of the amplitude spectrum, then after smoothing the spectrum let us apply the Discrete Cosine transform to have the Cepstral coefficients (see figure).

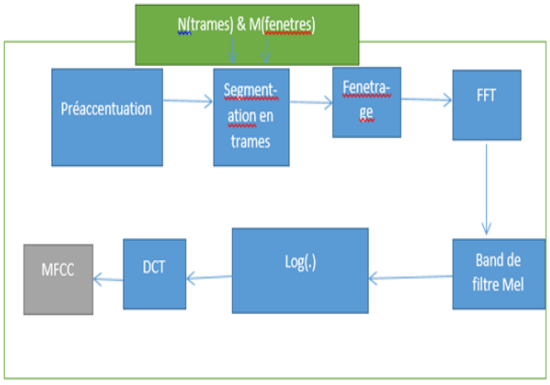


Fig .1. Mel coefficient calculation process

The extraction of the MFCC coefficients consists of six steps as mentioned in the previous figure (Fig. 1) [6]:

Step 1: Pre-emphasis:

This step in the process is to emphasize the high frequencies, this result in increasing the energy at the higher frequencies.

$$Y(n) = X[n] - a.\,X[n-1] \qquad (1)$$

Stage 2: Segmentation into frames: this stage consists in fragmenting the signal into frames of 20 to 40 ms. The speech signal is split into N samples. Adjacent samples are spaced by M (M<N), typically the values used are M=100 and N=256.

Step 3: Windowing with Hamming: Discontinuities related to segmentation can be overcome by multiplying each frame by a Hamming window. The Hamming window is given by the following equation:

If the window is defined as W (n), $0 \leq n \leq N-1$ such that:

N: number of samples in each frame

Y[n]: Output signal

X[n]: Input signal

W (n): The Hamming window, so the result will be:

$$Y(n) = X(n).Y(n) \qquad (2)$$

W (n) =0.54-0.46 cos (2πn/ (N-1)) (3)

0≤n≤N-1

Step 4: The fast or short-term Fourier transform:

To go from the time domain to the spectral domain, a Fourier transform is applied to each frame of N samples. The FFT is shown at the bottom:

Y(w)=FFT[h(t)*x(t)] =H(w)×X(w) (4)

Step 5: Mel Filter Bank

Step 6: Application of the iDCT (Inverse Discrete Cosine Transform)

2.2 LPC (Linear Predictive Coding)

LPC analysis is based on the speech production model mentioned in the figure below [7]. Starting from the hypothesis modeling the speech by a linear process, then It is a linear prediction at an instant n of the p previous samples. However, the non-linearity of speech requires the existence of an error denoted e(n) introduced to correct this error [2].

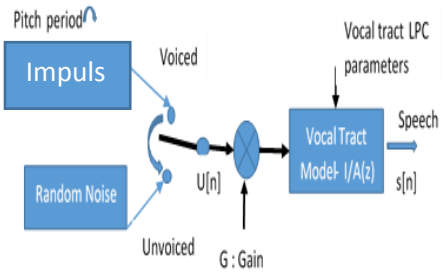


Fig.3. Speech production model

The LPC consists in calculating the coefficients a_k by minimizing the error. The following equation presents the process:

$$s(n) = \sum_{k=1}^{p} a_k.s(n-k) + G_u(n) \qquad (5)$$

The preacher's equation is:

$$s'(n) = \sum_{k=1}^{p} a_k.s(n-k) \qquad (6)$$

The prediction error is calculated by the following equation:

$$e(n) = s(n) - s'(n) =$$

$$s(n) - \sum_{k=1}^{p} a_k.s(n-k) \quad . \qquad (7)$$

The problem that troubles researchers is: how to determine p "optimal" coefficients a_k knowing N samples of a certain signal x[n] such that the error e(n) is the smallest possible. To do this, we minimize the energy of the prediction error e (n), over the duration of the block of length N. So we need to minimize:

$$E = \sum_{n=0}^{N-1} e[n]^2 =$$

$$\sum_{n=0}^{N-1} \begin{pmatrix} x[n] \\ \sum_{k=1}^{p} a_k.x[n-k] \end{pmatrix}^2 \qquad (8)$$

We get there by setting $\partial E/\partial a_k = 0$ and for each a_k. This generates a system of p equations with p unknowns (the a_k), which can then be solved to obtain the a_k. The system of equations that will allow us to calculate the coefficients a_k is:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ R(2) & R(1) & R(0) & \dots & R(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \dots \\ R(p) \end{bmatrix}$$

Fig.4. Yule-Walker matrix

Such as:

$$R(k) = \sum_{m=0}^{n-1-k} s(m).s(m+k) \qquad (9)$$

The transfer function of the filter is determined by the following equation:

$$A(z) = \frac{w(z)}{s(z)} = \sum_{k=1}^{p} a_k \cdot z^{-k} \qquad (10)$$

2.3. ¢ he PLP technique

PLP (Perceptual Linear Prediction) is a parametrization technique based on the human auditory system, it is an improvement of the one named LPC which estimates the spectrum over the entire audible band and can miss certain spectral details. The PLP estimates the parameters of an all-pole autoregressive filter, allowing a better modeling of the auditory spectrum by introducing critical bands at the level of the power spectrum with a bank of 17 filters whose central frequencies are linearly spaced according to the Bark scale which simulates the perception of the human ear [3, 4], whose audible frequencies range approximately from 20 Hz to 22 kHz much closer to perception than the linear Hertz scale (1 Bark = 100 Mels) [4].

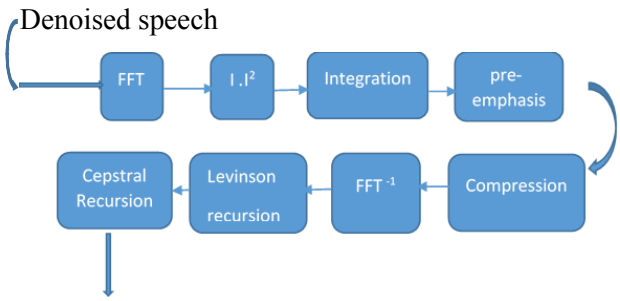


Fig.4. PLP coefficients

2.4. ¢ he PLP RASTA technique:

PLP RASTA is a hybrid parametrization technique between Perceptual Linear Prediction (PLP) and Relative Spectral Prediction (RASTA). The RASTA technique allows the identification of the (interesting) zones by comparing the temporal evolution of the spectral components with respect to the vocal tract and removes the others that do not correspond to them, which are not speech (noise), or the signal speech is often stained with noise having a slow variation, RASTA uses a bank of filters eliminating stationary signals, this technique makes it possible to reduce the sensitivity of speech analysis in the face of slow changes, a band-pass filter is applied to each spectral component according to a frequency representation in the critical band. The transfer function is:

$$H(z) = 0.1z^{-4} * \left(\frac{2.z^{-1} - z^{-3}2.z^{-4}}{1 - 0.98z^{-1}}\right) \qquad (5)$$

This method gives results against distortions and its lower quality for additive noises [8].

3. Experimental Results

Two essential steps to carry out the classification of the pathological paths and those healthy, to be done the first step is the parametrization (matrix of the relevant parameters), or still acoustic vectors extracted starting from corpus of sounds of the TIMIT database and from other people with vocal pathologies, these vectors are injected at the input of the SVM classifier, the first step is learning ", and the second step is the test, this is why the validation base is divided into two sub-bases one for learning (3/4) and one for testing (1/4). After a certain number of executions of the two stages, we can distinguish the voices of healthy people from the voices of people who have difficulties during the production of speech (cold for example). In the following, we present the different analyzes:

-The LPC analysis represents the speech signal by these LPC linear predictive coding coefficients and is carried out in 4 steps:

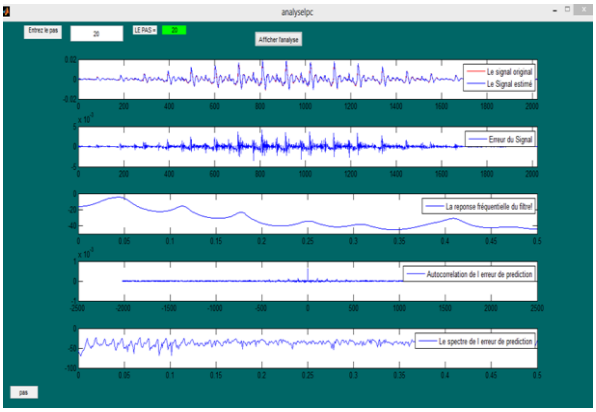


Fig.5. LPC analysis of a few samples of a signal

Then we present the broad and narrow band spectrogram

- The broadband spectrogram which is obtained with a window of short duration (3 ms in our project), it makes it possible to follow the evolution of the formants, the voiced periods appear there in the form of dark bands which are vertical.

- The narrow band spectrogram: it is obtained with a larger window (30 ms), it makes possible to visualize the harmonics of the signal in the voiced zones, and they appear in the form of horizontal bands

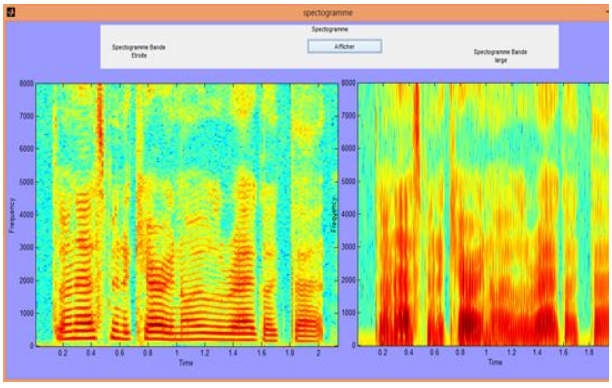


Fig.6. Representation of broadband (right) and narrowband spectrograms

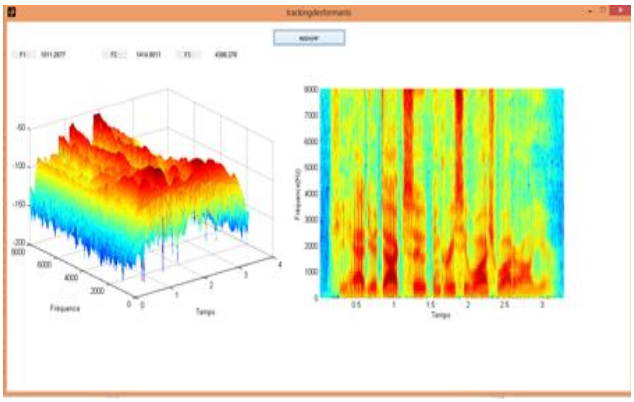


Fig.7. Formants display

- PLP and PLP-RASTA technique analysis

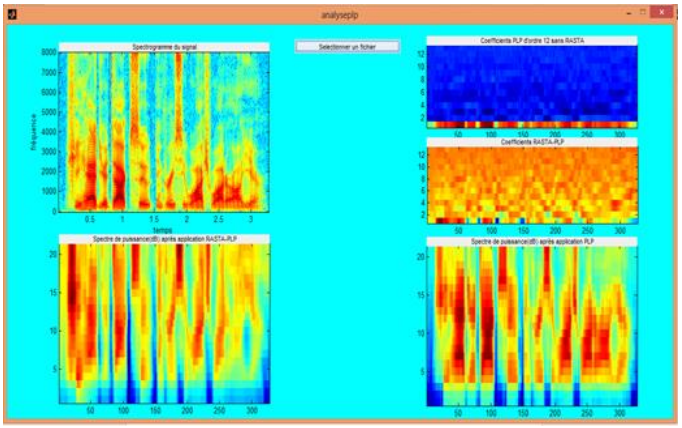


Fig.8. PLP and PLP-RASTA analysis

In the following we will present the parametrization matrices of the different techniques and establish a comparison

- Parametric matrix 1:

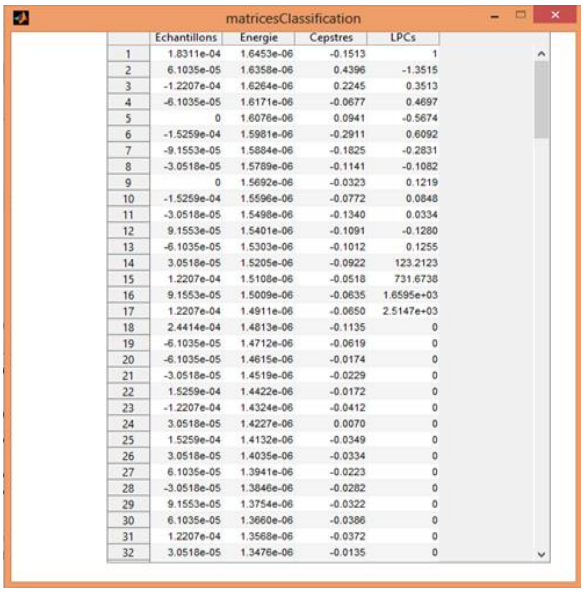
This matrix is the first that we will use in the classification in order to make a comparison between the performances of the different methods of parameterization.

This matrix contains 4 columns and 200 rows, the columns contain the parameter types and the rows contain the values:

- ✓ First column: this column contains samples of the signal
- ✓ Second column: this second column contains the short-term energy of the signal.
- ✓ Third column: contains the cepstral coefficients.

- ✓ Fourth: This column contains the first 12 cepstral coefficients plus the pitch (F0) and the first three formants (F1 F2 F3)

The following figure shows this matrix:



	Echantillons	Energie	Cepstres	LPCs
1	1.8311e-04	1.6453e-06	-0.1513	1
2	6.1035e-05	1.6358e-06	0.4396	-1.3515
3	-1.2207e-04	1.6264e-06	0.2245	0.3513
4	-6.1035e-05	1.6171e-06	-0.0677	0.4697
5	0	1.6076e-06	0.0941	-0.5674
6	-1.5259e-04	1.5981e-06	-0.2911	0.6092
7	-9.1553e-05	1.5884e-06	-0.1825	-0.2831
8	-3.0518e-05	1.5789e-06	-0.1141	-0.1082
9	0	1.5692e-06	-0.0323	0.1219
10	-1.5259e-04	1.5596e-06	-0.0772	0.0848
11	-3.0518e-05	1.5498e-06	-0.1340	0.0334
12	9.1553e-05	1.5401e-06	-0.1091	-0.1280
13	-6.1035e-05	1.5303e-06	-0.1012	0.1255
14	3.0518e-05	1.5205e-06	-0.0922	123.2123
15	1.2207e-04	1.5108e-06	-0.0518	731.6738
16	9.1553e-05	1.5009e-06	-0.0635	1.6595e+03
17	1.2207e-04	1.4911e-06	-0.0650	2.5147e+03
18	2.4414e-04	1.4813e-06	-0.1135	0
19	-6.1035e-05	1.4712e-06	-0.0619	0
20	-6.1035e-05	1.4615e-06	-0.0174	0
21	-3.0518e-05	1.4519e-06	-0.0229	0
22	1.5259e-04	1.4422e-06	-0.0172	0
23	-1.2207e-04	1.4324e-06	-0.0412	0
24	3.0518e-05	1.4227e-06	0.0070	0
25	1.5259e-04	1.4132e-06	-0.0349	0
26	3.0518e-05	1.4035e-06	-0.0334	0
27	6.1035e-05	1.3941e-06	-0.0223	0
28	-3.0518e-05	1.3846e-06	-0.0282	0
29	9.1553e-05	1.3754e-06	-0.0322	0
30	6.1035e-05	1.3660e-06	-0.0386	0
31	1.2207e-04	1.3568e-06	-0.0372	0
32	3.0518e-05	1.3476e-06	-0.0135	0

Fig.9. Parameterization matrix 1

- o Parametric matrices 2

This part groups together the 3 most used parameterization methods in all that is voice recognition. Each represented by a matrix.

The dimension of these matrices is 13 columns and 214 rows. That is 214 frames or vectors and each with 13 coefficients.

- ✓ First matrix: this matrix contains the PLP coefficients without the RASTA filtering.
- ✓ Second matrix: it contains the MFCC coefficients.
- ✓ Third matrix: This third and last contains the PLP-RASTA coefficients.

The Fig.9, presents these 3 matrices:

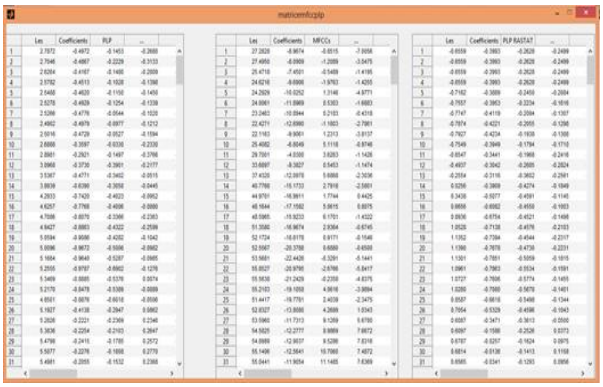


Fig.9. Classification of pathological signals VS Normals

It is a question of using an SVM classification in order to make a distinction between speech signals coming from people who suffer from vocal pathology (Nodule or Oedema) and normal signals (no pathology).

In our application, two types of classifications were used:

- A classification into two classes: in which we used corpus samples from pathological signals (Nodule and Oedema) and others from normal signals to make this classification, the figure (Fig.) shows us the principle of classification.

- A multi-class classification: in which we took samples from each type of pathology among the two we have (Nodule and Oedma) to constitute the first and second classes and samples from normal signals which is for the third class. To perform a multi-class classification, we used a OneVSall-type algorithm, i.e. "one against all", this algorithm consists of taking a signal and comparing it with all the classes.

1. Learning phase

The learning phase consists of creating a basic model on which the subsequent classification of signals is based.

This involves taking a speech signal, extracting its coefficients (MFCC, PLP or PLP-RASTA) and applying the function dedicated to learning

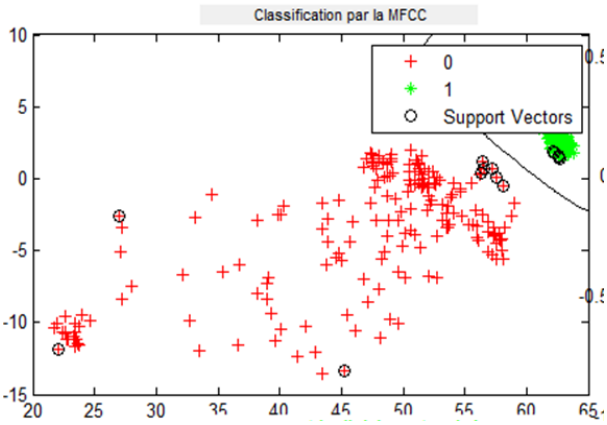


Fig. 10. Learning phase

The red color corresponds to the parameters extracted from a signal of a healthy individual (0) and the green color corresponds when it to the pathological signals (1)

2. Test phase

The test phase consists of recovering the matrix resulting from learning in order to predict or generate a decision, arguments of SVM classifier:

- Train matrix or learning matrix:

- Data N: is a matrix of the same size as the matrix used when learning the model, it is a matrix of data to be classified. Thus the system displays a message to say that the voice comes either from a healthy person in terms of voice production or suffers from a pathology.

The following figure shows an overview of a classified signal.

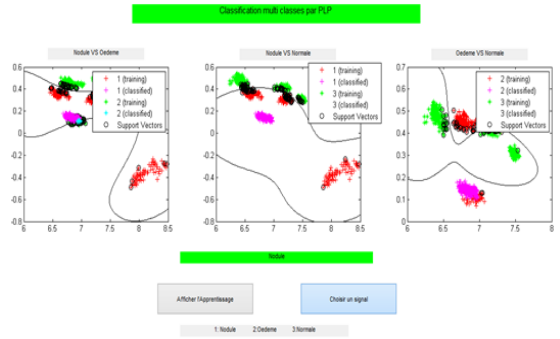


Fig.11. Example of a classified signal

The following table shows the results obtained after applying several parametrization methods, it should be noted that the signals used include male and female voices.

Table 1. number of signals for validation from TIMIT

Pathological signals		Normal signals	Total signal	Training signals
Nodule	Oeudem			
13	12	9	34	8

Table 2. Results of recognition

Recognition rate /2classes		Multi-class recognition rate	
PLP	MFCC	PLP	MFCC
88.23%	76.4%	85%	58%

In this table we have not presented the PLP RASTA method, because the latter classifies all the signals as being normal, and this leads us to say that the PLP RASTA eliminates the noise which caused the pathological signals to be considered as such, from suddenly these signals become normal following RASTA filtering.

4. Conclusion

In this paper we have developed an application in Matlab which aims to make a parametrization in order to perform a recognition of pathological voices.

This recognition is done using the SVM classification with several types of acoustic vectors (PLP, MFCC, and PLP-RASTA). According to the results obtained during the tests, we were able to observe that the parameters generated by the PLP-RASTA method give a less satisfactory result compared to the other two methods.

References

[1] René Boite, Hervé Boulard, Thierry Dutoit, Joël Hancq and Henri Leich, “speech processing book (ISNB: 2-88074-388-5)” Presses Polytechniques et Universitaire Romandes. 2000.

[2] : http://fr.wikipedia.org/wiki/Fenêtrage

[3] Zied Hajaiej, Kais Ouni, Noureddine Ellouze, “Speech Parametrization Based on Cochlear Filter Modeling: Application to PAR”. Article from the Signal Processing and Systems Laboratory (LSTS). ENIT.

[4] Julien PINQUIER “Sound indexing: search for primary components for audiovisual structuring” Thesis in Computer Science. University of Toulouse III – Paul Sabatier. December 20, 2004

[5] Houda HOSNI, Zied SAKKA, Abdennaceur KACHOURI and Mounir SAMET.« Étude de la Paramétrisation RASTA PLP en vue de la Reconnaissance Automatique de la Parole Arabe »

[6] Linda salwaMuda, MumtajBegam et I. Elamvazuthi « Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques »Paper from the Journal of Computing, Volume 2, Issue 3. 3 mars 2010

[7] CHERIF Adnen: Faculty of Sciences of Tunis El Manar “Audio processing and transmission course: digitization of the audio signal”

[8] Houda Hosni, Zied Sakka, Abdennaceur Kachouri and Mounir Samet. "Study of the Rasta PLP configuration for automatic recognition of Arabic speech", LETI laboratory of National School of Engineers in Sfax. 2009