Available online www.jsaer.com

Journal of Scientific and Engineering Research, 2016, 3(4):417-422

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Automatic Indexing of Multimedia Documents by Neural Networks

Dabbabi Turkia[1], Lamia Bouafif[2], Ellouze Noureddine[2]

[1]Signal and information processing Laboratory, Science Faculty of Tunis, UTM
[2]Image and Signal Processing Laboratory, ENIT, University of Tunis Manar

Abstract In this paper, we will present a new sound indexing of multimedia documents in order to accelerate the audio documents retrieval in the net or their research in digital libraries. The main difficulty of this operation is the parameterization, modelling and discrimination of the sound track and its transcription. To resolve this problem, we have developed a hybrid method constituted by four procedures: parameterization, training, modelling and classification. The first step is based on the extraction of new temporal and spectral features and descriptors, however the others procedures are associated with MMG and RN classifier. The implementation of our algorithm is integrated in an embedded Matlab platform. The system performances are evaluated on a database constituted of music songs with speech segments under several noisy environments.

Keywords audio indexing, classification, HMM, ANN, Speech-music discrimination.

Introduction

The important development of the Internet and digital database with several multimedia documents requires new smart tools for structuring and indexing these data in order to reduce the time and to enhance the classification ratio. The automatic indexing aims to extract from the digital stream several descriptors to access to the information by its content. For musical signals, we extract descriptors allowing deducing the original partition, the kind of the song, the signature or "Audio summary" of the audio document. This operation may consist, for example, to search for a particular document in a collection from a description of another document; to navigate in a collection to browse through its contents, to summarize the collection, to automatically describe the documents for the archives, to use these descriptions to produce new documents or new services (radio and TV emission TV, films etc.).

In the last century, the first one to be interested in this problem is Moorer in 1975 [1]. His objective was the automatic transcription of sounds, but the developed procedure was manually and not automatic. Later in 2004, Pinquier [2] developed an indexing multimedia system applied on a TV broadcasting emission. This work uses GMM and SVM methods. In 2010 Rahona [3] and Durrieu [4] published two other studies on audio classification, music discrimination and transcription based on new descriptors and training. However, the performances of these techniques vary considerably between a database to another and cannot be verified nor recommended.

Indexing methods

In this framework, several approaches of indexing and structuring the audio tape of audiovisual documents are proposed. Their goals are to detect the primary components such as speech and music.

Speech aspects

For the speech signal, the information source is consisting of a periodic signal (glottis pulses characterized by the pitch Fo) and a noisy signal representing the unvoiced speech like fricatives. The vocal tract is an acoustic cavity equivalent to a RII filter. Its function is to transform the glottis signal by resonance phenomena to a formantic timbre characterized by formants F1, F2, F3. Figures 1 and 2 give an illustration of the temporal and the time-frequency evolution of speech and music signals. However, the traditional music is characterized by a harmonic structure, stationary, repetitive rhythm, and a lack of silence [5]. For example, in figure 2 , we observe that the music track have a harmonic aspects in the spectrogram illustration..

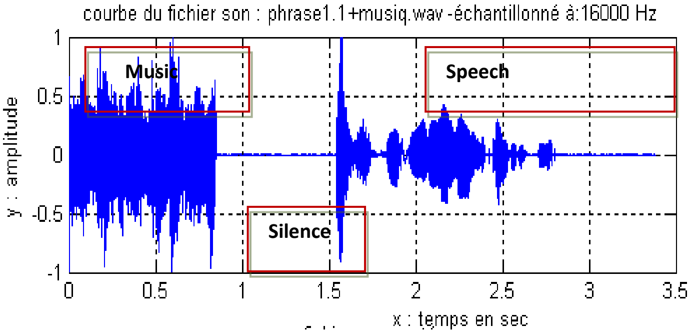

Figure 1: A speech and music signal pronounced by a male speaker

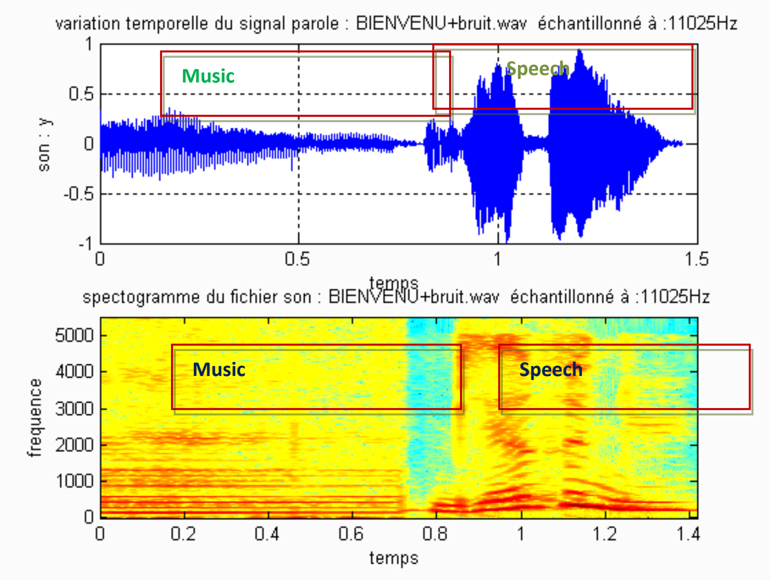

Figure 2: Spectrogram of an audio file (speech + music)

Audio descriptors

The descriptors of audio signals can be classified into four categories:

- Spectral descriptors (centroid, spectral flat, MPEG7, Energy bands,…)
- Temporal descriptors (ZCR, statistical moments, Autocorrelation coefficients,…)
- Cepstral descriptors (MFCC)
- Perceptual descriptors (PLP, pitch, Jitter, Shimmer)

Classification Based on GIM method: “Gaussian Incremental Modeling”

The objective of this work is to propose an approach of classification sound that can be easily and automatically adapted to the multimedia content and application. The proposed approach is based on a short-term model based on a Gaussian incremental modeling procedure which uses psychoacoustics features [6]. The GIM model uses a neural network classifier applied on four classic problems of sound classification: the classification in music/speech, man/woman, action/ non-action and finally the recognition of the kind of music to generate the appropriate format. Two applications have been developed in this project. The first one concerns the structuring of a video in scenes sound in order to facilitate the search and navigation. The second one is the implementation of a CYNDI [6] sound indexer and its architecture is described in figure 3. The CYNDI indexer consists of the following modules:

1. A module of demultiplexing which separates the audible signal of an audiovisual program to the entry;
2. A module of segmentation by speech or no-speech, man or woman, with a delay of 15 seconds;
3. A module of Indexing/segmentation speech/music;
4. A module for indexing the music;
5. A module for the generation of MPEG7 file.

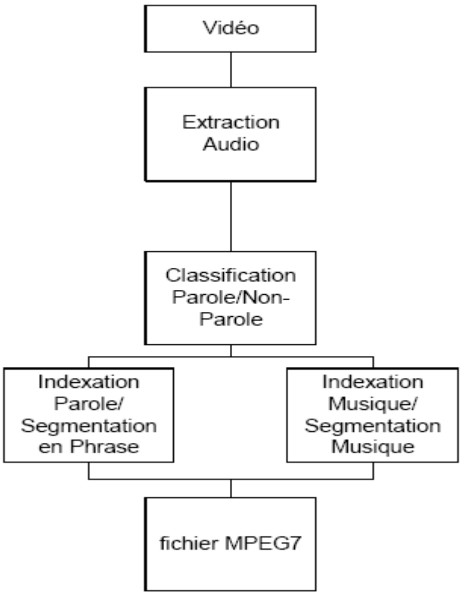
Vidéo
Extraction Audio
Classification Parole/Non-Parole
Indexation Parole/ Segmentation en Phrase
Indexation Musique/ Segmentation Musique
fichier MPEG7

Figure 3: CYNDI Architecture

Classification by HMM: Hidden Markov Models

This technique is very useful for indexing speakers in audio documents. The indexing in speakers is to determine the number of speakers, the time of their interventions and the discrimination of them. No information on speakers is available: nor their number or their identity, nor any sample of their voice.

The first task is to segment each document independently of each other. The second task is to identify the speakers appearing in several documents using the information contained in the segmentations produced by the first task. This task is to build an "index" or a codebook. The key of the index will be the speaker identifier in the multimedia document [7]. To solve this problems, we used the HMM method. The States of the HMM represent the speakers of the document the transitions between these states model changes in speaker. The process of segmentation is iterative: the states of the HMM, are added one by one to each iteration. In particular, the methods of acoustic parameterization of the signal, the models of Speakers by model multi-Gaussian (GMM) and the learning methods have been adapted to the task. This statistical framework allows keeping acceptable maximum likelihood values during the different stages of the segmentation [8].

Audio Indexing by artificial neural network ANN

We have developed an appropriate module based on ANN classifier. This interface is divided into three procedures parameterizations, training, and indexing-classification [9]. The structure of the multilayer perceptron (MLP-ANN) is illustrated by figure 4.

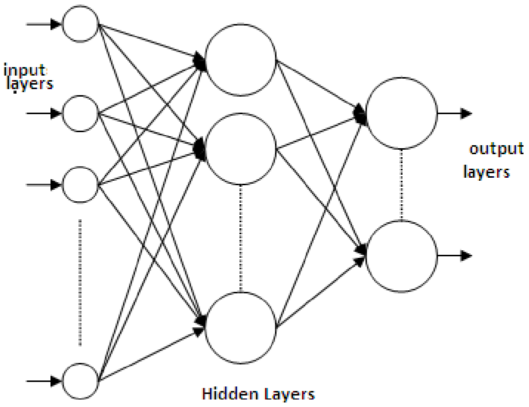

Figure 4: multilayer perceptron architecture

The number of neurons belonging to a layer has been fixed as follows:

- For the input layers, the number of neurons is equal to the total number of cepstral parameters (MFCC).
- For the output layers, the number of neurons is equal to the number of signals index.

- For Hidden layers, we start with a small number and then gradually we increase the number of neurons until we obtain the optimized recognition ratio and a minimal quadratic error.

- The activation function has the shape of an exponential sigmoid.

Learning phase

During this phase, the aim is to compute the weight functions and then to establish the model. The weights are adjusted in order to discriminate the training data by using gradient retro spread algorithm. The process is repeated for all examples of the codebook until we obtain a negligible output error [10]. Figure 5 demonstrates that with 13 epochs we obtain an output error of 10^{-5}.

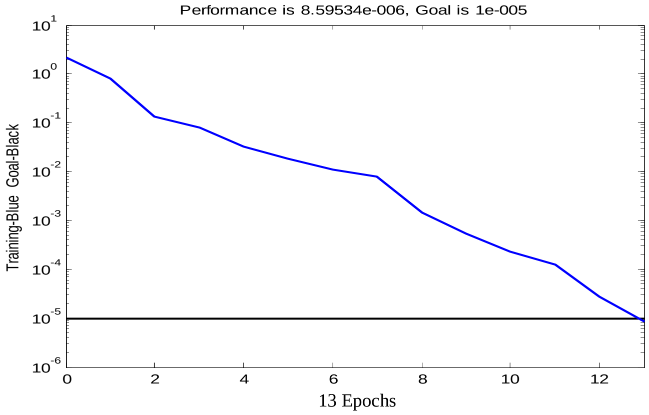

Figure 5: The mean quadratic error between actual and desired outputs.

Simulation and Experiments

We have implemented under Matlab a GUI interface that allows us to choose different parameters such as: the size of the segmentation frame, the recovery ratio, the number of neurons in the hidden layer, the number of iterations and the number of input and output layers.

In order to optimize our proper architecture, we conducted a series of programming and tests by varying the number of neurons of the hidden layer, the number of training and the number of iterations. This will allow us to reduce the computing and training times and to increase the indexing ratio.

The effects of number of neuron in the hidden layer on the performance of the system

Table 1 and figure 6 summarize the influence of the number of neuron in the hidden layer on the indexing error computed on our system.

Table 1 : Evaluation of the indexing error and the training time

Number of neurons	20	30	50	60	70	80	90	100	110	120
Training time(s)	134	164	180	192	226	245	284	334	369	384
Indexing Error (%)			21	19	17	16	14	12	12	12

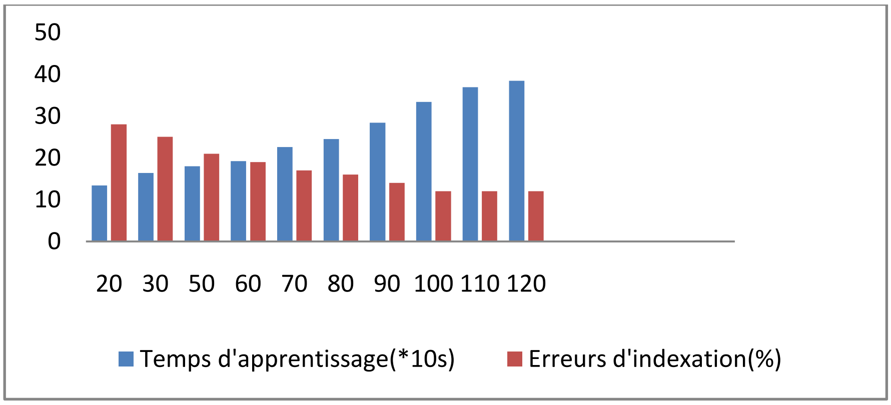

Figure 6: Evolution of the indexing error and the training time vs the number of neuron in the hidden layer

Figure 6 shows that the indexing error decreases when we increase the number of neurons from 20 to 100, and then remains constant. Therefore we have selected the value of 100 neurons as the number of the optimal neuron in the hidden layer.

Study of the influence of the variation of the Training codebook size

The table 2 summarizes the effect of the variation of training parameter η on the indexing error for music.

Table 2: Error of indexing and the training time in function of the training

No training (η)	0.001	0.005	0.007	0.009	0.01	0.02	0.04
Indexing error (%)	45	38	31	23	17	12	13
Training time (s)	167	168	171	166	169	170	164

Training (η)	0.05	0.09	0.1	0.5	0.6	0.7
Indexing error (%)	12	12	16	18	20	22
Training time (s)	166	164	165	166	172	165

According to figure 7, we note that the error of the indexation of our system decreases each time that η increases up to wait the optimal value (η =0 ,04) then increases when η exceeds this value.

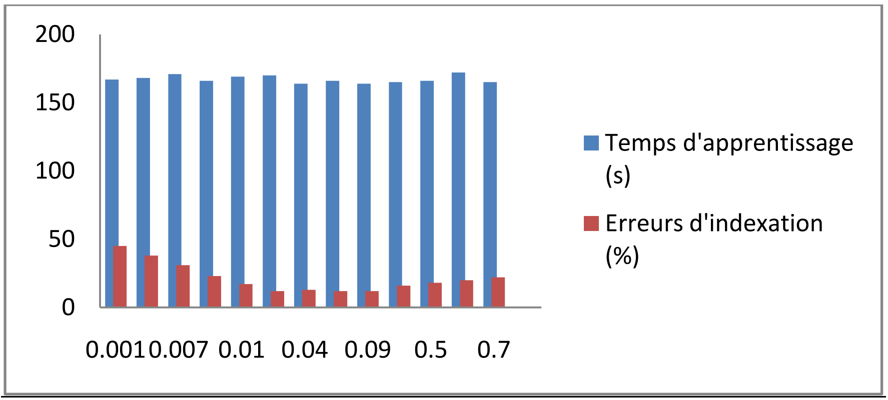

Figure 7: Evolution of the indexing errors vs the training (η)

Study of the variation of the number of iterations in function of the retro-propagation

We vary the number of iterations of the back-propagation algorithm, and each time it performs the indexing test to our codebook. Table 3 and figure 8 illustrate the results of these tests.

Table 3: Indexing error and training time vs the iterations number (N)

Number of iterations	100	150	200	300	500	1000	1500	2000	3000
The learning time(s)	21	27	33	55	68	122	178	235	390
Errors of indexing (%)	29	26	22	18	16	11	11	11	11

According to the figure 8, we note that 1000 iterations (cycles of learning) are required for a good convergence of the network with an indexing error of 11% .

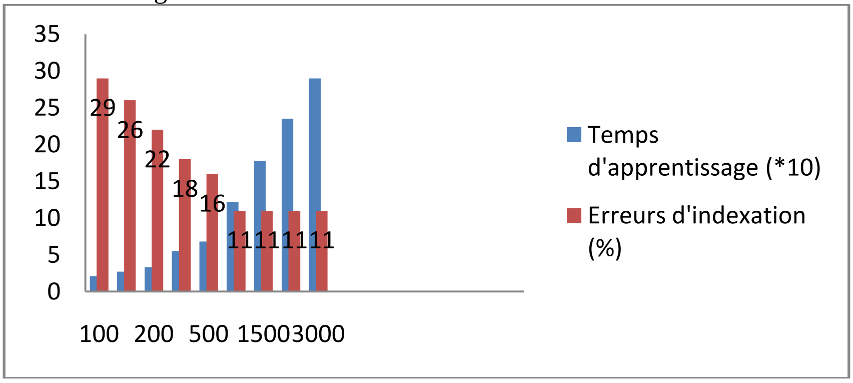

Figure 8: Evolution of the indexing vs the number of iterations

Optimal results

The table 4 summarizes the optimal settings. The tests carried out on the speed of convergence and the error for the indexing of the network, allowed us to next optimized parameters:

- The value of the no programming = 0.01
- The number of iterations = 1000
- The number of hidden layers = 100 neurons

- The minimal error of indexation = 7 %
- Training time = 136 seconds.

Table 4: Optimal Settings

Parameters	**Optimal value**
Number of neurons in the hidden layer (**N)**	100
No learning (η)	0.04
Number of iterations (N**)**	1000
Indexing error (%)	7

Conclusion

During this study, we presented and implemented an automatic audio indexing system based on ANN classification which is intended for the automatic multimedia research documents. The developed system presents a parameterization module (feature extraction: MFCC, …), a learning tool which is based on the algorithm of the retro-spread gradient, and an indexation tool based on ANN.

However, the choice of the ANN architecture requires to be optimized. In fact, we cannot choose random values of the number of neurons of the hidden layer, or those of training and iterations. Only the experience allows us to answer this question. In this study, our measurements and simulations results allowed us, not only to choice the optimal parameters, but also to follow the effect of their variations on the system and indexing performances.

The audio database is a set of speech and songs. The experiments and tests conducted on our platform, showed that we have enhanced the indexing ratio to 93% (with an error of 7%).

References

[1]. S. Rossignol (2000), Segmentation and indexing of sound signals musical, *PHD Thesis*, IRCAM-University of Paris VI, France.

[2]. Pinquier. I. (2004). Indexing sound: search for the primary components for an Audiovisual structuring" *PHD Thesis,* IRIT, University of Toulouse, France.

[3]. Rahona M (2010). Classification automatique de flux radiophoniques par SVM, *PHD Thesis*, Telecom Paris-Tech, France.

[4]. Durrieu. J.L (2010). Transcription et séparation automatique de la mélodie principale dans les signaux de musique, *PHD Thesis*, Telecom Paris-Tech, France.

[5]. Geoffroy P (2013). Indexation automatique de contenus audio-musicaux, *HDR à l'Université de Paris VI*, France.

[6]. Harb H (2001). Classification of the audible signal with a view to a indexing by the content of multimedia documents, *PHD Thesis, Liris*, Central School of Lyon, France.

[7]. Ezzaidi H (2002). Discrimination speech-music and study of new settings and models for a system for the identification of the Speaker in the context of telephone conferences, *Doctoral thesis*, University of Quebec in Chicoutimi, Canada.

[8]. Read L, H. Jiang, and H. Zhang (2001). A robust audio classification and Segmentation Method. *ACM International Conference on Multimedia*, pages 203-211. Ottawa, Canada.

[9]. Scheirer E. and M. Slaney (1997). Construction and evaluation of a robust Multi-feature Speech/Music discriminator. *IEEE International Conference on audio, speech and signal processing*, pp 1331-1334. Munich, Germany.

[10]. Auder.J and Mallat S (2011). Multi-scale scattering for audio classification. *In proceedings of ISMR2011*, pp 657-662. Florida, USA.