# Innovative Automatic Discrimination Multimedia Documents for Indexing using Hybrid GMM-SVM Method

Debabi Turkia[1], Bousselmi Souha[2], Cherif Adnen[3]
Laboratory Analysis and Processing of Electrical and Energy Systems
Faculty of Sciences of Tunis, FST
Tunis, Tunisia

***Abstract*—In this paper, a new parameterization method sound discrimination of multimedia documents based on entropy phase is presented to facilitate indexing audio documents and speed up their searches in digital libraries or the retrieval of audio documents in the network, to detect speakers in purely judicial purposes and translate films into a specific language. There are four procedures of an indexing method are developed to solve these problems which are based on (parameterization, training, modeling and classification). In first step new temporal characteristics and descriptors are extracted. However, the GMM and SVM classifiers are associated with the other procedures. The MATLAB environment is the basis of the simulation of the proposed algorithm whose system performance is evaluated from a database consisting of music containing several segments of speech.**

***Keywords*—Audio indexing; classification; GMM; SVM; entropy; Speech-music discrimination***

## I. Introduction

The significant development of the digital database and the Internet containing several multimedia documents requires new intelligent tools to structure and index this data to reduce delays and improve the classification report. The automatic discrimination objective at extracting several descriptors of the digital flow, allowing bout to the information via its contents. Extraction of descriptors, for musical signals, allows deducing the original score, the type of song, and the signature of the sound document. To search for a special document within a collection from a description of another document, this operation may consist of browsing a collection to browse its contents, summarizing the collection, archiving the automatic description of documents and using these descriptions to produce new documents or new services (TV emission, films and radio, etc.). In [1], the authors developed a multimedia indexing system in [12], using the GMM in [23] and SVM methods in [2]-[3]-[11] and [27], applied to a television broadcast. In [13]-[17]-[18]-[19]-[25] and [26], presented works introducing semi-automatic segmentation, music classification, discrimination and transcription founded on novel descriptors and training, published two other studies on audio classification, in [9] and [10]. In [4]-[14]-[15]-[20] and [21], the authors differentiate two technical classifications: Gaussian Mixture Models (GMM) and Vector Support Machines (SVM) used for indexing audio tasks. However, there is considerable variation in the performance of these techniques from one database to another.

In this framework, a new parameterization method sound discrimination of multimedia documents based on entropy phase is presented with several approaches of discrimination and structuring audio documents are proposed for detecting the primary components as speech and music.

The aim of this work is to propose an easily and automatically adaptable sound classification approach to multimedia application. A hybrid model that is the basis of the proposed approach using a Gaussian mixing classifier and support vector machines applied to the sound classification: classification (class/non-class) and finally recognition of the type of audio document. In this project, two developed applications: The first concerns discrimination between music and non-music (music/speech). The second is to index the type of multimedia documents to facilitate search and navigation (speaker1/speaker2, music/song).

The organization of this article is as follows: Section 2 focuses on the indexation models; section 3 describes the proposed entropy phase discrimination method with hybrid model GMM/SVM; section 4 is devoted to simulation results analysis; finally, concluding remarks are discussed in section 5.

## II. Indexation Models

In this framework, segmentation and indexation approaches to sound documents are proposed.

Their objective is to spot the primary components for speech and music. Fig. 1 illustrates the indexing system for audio documents.

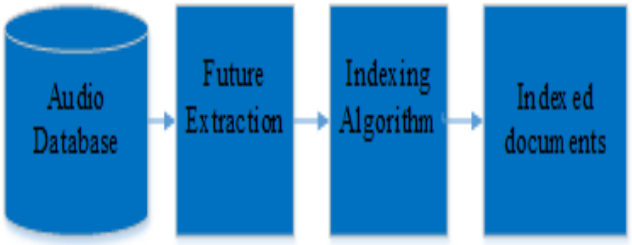


Fig. 1. Indexing System.

### A. Discrimination Speech/Music: Future Extraction with Entropy

Shannon determines Hof's entropy as being a discrete random variable X with possible values {x1… xn} and probability mass function P(X) as [6]:

$$H(X) = \mathbb{E}[I(X)] = \mathbb{E}\left[-\ln\left(P(X)\right)\right] \quad (1)$$

$\mathbb{E}$: the operator of the expected value and I: the content of X and I (X): a random variable.

Entropy can be expressed with:

$$H(X) = \sum_{i=1}^{n} P(x_i)I(x_i) = -\sum_{i=1}^{n} P(x_i)\log_b P(x_i) \quad (2)$$

When b is the base of this logarithm, the common values used for b are 2, e and 10 being the Euler number, and for the entropy units are the bits for b = 2, for b = e and the bans if b = 10. If P (xi) = 0 for the same I, the value of the sum for the corresponding sum 0 logb (0) is equal to 0, which is consistent with the limit.

$$\lim_{p \to 0+} plog(p) = 0$$

The definition of the conditional entropy of two events X and Y respectively admitting the values xi and yi, as follows:

$$H\left(\frac{X}{Y}\right) = -\sum_{i,j} p(x_i, y_i)\log\frac{p(x_i,y_i)}{p(y_i)} \quad (3)$$

The probability (xi, yi) for X = xi and Y = yi is understood as the amount of randomness for X the random variable considering Y which is the event.

### B. Gaussian Mixture Models

The GMM, in pattern recognition, is based on Fisher's algorithm [5] and [8], he envisions that each vector is part of a class, a probability distribution function (pdf) is a model for each class it type Gaussian Mixture Model:

Training phase: is the phase of estimating pdf templates (or parameters) for each class.

Classification phase: This is the decision phase by calculating the maximum log-likelihood criterion for each test observation.

k Gaussian laws are combined for a Gaussian Mixture Model (GMM). For that, the weighting of each law is (pk) and two parameters are specific: the average μk and the covariance matrix ∑k.

$$f(x) = \sum_{k=1}^{K} p_k N(x, \mu_k, \sum_k) \quad (4)$$

$$p_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} p_k = 1 \quad (5)$$

A simpler approach is used to detect the two basic components: music and speech. In this context, we define for the classification of the two systems, one to detect the music and one for the speech using class/non-class for the classification approach. For this purpose, the results of the two systems are merged seeking segments containing speech/music. A GMM classification system is well defined for each type of sound: speech/non-speech, Fig. 2 summarizes this system.

### C. Support Vector Machine

For classification or regression challenges, the supervised machine learning algorithm: Support Vector Machine (SVM) is used [7]-[16]-[22] and [24]. Indeed, in classification problems it is commonly used. In a space with n dimensions it draws each data or entity in the form of a point, each value of a particular coordinate is the value of each entity. Then, he proceeds to a classification: he searches for the hyper-plane which distinguishes in a certain way the two classes Fig. 3.

The resolution of an SVM, it is necessary to find the function of decision which makes it possible to classify the data and any vector xi must satisfy:

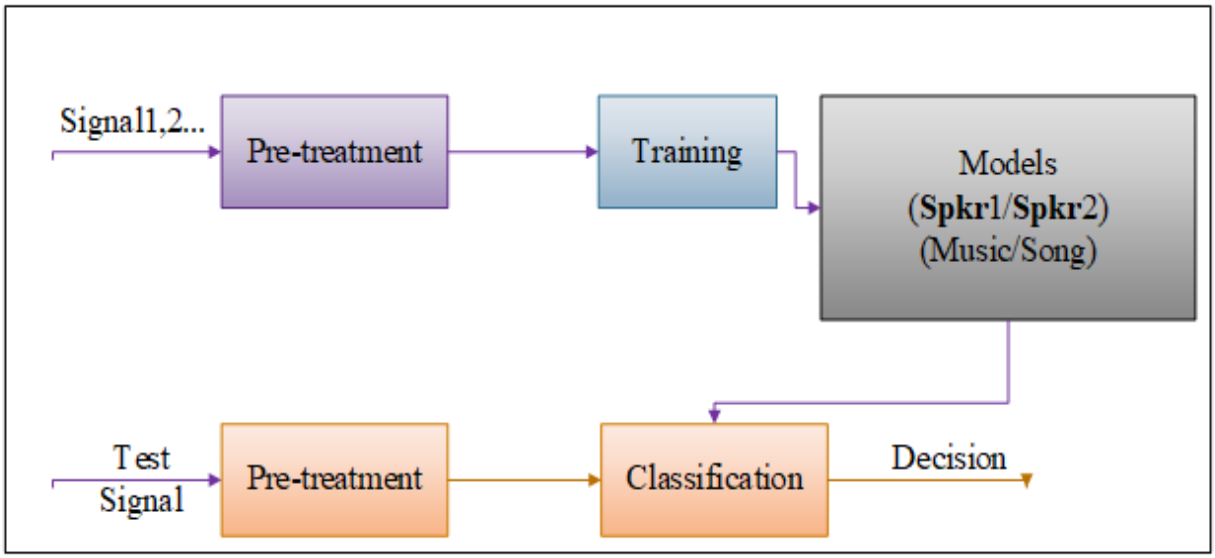


Fig. 2. Gaussian Mixture Models (Spkr: Speaker).

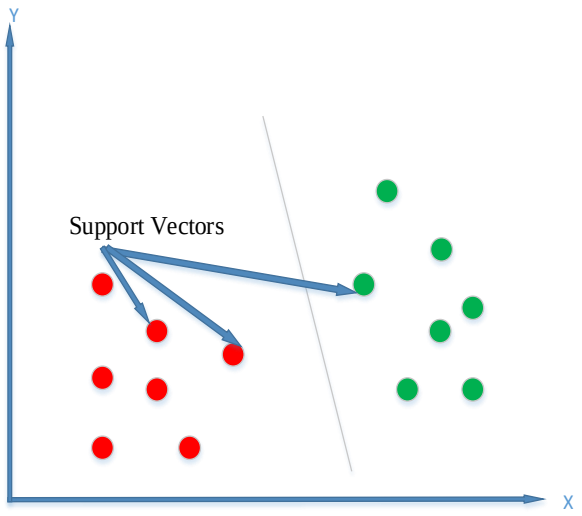


Fig. 3. SVM Classification.

$$\begin{cases} w.x_i + b \geq +1 & \text{if} \quad y_i = +1 \quad (6) \\ w.x_i + b \leq -1 & \text{if} \quad y_i = -1 \quad (7) \end{cases}$$

The restrictions are expressed by:

$$y_i(w.x_i + b) \geq 1 \quad (8)$$

$1/\|w\|$ is the distance between the contour for each class and the separation function. To solve the problem of an SVM one must minimize $\|w\|$ subject to (8), returns to a quadratic problem. The resolution of such a problem consists of converting it into a double expression by using the Lagrange multipliers method, which is a typical approach with the following form:

$$L_0 = \sum_{i=1}^{l} \propto_i - \frac{1}{2}\sum_{i,j=1}^{l} \propto_i \propto_j y_i y_j K(x_i . y_j) \quad (9)$$

$$\sum_{i=1}^{l} \propto_i y_i = 0 \quad (10)$$

Under constraints:

$$0 \leq \propto_i \leq C \quad i = 1, \ldots, l \quad (11)$$

With xi = training vectors, yi = class label of xi, K (....) = is the function of the kernel, C = is compromised between the erroneous classification of the learning data and that of the margin reached.

The SVM decision function:

$$f(u) = \text{sign}\left(\sum_{i=1}^{l} \alpha_i y_i K(u, x_i) + b\right) \quad (12)$$

## III. Entropy Phase Discrimination Method with Hybrid Method GMM/SVM

In this research works, discrimination algorithm for indexing audio documents is performed using four most habitually steps: parameterization (calculate entropy phase of the signal), training (hybrid method GMM and SVM classifiers), modeling and classification (Fig. 4).

Our work is to offer an easily and automatically adaptable sound discrimination approach to multimedia content and application. The proposed approach is based on an entropy phase summarized with the following Fig. 5.

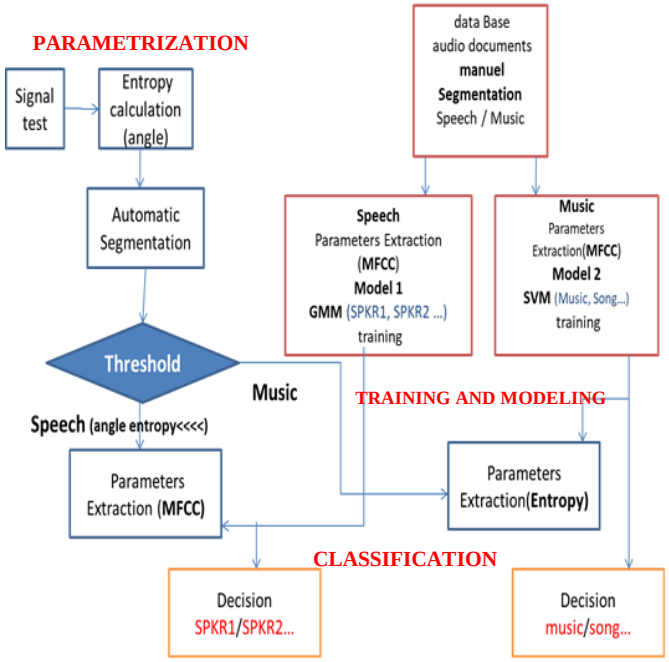


Fig. 4. Discrimination Algorithm for Indexing with Hybrid Model.

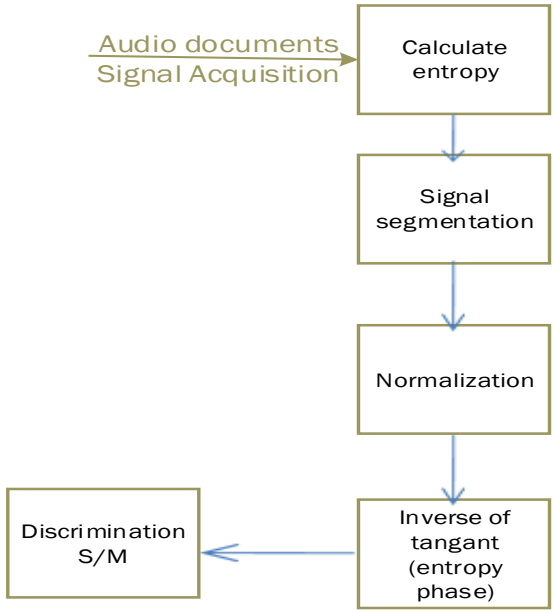


Fig. 5. Audio Documents Automatic Discrimination with Calculating Entropy Phase for Parameterization.

## IV. Simulation Results Analysis

STEP1: In this step, after acquisition of audio documents signal, entropy is applied (Fig. 6) and is divided into stationary frames (in which is projected entropy on the x-axis) (Fig. 7), then the normalization of the signal is tried (Fig. 8) and inverse of a tangent to find the angle of the signal (Entropy phase) (Fig. 9) and finally (Fig. 10) resumes discrimination (S/M) with a global thresholding is applied.

The following Fig. 6 illustrates a typical plot of the original signal and the entropy that demonstrates: entropy decreases with speech then with music.

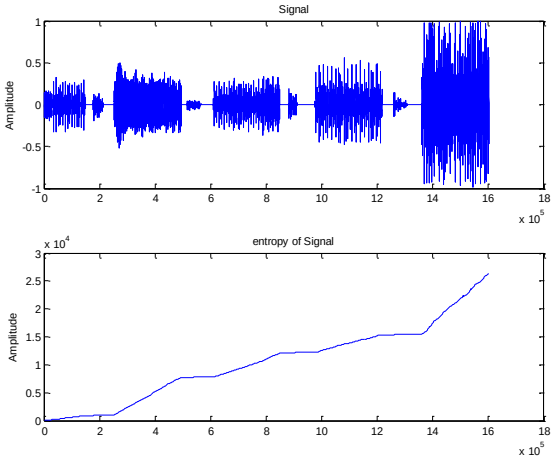


Fig. 6. Entropy of the Signal.

Fig. 7 illustrates a typical plot of the original signal and the segmentation of the signal. In which, the entropy is projected on the x-axis to show clearly the comparison between entropy phase with speech then music.

The figure below (Fig. 8) described a typical plot of the original signal and signal normalization.

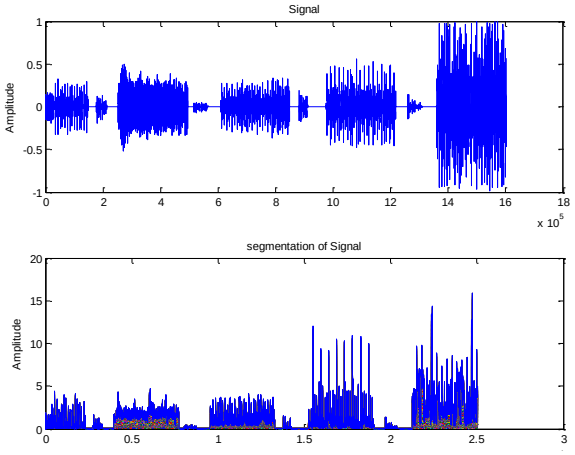


Fig. 7. Signal Segmentation.

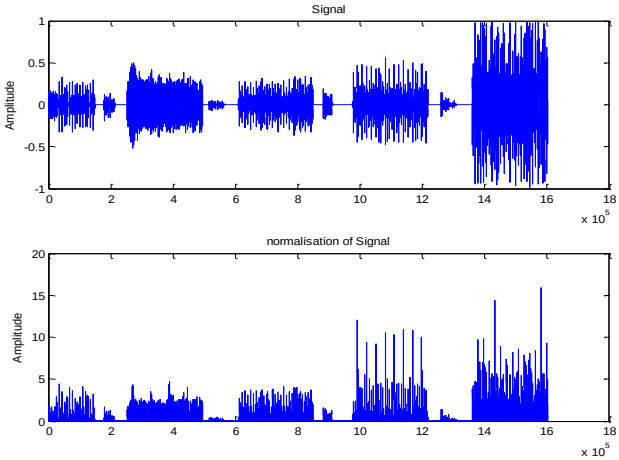


Fig. 8. Signal Normalization.

The following Fig. 9 illustrates a typical plot of the original signal and the entropy phase. As observed, this figure proves that music has an angle of entropy greater than that of speech.

Fig. 10 below illustrates a typical plot of the original signal and the angle of entropy. As seen, this figure shows automatic discrimination of the signal (music and speech) with a threshold calculated using the mean and the maximum of the variation of the entropy angle.

Fig. 11 illustrates a test sample that is composed of an alternation of voices (male and female speaker) and music. On the waveform (signal), the speech, music and song sections are indicated, and the instant musical probability is marked on the lower graph which is compared to a threshold calculated automatically.

Fig. 12 demonstrates clearly the efficiency of our automatic segmentation and indexation method, applied to 100 audio documents (constituted of music, songs with several speech segments), that 3 documents present a discrimination error: for 01 h: 26 m: 24 s the indexation error is of the order of 12 s.

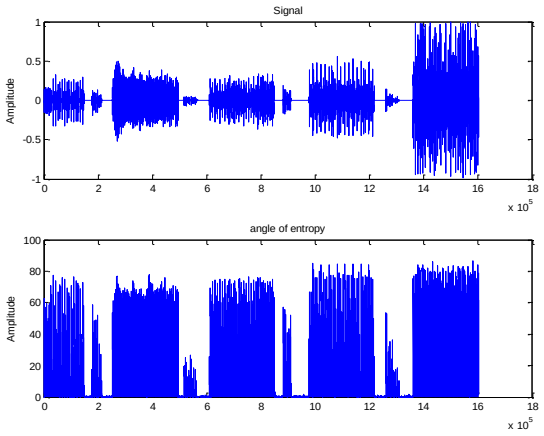


Fig. 9. Angle of Entropy.

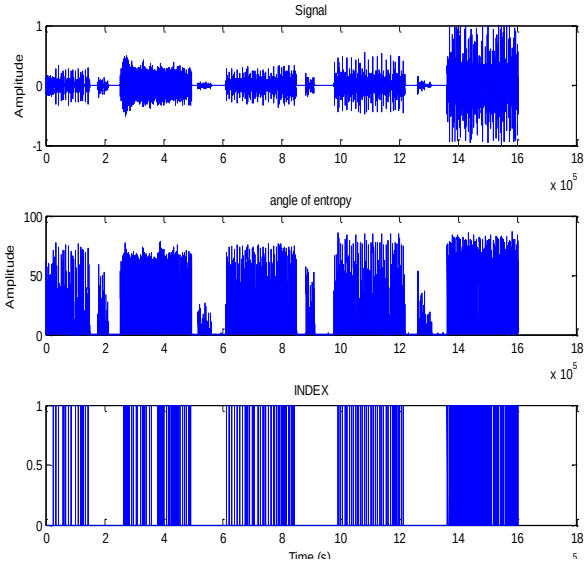


Fig. 10. Automatic Discrimination of Music and Speech.

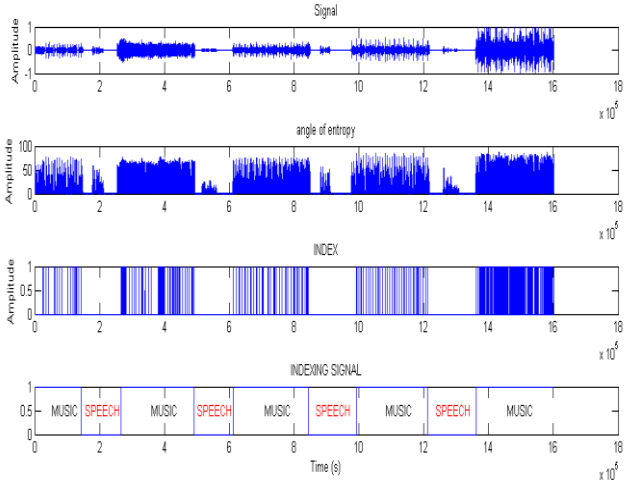


Fig. 11. Discrimination of Music and Speech.

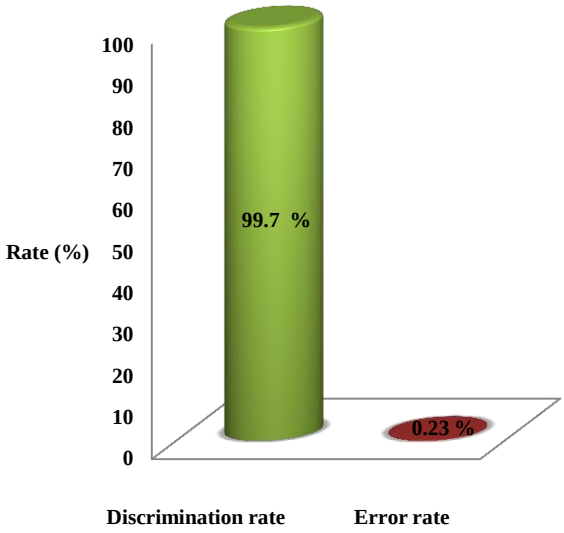


Fig. 12. Percentage of Discrimination Rate and Error.

- STEP2: After performing the discrimination (parameterization) of audio documents with threshold value, in second step, training techniques used in audio indexing tasks: Gaussian Mixture Models (GMM) [1] with speech (window frame: 1024, Gaussian number: 4, MFCC coefficients: 13, iteration number: 10) and Support Vector Machines (SVM) for music.
- STEP3: The third step focuses on modeling speech and music, is to propose an approach of modeling sound adaptable in an easy and automatic way to the multimedia content and application. A hybrid model of the proposed approach is based on the use of a Gaussian mixing classifier and support vector machines applied of sound modeling: the modeling in music/song, man/woman(SPKR1, SPKR2), and class/non-class.
- STEP4: In this step, the proposed approach is based on a hybrid model GMM and SVM applied of sound classification to indexing multimedia documents in order to facilitate the search and navigation.

In training and modeling techniques data base TIMIT, music and songs from the internet are used.

Finally, Fig. 13 shows the index for different speakers and types of musical documents (music/song).

Comparing the error rate with another method of discrimination and indexing, we find with our algorithm a very negligible error rate, Fig. 14 comparing the error rate with another method of discrimination and indexing, we find with our algorithm a very negligible error rate, Fig. 14 summarizes the indexation rate.

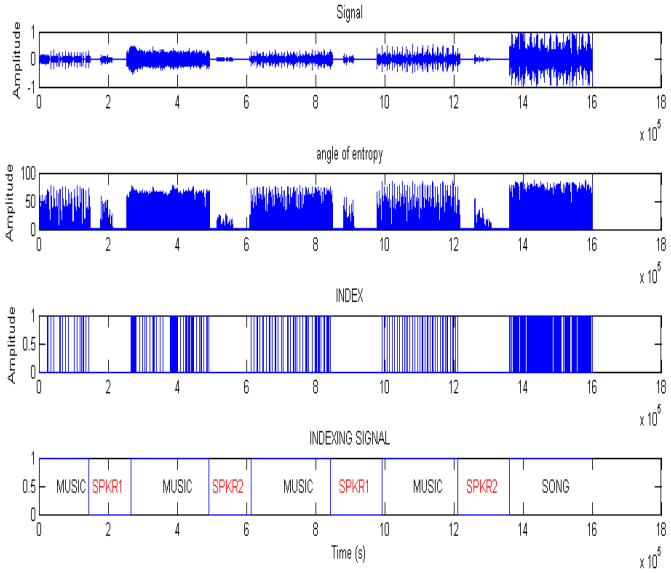


Fig. 13. Indexing Class of Speech and Music with GMM/SVM.

SPKR1:Speaker1

SPKR2: Speaker2

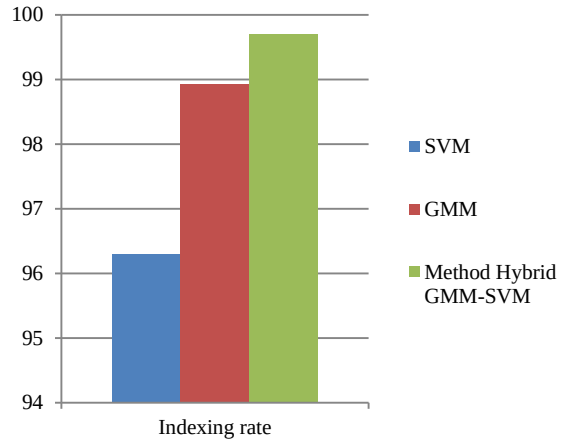


Fig. 14. Comparison of Indexing Rate with Differents Methods.

## V. CONCLUSION

In this paper, a new parameterization method audio discrimination system based on entropy phase is presented and implemented with an hybrid model GMM/SVM classification which is intended for automatic multimedia search documents: The system that has been developed has a discrimination tool which is based on Entropy phase calculation to separate SPEECH/MUSIC and an indexation tool based on GMM/SVM four procedures (parameterization, training, modeling and classification) to indexing different speakers and kind of music (music/song) in audio documents.

In our study, measurements and simulations are automatically determined by optimal parameters (threshold); on the system to indexing performances for the audio database is a set of speech (speaker1, speaker2), music and songs.

In conclusion to have a better indexation, we must find a good discrimination; this is the case of our work.

## REFERENCES

[1] Pinquier I., "Evaluation of classification techniques for audio indexing", IRIT, University of Toulouse, France, 2018.

[2] Rahona M., "Automatic classification of radio streams by SVM" , PHD Thesis, Telecom Paris-Tech, France, 2010.

[3] Durrieu J.L., "Transcription and automatic separation of the main melody in music signals", PHD Thesis, Telecom Paris-Tech, France, 2010.

[4] José Anibal Arias, "Evaluation of technical classification of audio indexing", 2015.

[5] Santiago Álvarez-Buylla Puente, "Single and multi-label environmental sound classification using convolutional neural networks", Audio Technology Group Chalmers University of Technology Gothenburg, Sweden, 2018.

[6] Damien Nouvel., "Information theory and Entropy Measures", National Institute of Oriental Languages and Civilization.

[7] Ma, Y., & Guo, G. , "Support Vector Machines Applications", Vol. 9783319023007, pp. 1–302, Springer International Publishing, 2014.

[8] B. Fernando, E. Fromont, D. Muselet, and M. Sebban, "Supervised Learning of Gaussian Mixture Models for Visual Vocabulary Generation", Pattern Recognition, 45(2) :897–907, 2012.

[9] M. Fradet, "Contribution to the Segmentation of Image Sequences in the Sense of Motion in a Semi-automatic Context", PHD thesis, Universite de Rennes 1, France, 2010.

[10] C. Joder, S. Essid, and G. Richard, "Temporal Integration for Audio Classification With Application to Musical Instrument Classification", IEEE Transactions on Audio, Speech, and Language Processing, 17(1) :174‑186, 2009.

[11] Dhanalakshmi, P., S. Palanivel, and Vennila Ramalingam, "Classification of audio signals using SVM and RBFNN," Expert systems with applications 36.3, 2009: 6069-6075.

[12] Lu, Goujun, "Indexing and retrieval of audio: A survey," Multimedia Tools and Applications 15.3, 2001: 269-290.

[13] Kiranyaz, Serkan, Ahmad Farooq Qureshi, and Moncef Gabbouj, "A generic audio classification and segmentation approach for multimedia indexing and retrieval," IEEE Transactions on Audio, Speech, and Language Processing 14.3, 2006 : 1062-1081.

[14] Agostini, Giulio, Maurizio Longari, and Emanuele Pollastri, "Musical instrument timbres classification with spectral features," EURASIP Journal on Advances in Signal Processing2003.1, 2003 : 943279.

[15] Theodorou, Theodoros, Iosif Mporas, and Nikos Fakotakis, "Automatic sound classification of radio broadcast news," International Journal of Signal Processing, Image Processing and Pattern Recognition 5.1, 2012 : 37-47.

[16] Chen, Lei, Sule Gunduz, and M. Tamer Ozsu, "Mixed type audio classification with support vector machine," Multimedia and Expo, 2006 IEEE International Conference on. IEEE, 2006.

[17] Richard, Gaël, Mathieu Ramona, and Slim Essid, "Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams," Acoustics, Speech and Signal Processing, 2007, ICASSP 2007, IEEE International Conference on. Vol. 2. IEEE, 2007.

[18] Lavner, Yizhar, and Dima Ruinskiy, "A decision-tree-based algorithm for speech/music classification and segmentation," EURASIP Journal on Audio, Speech, and Music Processing2009, 2009 2.

[19] Lu, Lie, Hong-Jiang Zhang, and Stan Z. Li., "Content-based audio classification and segmentation by using support vector machines," Multimedia systems 8.6, 2003 : 482-492.

[20] Liang, Bai, et al,, "Feature analysis and extraction for audio automatic classification," Systems, Man and Cybernetics, 2005 IEEE International Conference on. Vol. 1. IEEE, 2005.

[21] Lyon, R. J., et al, "A study on classification in imbalanced and partially-labelled data streams," Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013.

[22] Shanahan, James G., Norbert Roma, and David A. Evans, "Method and apparatus for adjusting the model threshold of a support vector machine for text classification and filtering," U.S. Patent No. 7,356,187, 8 Apr, 2008.

[23] Bocklet, Tobias, et al, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," ICASSP, 2008.

[24] Moreno, Pedro J., Purdy P. Ho, and Nuno Vasconcelos, "A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications," Advances in neural information processing systems, 2004.

[25] El-Maleh, Khaled, et al, "Speech/music discrimination for multimedia applications," icassp. IEEE, 2000.

[26] Wang, W. Q., W. Gao, and D. W. Ying, "A fast and robust speech/music discrimination approach," Information, Communications and Signal Processig, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on. Vol. 3. IEEE, 2003.

[27] Bahatti, Lhoucine, et al, "An Efficient Audio Classification Approach Based on Support Vector Machines," INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 7.5 , 2016: 205-211.