Voice Detection in Traditionnal Tunisian Music using Audio Features and Supervised Learning Algorithms

Wissem Ziadi, Hamid Amiri
Signal, Images, Information Technologies (LR-SITI)
Tunisian National School of Engineering, ENIT
Tunis, Tunisia

***Abstract*—The research presented in this paper aims to automatically detect the singing voice in traditional Tunisian music, taking into account the main characteristics of the sound of the voice in this particular music style. This means creating the possibility to automatically identify instrumental and singing sounds. Therefore different methods for the automatic classification of sounds using supervised learning algorithms were compared and evaluated. The research is divided into four successive stages. First, the extraction of features vectors from the audio tracks (through calculation of the parameters of sound perception) followed by the selection and transformation process of relevant features for singing/instrumental discrimination. Then, using learning algorithms, the instrumental and vocal classes were modeled from a manually annotated database. Finally, the evaluation of the decision-making process (indexing) was applied on the test part of the database. The musical databases used for this study consists of extracts from the national sound archives of Centre of Mediterranean and Arabic Music (CMAM) and recordings made especially for this research. The possibility to index audio data (classify/segment) into vocal and instrumental recognition allows for the retrieval of content-based information of musical databases.**

Keywords*—Tunisian voice timbre; audio features extraction; singing voice detection; sung/instrumental discrimination; supervised learning algorithms

I. INTRODUCTION

Faced with the increasing availability of sound data broadcasted online, the importance to search these immense volumes of data by musical, sound and information-based content have become apparent for archiving and classification purposes. Several multidisciplinary researches have been carried out for this purpose, notably on the description of sound and musical contents. Among others the QUAERO[1] project, a collaborative industrial research and innovation program addressing automatic analysis and enrichment of digital, multimedia and multilingual content.

This project gathers 32 French and German partners. The CUIDADO[2] project from IRCAM is another major development scheme for the description of audio-visual content (the MPEG-7standardization process).

In Tunisia, the Telemeta[3] platform [1] is being developed in the national sound archives of the Centre of Mediterranean and Arab Music[4] (CMAM). The project is conducted under the direction of the Ethno-Musicology Research Centre in France. Telemeta is a collaborative platform for CNRS sound archives to analyse, identify and index digital sound resources.

In a piece of music, the singing voice is the main element. She carries the message. In order to automatically detect the sung voice in an audio stream, multiple methods have been implemented so far, but never on a corpus of traditional Tunisian music. This research explores different methods using supervised learning algorithms to extract the vocals in this particular music: K-nearest neighbours, the support vector machine and the Gaussian mixture model. The method chosen to locate the sung voice in an audio stream is inspired by the systems developed by Tong Zhang [2] and Peeters Geoffroy [3], [4]. This is a statistical method based on two phases: a learning phase and a classification phase. This process of supervised classification is based on a succession of four stages: First the extraction of relevant audio features [5], [6]. Followed by the selection and transformation of these feature vectors to minimize redundancy and reduce the dimensions. Then a modelling procedure of the instrumental and vocal classes is needed using learning algorithms [5], [7], [8]; finally, ending with an indexing phase. The training phase was carried out on a training database. A second database was used for the test phase (indexing). The extracts of Tunisian music used during this research were obtained from the funds of National Phonetics collected and saved by the National Archive of the Centre for Arab and Mediterranean Music. Another database was composed of live recordings and recordings made specifically for this study at Ixir studio. This was done in order to compare the outcome of different recording techniques and acoustic elements. The whole corpus was manually annotated indicating the sung and instrumental parts in the audio tracks.

This paper includes, in the first section, the musical context and a presentation of the audio features used for sung/instrumental discrimination. Then, in the second section, it presents a few methods for dimension reduction and a brief description of the supervised learning algorithms aiming to model the two instrumental and singing classes. The final

[1] www.quaero.org
[2] anasynth.ircam.fr/home/english/projects/cuidado

[3] telemeta.org
[4] phonotheque.cmam.tn/

section is reserved to the experimental results including corpus, post-treatments and evaluation.

II. Musical Context and Audio Features

There are different styles of traditional Tunisian music, including Soufi, Mezoued, Stambeli, Salhi combining spirituality, poetry, festivity and religion. The best known style though is Mâlouf. It is influenced by Arabic poetry brought to Tunisia by muslim-Andalusian immigrants in the 13th and 14th century. It is organized by quartertones following a classical Arab mode called 'maqâm' and carries Berber and Turkish elements in its rhythm. It features instruments like the violin, various percussion instruments, the 'Ud', and flutes, but it is essentially carried by the sung voice. The acoustic characteristics and timbre of the singing voice are genre-specific and distinguished by its harmonic sound and its vibration [9]. Styles differ greatly from region to region, each presenting a range of acoustic characteristics specific to each environment, instrumental technique and peculiarity of the vocals.

The first stage of the research tried to identify the typical aspects that characterize the sung voice in traditional Tunisian music. The analysis of the voice was carried out through the visualization of the items extracted from the sound signal (using VAMP plugin with Sonic Visualizer). Each detail is capable of describing the precise behaviour of an analysed signal. A multitude of features have been proposed in the literature, both about the field of speech processing as well as the classification of sounds, in our case singing vs. instrumental discrimination. The aspects best capable of distinguishing the singing voice from the rest of a musical sound stream of the Tunisian sung repertoire were identified. The most relevant features are: auto-correlation, Zero Crossing Rate (ZCR), Spectral Centroid, Mel -frequency Cepstral Coefficients (MFCC) and Harmonic Pitch Class Profiles (HPCP). Each feature describes one or more acoustic characteristics of sound and will be used by learning algorithms to establish a model for both instrumental and vocal classes [5], [7], [10].

1) *Auto-correlation*

Auto-correlation is used to compare the time lag of a signal with a delayed copy of itself. A periodic signal is perfectly correlated with itself if the delay time is the same as the duration of the signal. Autocorrelation is a relevant parameter for describing the mechanism of human listening, hence its relevance in the process of sound differentiation.

2) *Zero Crossing Rate (ZCR)*

This feature consists in locating the number of times the signal changes sign during a given time interval (in seconds), from negative to positive and vice versa. ZCR is the most relevant feature in voice / noise classification in the process of speech recognition and music information retrieval. The smaller the ZCR, the closer the sample is to the human voice.

3) *Spectral centroid*

This feature indicates the center of gravity of an audio signal. It is calculated as the weighted average of the frequencies present in the signal, determined using a Fourier transformation with their magnitudes as weight. The spectral centroid is used to estimate the brightness of a sound. It is a key aspect in describing the musical timbre.

4) *Mel-Frequency Cepstral Coefficients (MFCC)*

The MFCC is the most used feature in existing methods of automatic speech recognition and sound indexing. Bridle and Brown were the first to use it in 1974. The Mel-Frequency Cepstral Coefficients is able to simulate part of the speech production and perception. More exactly, the MFCC is a logarithmic representation of the loudness and pitch of a sound.

5) *Harmonic Pitch Class Profiles (HPCP)*

HPCP is a set of qualities commonly used for the recognition and identification of string instruments in an audio signal. This feature presents a sequence of vectors that describe the distribution of pitches in a single octave, specifying their tones and intensities according to a distribution over 12 temperate ranged classes. With HPCP you can determine the key of a song, search by similarity or, as in our case, index and classify sound signals.

III. Singing Voice Detection

A. Dimension Reduction

In this experimental study, audio features were extracted with a sliding window called "Hamming window" [11] with a size of 50 ms, the hop size set to 25 ms. Temporal modelling was applied using the mean and variance values over a 2s window with a hop size of 1s. After the extraction of features from an audio signal, the selection and transformation of the most relevant features was performed to describe the two classes (vocal and instrumental). A very large amount of features can cause confusion of the class models during the course of the experiment. Therefore the algorithms that select the features should be able to detect a minimal set of relevant (informative and meaningful) features in relation to the class models, avoiding a redundancy of data. The most popular filter type selection algorithm is the inertia ratio maximization using feature space projection (IRMFSP) algorithm [12]. A features space transformer was then applied. It aims to reduce the size of the features after the features selection phase. Several types of transformations are presented in the literature, such as:

- Box-Cox transformation which reduces the size of the features space while preserving the variance.
- The linear discriminant analysis (LDA) transformation reduces the space of audio features by maximizing class separation. LDA is based on predicting the belonging of a feature to a class depending on its characteristics measured using predictive variables.
- The principal component analysis method (PCA) extracts the principal component by a linear transformation, computed using singular value decomposition algorithms [13]. A new set of features are ordered according to their importance. This procedure consists of five steps: First, subtracting the mean from each audio feature vectors (mean is zero). Then, calculating the covariance matrix. After that, computing the eigenvalues and eigenvectors of the

covariance matrix. Then, order eigenvalue in descending order (the number of eigenvectors is equal to the dimension of audio features vectors). Finally, derive the new audio feature vectors (multiply the transpose of the audio feature vector to the left of the original data set. Final Features Vectors = RFE * RDM. Where RFE is the matrix with the eigenvectors (columns transposed) and RDM is the matrix mean-adjusted data transposed.

B. Supervised Learning Algorithms

The supervised learning phase of this system aims to model the two instrumental and vocal classes and subsequently permitting automatic classification of an unknown sound. The models are based on the information provided by the audio features vectors and manual annotations of the database, indicating the instrumental and vocal parts of each audio track. The most efficient methods chosen for the learning phase are:

1) *K-Nearest Neighbors (KNN)*

This method is considered the simplest and most popular of the supervised learning algorithms for the classification and automatic indexing of sounds [12], [14]. In our case, the data devoted to learning are formed by a set of vectors of audio features. Each sample contains a class label (instrumental or vocal) and is recorded in memory during the learning phase. During the indexing phase, the test samples are classified by assigning the class tags using the closest learning sample.

By definition, the implementation of this method results in two fundamental questions: the choice of neighbourhood (the value of K) and what distance to take into account. The results obtained by the method depend on these two criteria. We can arrive at totally different results depending on the choices made.

The most used method to determine the similarity between the samples is the Euclidean distance.

2) *Support vector machine (SVM)*

This method makes it possible to use the samples closest to the separation boundary, assuming that it provides the most useful information for the classification. These are called support vectors. SVMs were developed in the 1990s based on Vladimir Vapnik's theoretical considerations on the development of a statistical theory of learning: the Vapnik-Chervonenkis theory [15], [16].

This theory seeks to maximize the margin between the separation boundary and the closest samples. The problem then lies in finding the optimal decision surface for the separation between classes, and subsequently predicting which class a test sample belongs to.

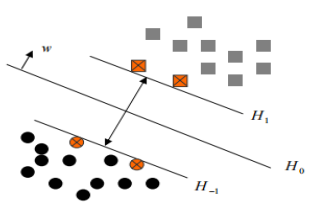

Fig. 1. SVM model.

In our case the issue is typical: the classification into two classes: instrumental and vocal, in which the samples (or feature vectors) are linearly separable. The separation boundary is used to classify a frame to one of two classes.

It is named in Fig. 1 the optimal hyper plan H_0.

We note the function f that has an input vector $x \in \Re^n$ and matches an output y: $f(x) = y$. In our case (question of two-class discrimination) $y \in \{-1,1\}$. The input vector $x = (x_1 \dots x_n)$ and a weight vector $= (w_1 \dots w_n)$. Which gives:

$$f(x) = w \cdot x + b \quad (1)$$

H_0 is the region of the vector x that checks the equation $f(x) = 0$. H_1 and H_2 are two parallel hyper plans to H_0 and which are defined respectively by $f(x) = 1$ and $f(x) = -1$.

The distance between the two hyper plans H_1 and H_2 is $\frac{2}{\|w\|}$

The intention is to maximize the margin. The decision hyper plan H_0 depends directly on the vectors closest to the two hyper plans H_1 and H_2 who we call support vectors.

3) *Gaussian mixture model GMM*

For its ability to approximate the global distribution of the features collection for each class [7], [8], the Gaussian Mixture Model was used. A GMM is used to model the distribution of data in the features space at D dimensions. This space is obtained from the weighted sum of N probability density function (pdf).

The probability to observe the feature vector x knowing its GMM is defined by the parameters $\lambda = \{\mu, \sum\}$ resulting in:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{N} p_i \, b_i(\vec{x}) \quad (2)$$

Where $\vec{x}$ is a D dimension of feature vector and p_i , i = 1…N, is the weight associated with each GMM. $b_i(\vec{x})$, i = 1…N, is the probability density, which can be written as follows:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_i|^{\frac{1}{2}}} \exp\{-\frac{1}{2} (\vec{x} - \vec{\mu}_i)' \sum_i^{-1} (\vec{x} - \vec{\mu}_i)\} \quad (3)$$

In the above equation $\vec{\mu}_i$ and $\sum_i$ represent the estimated means vector and the covariance matrix. The mixture weight satisfies the constraint $\sum_{i=1}^{N} p_i = 1$. The parameters of the GMM are the mean vector, the covariance matrix and the mixture weight. They can be represented as follows:

$$\lambda = \{p_i, \vec{\mu}_i, \sum_i\} \; ; \; i = 1, \dots, N. \quad (4)$$

These parameters will be initialized by the classical K-means algorithm and then iteratively adjusted by the Expectation-Maximization algorithm [17]. The model of the Gaussian mixture of the vocal class was noted as « λ_C » and that of the instrumental class as « λ_{In} ». These two models will be trained by the learning algorithm from the manual annotations of the database, distinguishing the vocal parts from the instrumental parts.

During the test phase, the classifier takes as input the features vectors of $T_{x-frame}$ $X = (x_1, x_2, \dots, x_{T_x})$, extracts from an audio track frame in the test data base, and outputs the log likelihood of each frame $log\ p(x_t|v_C)$ and $log\ p(x_t|v_{In})$; $1 < t < T_x$, for the vocal and Instrumental

GMM. Each frame was assigned to the corresponding class according the logarithmic probabilities. Depending on the choice of analysis interval, there are many variants and combinations that had to be taken into account during the decision making process [8]. In this study, decisions were made based on a fixed segment:

If $log\ p(x_t|v_C) > log\ p(x_t|v_{In})$, the frame is classified as vocal. Consequently if $log\ p(x_t|v_C) \leq log\ p(x_t|v_{In})$, the frame is classified as instrumental.

IV. Experimental Study

A. Corpus

The research corpus is composed of two databases; the CMAM database and the Ixir database. All tracks in both databases are 16-bit coded with a sampling frequency of 44.1 kHz. Both databases were annotated manually, dividing them into two classes, vocal and instrumental. The learning process was done on 2/3 of the database, called the training database. The test phase was conducted on the 1/3 of the database left, called the test database.

- The CMAM database: Was extracted from the Funds of National Phonetics, collected and saved by the Centre of Mediterranean and Arab Music (CMAM). In this centre, Tunisian phonographic collections (collected since the beginning of the 20th century) are catalogued and indexed. Access to the music records is obtained through a Telemeta platform. Our research database is composed of 403 tracks in wav format and corresponds to different Tunisian music styles and epochs. The duration of the extracts is between 2 and 60 seconds. All tracks combined represent about 3:50 hours (210 min) of music. Table I gives a description of the distribution.

TABLE I. Description of the distribution of the CMAM database

	Vocal	Instrumental	Total
Training	***155***	***139***	***294***
Test	***56***	***53***	***109***
Total	***211***	***192***	***403***

- The Ixir database: has been partly recorded in the studios of “el Xir Labs – studio” at the Centre for Music and Sound Studies in Tunis and partly live specifically for this research. We did so in order to compare the outcome of different recording techniques and acoustic environments. This database consists of 186 tracks. We opted for a variety of styles from the Tunisian repertoire. The covers are performed in voice and lute (‘ud’) by Wissem Ziadi. The audio tracks are between 2 and 60 seconds long, which makes the set about 1h30 (90 min) of music. Table II gives a description of the distribution of the Ixir database:

TABLE II. Description of the distribution of the Ixir database

	Vocal	Instrumental	Total
Training	***62***	***62***	***124***
Test	***31***	***31***	***62***
Total	***93***	***93***	***186***

B. Tests and Evaluation

The tests were done on a standard PC; Intel core i4, 2.3 Gh, 4 GB of RAM. Many simulations have been performed to evaluate the performance of the classification methods. All these simulations were made in python. The goal of this system is to assign each unknown analysis track from the test database to either vocal or instrumental.

1) *Post-treatment*

Due to the use of a very short decision window (50 ms), much less information is used for the calculation of each feature. This necessarily implies a greater variability of the estimated results. To reduce the noise and to refine the results after assigning the class probabilities (abrupt or accidental values), several types of filters are proposed in the literature [18], [19]. A median filter and a smoother using a hidden Markov model were applied. The median filter is a nonlinear filter. Therefore a median is a value m which serves to partition a set of values in two equal parts; on one side the smaller ones, on the other the bigger ones. In our case, the set of values is a probability distribution. The median can only be the value for which the density function is 0.5. The median filter applies to a distribution of N odd values to find the median that divides the samples into two equal groups. The application of the median filter brought a remarkable improvement in the performance of the classification system, but it remained blind to the nature of the classes. Therefore the Markov hidden model process was applied. The hidden Markov model smoothing process presented in this experiment is inspired by the famous Rabiner tutorial [20] and implemented in practice by Ramona [16]. Here, the purpose of the Markov model is to present the transitions between a set of states (the acoustic classes: vocal and instrumental.

2) *Résults*

a) Cross validation method

For each classification method, recall, precision and F-score were measured. The recall is the fraction of vocal frames (or instrumental) existing on all found frames. It is a measure of sensitivity. Precision is the fraction of the vocal frames found on all the vocal frames of the database. It is a measure of confidence.

F-score = 2 (Recall * Precision) / Recall + Precision.

This is the measure of competence.

For the evaluation of this system the cross validation method was used.

Table III first shows that the GMM model, applied to both databases, is the most reliable model for instrumental / vocal discrimination. The precision rate the GMM method has given ranges from 89.6% with the CMAM database up to 95.8% with the Ixir database. The Ixir database gave the best results with all three classification methods. This could be explained by the difference in the acoustic nature (recording techniques) of the records between the CMAM and Ixir databases, the complexity of the models and the large number of parameters related to the number of samples.

TABLE III. TEN-FOLD CROSS VALIDATION VOCAL CLASSIFICATION RATE OF KNN, SVM AND GMM METHODS FOR THE TWO DATABASES

	Database	
	CMAM	**Ixir**
KNN	***81.4%***	***91.5%***
SVM	***74.4%***	***88.3%***
GMM	***89.6%***	***95.8%***

The different values of the recall, precision and F-score factor should be analysed. These values depend directly on the distribution of the entire database. Running the CMAM database with the SVM learning model for example, F-score=74.4%. We got 94.4% as recall for the vocal class. While the precision rate is only 61.4%. This seems very unsatisfactory as a significant part of the vocal frames are classified as instrumental. But, in reality, this is due to the imbalance of the data composition between both databases since they don't have the same percentage of sung frames compared to instrumental frames. The CMAM database consists of 22620 frames belonging to the vocal class and 75875 frames belonging to the instrumental class. While the Ixir DB is composed of 7512 sung frames compared to 4375 instrumental frames.

b) Cross database validation

In this part of the test, a cross database validation method was applied. With this method the generality of the established classification systems is tested and checked if the system learned the general and specific acoustic characteristics of the sound signal. The procedure consists of starting the supervised learning process with the first database and later launching the test phase with the second and vice versa. The results of this test are given in the tables below.

Tables IV, V and VI show the classification results. They are declining and do not exceed 67.5% with the Gaussian mixture model. This is a fairly logical result due to the differences between the acoustic characteristics of both CMAM and Ixir records. The extracts retrieved from the CMAM database are recordings registered since the beginning of the 20th century. Hence the quality of the recording and post processing of the sound varies greatly from one era to another. While the Ixir database is registered under technically optimal conditions for this study. These differences imply a mutation of the extracted features from the audio frames and their distributions and therefor generate confusion in the classification process.

TABLE IV. CROSS-DATABASE VALIDATION (CMAM AND IXIR DATABASE) - KNN ETHOD

		Test database	
		CMAM	**Ixir**
Training Database	**CMAM**		***61.2%***
	Ixir	***60.5%***	

TABLE V. CROSS-DATABASE VALIDATION (CMAM AND IXIR DATABASE) - SVM ETHOD

		Test database	
		CMAM	**Ixir**
Training Database	**CMAM**		***56.5%***
	Ixir	***59.3%***	

TABLE VI. CROSS-DATABASE VALIDATION (CMAM AND IXIR DATABASE) - GMM ETHOD

		Test database	
		CMAM	**Ixir**
Training Database	**CMAM**		***65.7%***
	Ixir	***67.5%***	

V. CONCLUSION

This article presented the search to find the technical means and the appropriate tools to automatize the detection of the voice in traditional Tunisian music. An analytic spectrum of an audio signal to identify relevant audio features for vocal/ instrumental discrimination was established. Then selections and transformations algorithms to minimize redundancy and reduce the dimensions of the features vectors space extracted from audio tracks were implemented. The KNN, SVM and GMM methods were used to model the singing and instrumental classes in an approach that is based on firstly a learning and then a testing phase. The experimental results show that the supervised learning algorithm based on the Gaussian mixture models GMM have the best precision rate, resulting in 95.8% accuracy with the Ixir database. The Ixir database gave the best results with all three classification methods.

During this study some problems were encountered that distorted the results, like the complexity of the statistic models, the difference in the acoustic nature of the audio records and the large amounts of parameters in the classification process, related to the number of samples.

Despite these complications, the results were encouraging and open new perspectives in terms of sound analysis and supervised classification of Tunisian music through learning algorithms. However, the complexity and large amount of perimeters used during this study, makes it still a very time consuming process.

For further study, we would like to propose experimenting with the extraction of other audio features such as Linear Predictive Coefficients (LPC) or Perceptual Linear Prediction (PLP) [7] which give a spectral representation of the spoken voice and which is widely used in the field of speech-processing.

Other learning algorithms for supervised classification such as naive Bayesian or Artificial Neural Network classification have also given excellent results in other studies [20] and might be interesting to try with Tunisian music.

It would also be interesting to extend this study on all acoustic categories of Tunisian music; an analysis of the different timbres of the instruments used in the Tunisian tradition. This study would lead to the detection of any Tunisian instrument to classify and index a musical database (National Phonetics) in different acoustic categories to facilitate access and exploration of the musical and sound heritage.

REFERENCES

[1] Thomas Fillon, Josephine Simonnt, Marie-France Mifune, Stéphanie Khoury, Guillaume Pellerin, Maxime Le Coz, Estelle Amy de la Bretèque, David Doukhan and Dominique Fourer. Telemeta: An open source web framework for ethnomusicological audio archives

management and automatic analyses. Conference Paper, Journal of New Music Research, 2014.

[2] T. Zhang, "System and method for automatic singer identification". IEEE International Conference on Multimedia and expo, HPL, 2003.

[3] G. Peeters, "A generic system for audio indexing: Application to speech/music segmentation and music genre recognition". Proc. of the 10 international conference on digital audio effects (DAFx-07), Bordeaux, France, 2007.

[4] P. Herrera, G. Peeters and S. Dubnov, "Automatic classification of musical instrument sound". Journal of new music research, 2010.

[5] P. Herrera, G. Peeters and S. Dubnov. Automatic classification of musical instrument sound. Journal of new music research. Vol : 32, 2010.

[6] G.Peeters. Descripteurs audio: de la simple représentation aux modèles de connaissances. Geste sonore et paramètres. L'analyse musicale à l'heure des outils multimédia, Jan 2015, Paris, France, 2015.

[7] T. Ratanparaand N. Patel. Singer identification using perceptual features and cepstral coefficients of an audio signal from Indian video songs. Ratanpara and Patel EURASIP Journal on Audio, Speech, and Music Processing, 2015.

[8] W.Tsai and H.Wang. Automatic singer recognition of popular music recording via estimation and modeling of solo vocal signals. 2-MUSI Signal Processing for music, 2005.

[9] R. Miller and J. Franco. Analyse spectrographique de la voix chantée. NATS Journal, 1995.

[10] Inderjeet Singh, Shashidhar Koolagudi. Classification of Punjabi Folk Musical Instruments Based on Acoustic Features. Proceedings of the International Conference on Data Engineering and Communication Technology, pp.445-454, 2017

[11] Shruti and Bharti Chhabra. An Approach for Singer Identification Technique Using Artificial Neural Network. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, 2015.

[12] G. Peeters, "Automatic classification of large musical instrument data base using hierarchical classifiers with inertia ratio maximization". Audio Engineering Society, Convention Paper, USA, 2003.

[13] P. Huang, S.D. Chen, P. Smaragdis and M. Hasegawa-Johnso. "Singing-Voice separation from monaural recordings using robust principal component analysis". IEEE, 2012.

[14] M. Kalamani1, Dr.S.Valarmathy, S.Anitha. Automatic Speech Recognition using ELM and KNN Classifiers. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 3, Issue 4, 2015.

[15] Lhoucine Bahatti , Omar Bouattane , My Elhoussine Echhibat , Mohamed Hicham Zaggaf . An Efficient Audio Classification Approach Based on Support Vector Machines. International Journal of Advanced Computer Science and Applications pages 205 -211, 2016.

[16] Mathieu Ramona. Classification automatique de flux radiophoniques par Machines à Vecteurs de Support. Thésée à l'Ecole Télécom ParisTech , Spécialité : Signal et Images, 2010.

[17] Frédéric Santos. L'algorithme EM : une courte présentation. CNRS, 2015.

[18] Guillaume Noyel. Filtrage réduction de dimension, classification et segmentation morphologique hyper-spectrale. Thèse à l'Ecole des Mine de Paris, spécialité : morphologie mathématique, 2008.

[19] Lawrence Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. IEEE, 77(2): pages 257–286. (1989).

[20] Vikramjit Mitra, Chia-Jiu Wang. Content based audio classification: a neural network approach. Soft computing, Methodology and Application, volume 12, issue 7, pp 639–646. 2008