Speech-Music-Noise Discrimination in Sound Indexing of Multimedia Documents

Lamia Bouafif[1], Noureddine Ellouze[2]
[1] National Institute of Biomedical Studies of Tunis, 1092, Tunis, Tunisia
[2] Image and Signal Processing Laboratory, ENIT BP 37, University of Tunis El Manar, 1064, Tunisia

Sound indexing and segmentation of digital documents especially in the internet and digital libraries are very useful to simplify and to accelerate the multimedia document retrieval. We can imagine that we can extract multimedia files not only by keywords but also by speech semantic contents. The main difficulty of this operation is the parameterization and modelling of the sound track and the discrimination of the speech, music and noise segments. In this paper, we will present a Speech/Music/Noise indexing interface designed for audio discrimination in multimedia documents. The program uses a statistical method based on ANN and HMM classifiers. After pre-emphasis and segmentation, the audio segments are analysed by the cepstral acoustic analysis method. The developed system was evaluated on a database constituted of music songs with Arabic speech segments under several noisy environments.

Keywords: Speech processing, audio indexing, training and recognition.

1 Introduction

Currently, the high volume of audiovisual materials needs selection and automatic classification of the multimedia contents. This can be achieved by speech processing and HTML-XML languages where one can extract a movie sound track not only by keywords but also by a sound identification. In order to facilitate the access to the recording audio databases, the multimedia documents is segmented then transcribed and annotated in function of its contents. This step requires at first a good discrimination between speech, music and noise and further discrimination between speakers.

Several works on this domain was conducted such as the NIST (USA) project of Rich Transcription[1]. It contains 78 hours of English meeting speech, reference transcripts and other materials. Rich Transcription is broadly defined as a fusion of speech-to-text and metadata extraction technologies providing the bases for the generation of more usable transcriptions of human-human speech in meetings. In 2004, another interesting work was done by Pinquier[2] and Aguilar in 2004[3] at the Research Informatics laboratory of Toulouse. A classification and modeling approach was carried out using Gaussian Mixture Model and Support Vector machines. Their evaluation was conducted on a TV broadcasting emission. In 2008, Mirrezaie & Ahadi developed a new speaker diarization technique based on genetic algorithms. Both segmentation and indexing experiments were carried out using the PSO, GA and DISTBIC algorithms[4].

Pinquier (2003) used the time-frequency features such as spectral gravity and energy for feature extraction part and used GMM as a classifier. In the same year, Harb and Chen, stated that the MFCC could be used also in audio classification as a strong feature. Later in 2012, Moattar & Al[5] developed new speaker diarization systems based on evolutionary computation algorithm and a new TLBO algorithm in speaker clustering stage for speaker diarization of Broadcast News.

In fact, the sound indexing is complex because of the correlation between speech, noise, music and other components. Many references and works used several methods to resolve this problem. There are two kinds of discrimination and indexing techniques: The first one is a parametrical method which is based on features extraction of the audio file and its comparison with standard values. However, the second technique is statistical and is based on modeling, training and recognition such as Neural networks NN, GMM, HMM and SVM. The implementation of these methods has relieved the following problems:

- Automatic segmentation of the soundtrack in Speech-Music-Noise context
- The choice of suitable discriminative method (parametric modeling or a statistical approach)
- Indexing by speakers to detect the speakers present in a document or a collection of documents and annotate their content following these speakers.

To resolve these problems, the first step is the Speech-Music-Noise segmentation and the development of the indexing engine. It is based on automatic segmentation of audio in music and/or speech followed a new segmentation into “sentences” on the identified speech zones, allowing easy navigation in an audiovisual document and rapid retrieval. The segmentation technique of sentences is based on statistics of the normal size of a sentence. Detecting phrases edges is based on automatic thresholding of the Kullback-Leibler distance values[6].

2 Theory and methodology

2.1 Speech aspects

The speech signal is characterized by its variability in amplitude and phase, its stochastic content and its non-stationary behaviour. It is the result of a convolution between a phonation (glottis source) and an articulation (filter: Vocal tract). The source is characterized by the pitch F_0, yet he vocal tract is characterized by a formantic structure which reflects the resonance of the vocal tract given by the formants (F_1,F_2,F_3,..). However, the traditional music is characterized by a harmonic and stationary structure, a repetitive rate/rhythm and an absence of silence[7]. The band-width of the musical production is extended until the higher limit of the hearing response (20 kHz), whereas for the speech word the limits are around 8 kHz[8]. For example, Figures 1 to 4 represent an illustration of the waveform, spectrogram, pitch and formants parameters of speech and music signals.

Figures 1 and 2 represent the parameters of a female speech signal “bientot.wav” sampled at 11025 Hz. The mean pitch frequency is about 245 Hz with null values between 0.25 seconds and 0.4 seconds indicating a silence zone. The wide band spectrogram of figure 1 shows the formantic character of the speech illustrated by the red curves.

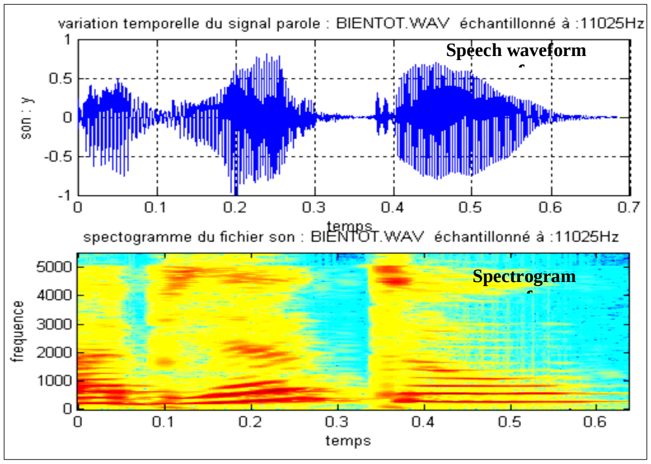


Figure 1. Speech waveform & spectrogram.

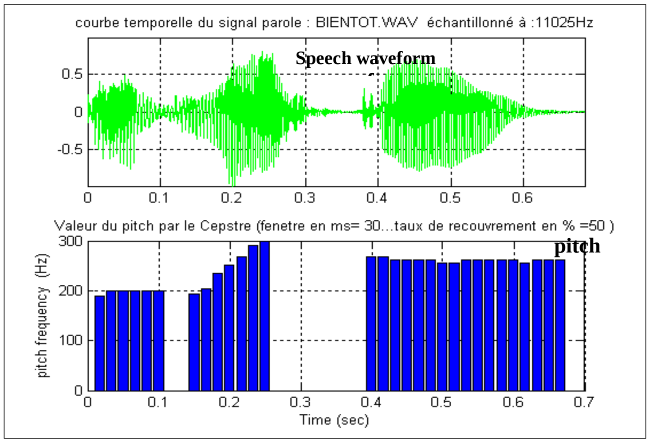


Figure 2. Pitch evolution of a female speech "bientot.wav" sampled at Fs=11025 Hz.

2.2 Music aspects

Figures 3 and 4 represent a music waveform: We can observe a continuous harmonic curve (F=152 Hz). Finally, because of the ambiguity of the pitch in mixture audio files[9], we have mixed in Figure 5, music (from 0 to 0.8 seconds) and a speech sound (0.85 seconds to 1.4 seconds). We can observe either in Figures 5 or 6 a clear discrimination between speech and music segments by pitch curves values or by formants deduced and computed from WB spectrogram.

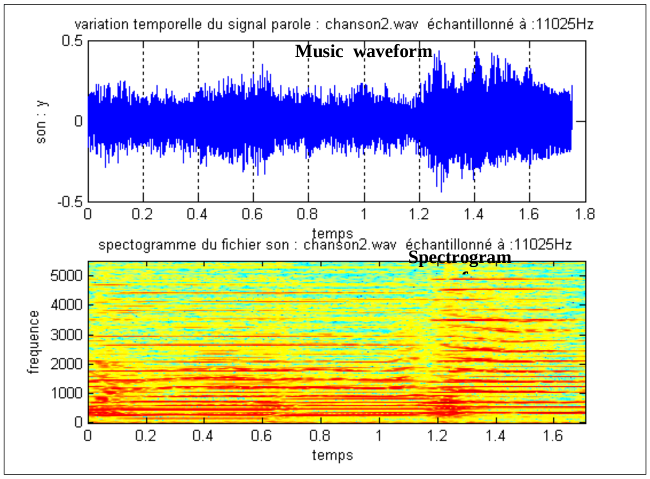


Figure 3. Music waveform & spectrogram.

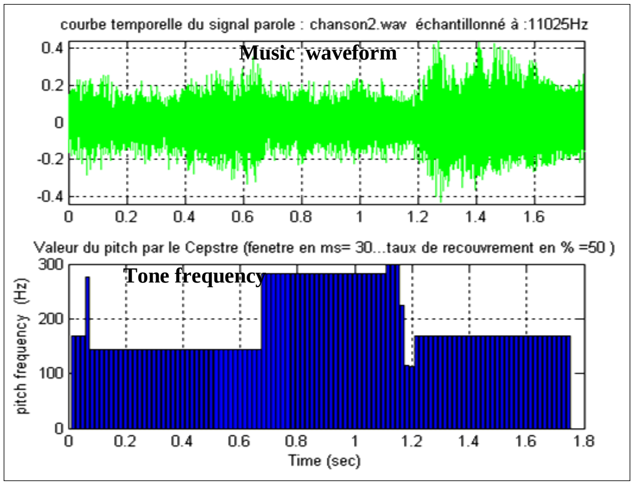


Figure 4. Tone frequency evolution of a music "chanson.wav" sampled at Fs=11025 Hz.

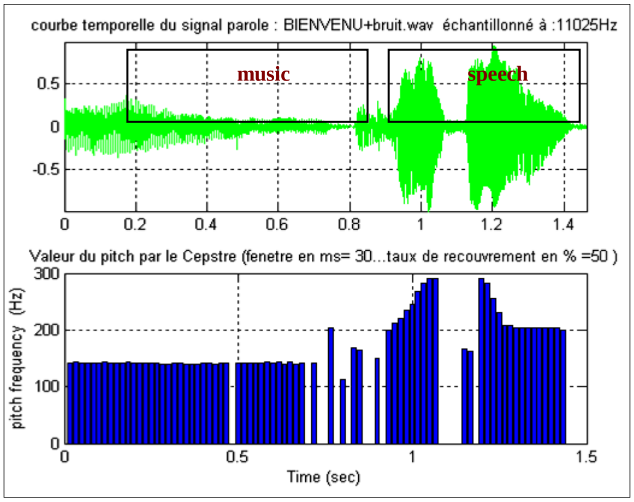


Figure 5. Waveform & pitch of a Mixed Speech & Music.

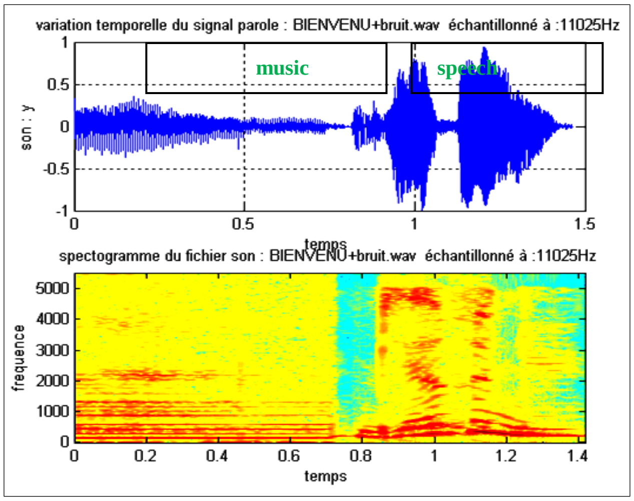


Figure 6. Waveform & spectrogram of a Mixed Speech & Music.

The previous results are simulated by our developed GUI interface under Matlab environment In order to verify and validate these analysis, we have applied the same files and data on an another speech processing tool which is the famous "Speech Analyzer" software developed by SIL Int (USA). Figures 7 and 8 give illustrations of the evolution and variations of speech and music parameters. The results demonstrate that the speech tracks are characterized by a continuous raw pitch curve Fo, formantic curves (F1=409 Hz, F2=1324 Hz, F3= 1962 Hz, F4=2439 Hz, …). However, the music signal does not have any formants and has a disturbed pitch curve, but is characterized by notes frequencies and harmonics which are presented by red colors in the spectrogram figure.

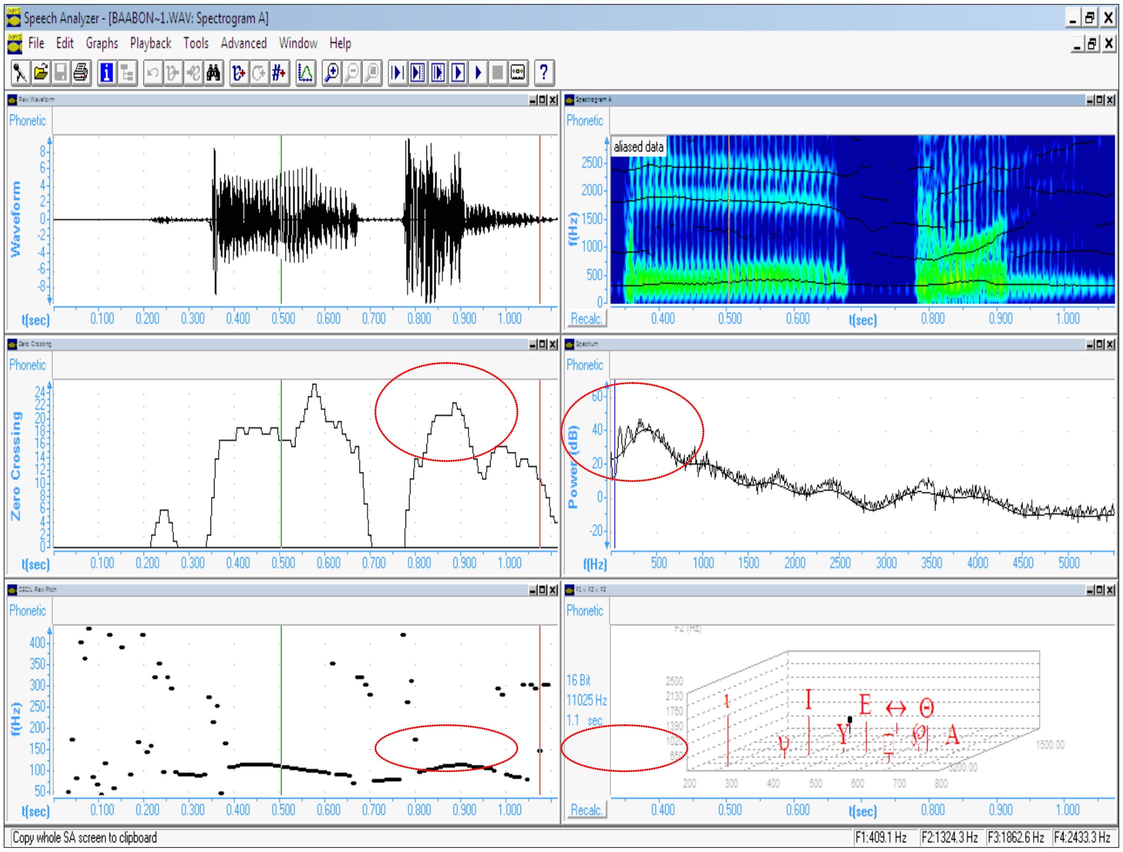


Figure 7. Waveform, zero-crossing, pitch, spectrogram, spectrum and formants of a Speech file.

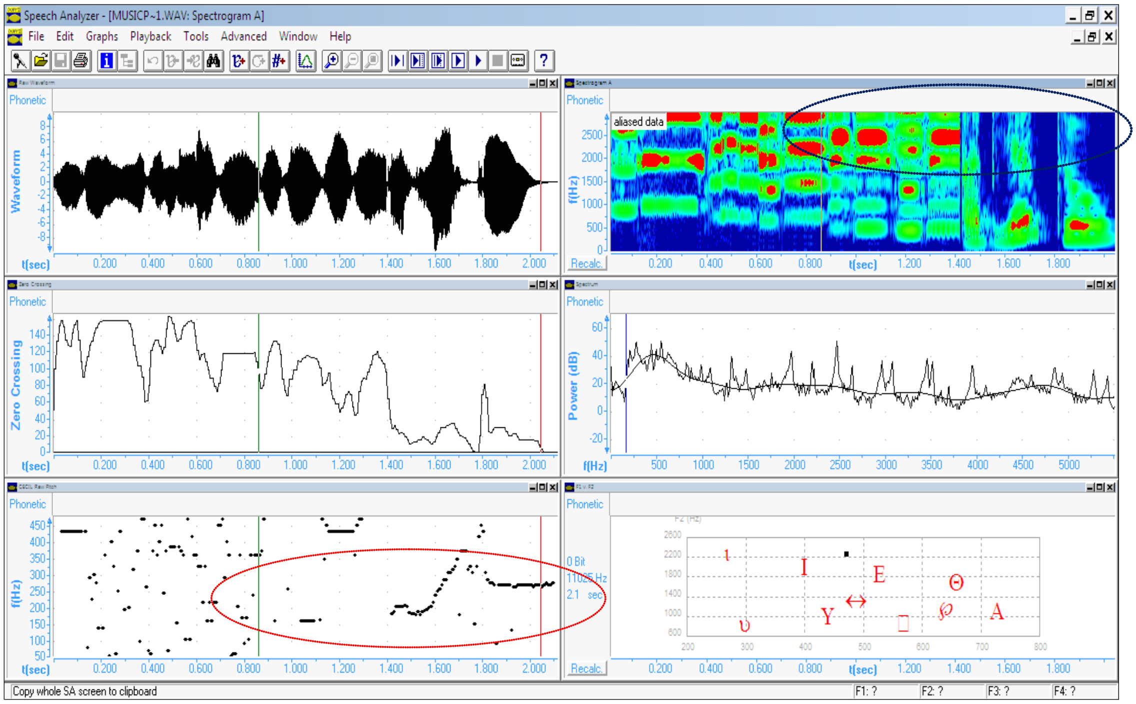

Speech Analyzer - [MUSICP~1.WAV: Spectrogram A]
File Edit Graphs Playback Tools Advanced Window Help
Phonetic
Waveform
t(sec)
aliased data
f(Hz)
Recalc.
Zero Crossing
Power (dB)
0 Bit
11025 Hz
2.1 sec
Copy whole SA screen to clipboard
F1: ?
F2: ?
F3: ?
F4: ?

Figure 8. Waveform, zero-crossing, pitch, spectrogram, spectrum and formants of a Music file.

2.3 Classification algorithms

The audio indexing algorithm is given by Figure 9. After a first step of Pre-Emphasis and speech segmentation with a sliding hamming window of 10 ms, the program calculates the speech activity with a VAD test (Vocal Activity Detector). This step is necessary to detect silence from speech[10]. The second step is the parameters extraction such as MFCC, ZCR, SF and SC. Finally, a classification operation is applied by using a training-recognition procedure (ANN, MMC or SVM) in order to discriminate S-M-N and to identify the speaker segments.

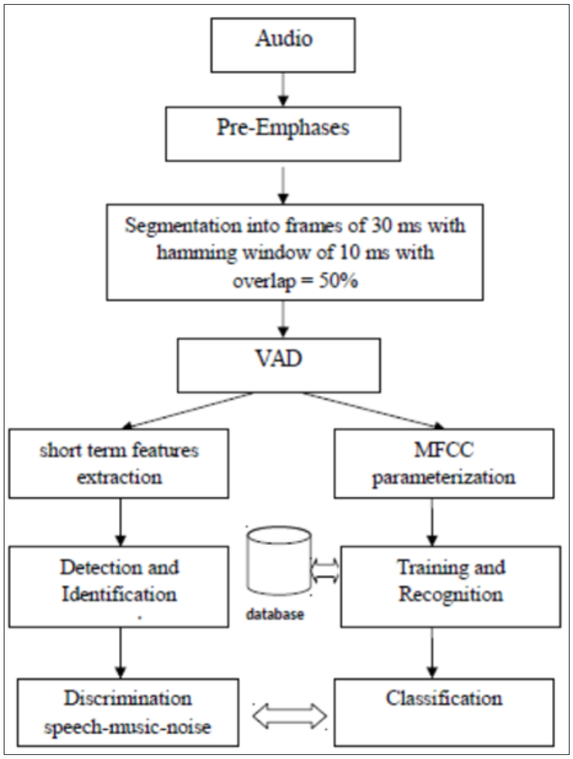

Audio
Pre-Emphases
Segmentation into frames of 30 ms with hamming window of 10 ms with overlap = 50%
VAD
short term features extraction
MFCC parameterization
Detection and Identification
database
Training and Recognition
Discrimination speech-music-noise
Classification

Figure 9. Audio indexing algorithm.

2.4 Audio parameters and featutes

Several parameters can be used to characterize the audio signals and thus to reduce the quantity of information to be treated. These parameters can be temporal, spectral or statistical such as the short time energy, zero-crossing, spectral flux, spectral centroids and entropy. For a speech signal, we can add others parameters such as the pitch, the formants, and MFCC coefficients.

MFCC coefficients

MFCC is the most used for speech feature extraction and parameterization. The MFCC algorithm which is represented by Figure 10 can be expressed as[11]:

$$MFCC(i)=\sqrt{\frac{2}{N}}.\sum_{k=1}^{N}\log(E_k).\cos[\frac{\pi.i}{N}(k-\frac{1}{2})] \qquad (1)$$

Where: N is the number of band pass filters and Ek is the band pass filter output energy.

The number of MFCC coefficients is equal to 14. We added to this vector, the 1st derivative Δ and the 2nd $\Delta\Delta$ in order to follow the dynamic variations of the audio features.

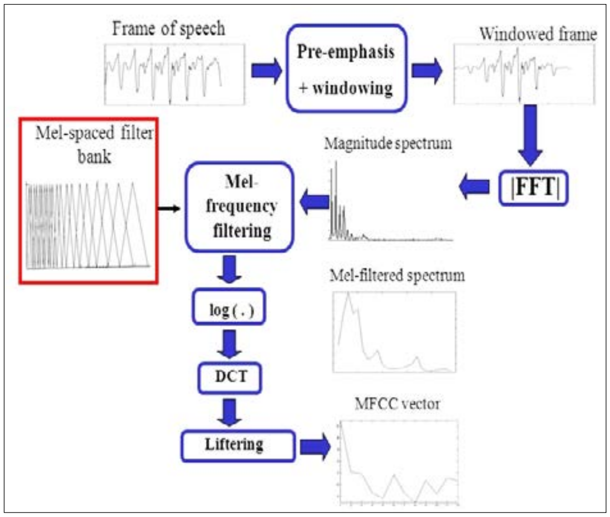


Figure 10. Feature extraction: MFCC algorithm.

Zero-crossing ratio: ZCR

The ZCR gives an idea on the signal dynamics and vocal activity. For noisy signals or unvoiced speech, ZCR values have increased contrarily for voiced speech sounds. However for silent, ZCR is around zero. This parameter can be expressed as[8]:

$$z(n)=\frac{1}{2N}\left(\sum_{i=1}^{N}\left|sign(x_n(i))-sign(x_n(i-1))\right|\right) \qquad (2)$$

In practise, ZCR can be used for Unvoiced/Voiced speech separation, or Music-noise, Music-unvoiced speech segmentation, but it is ineffective neither for a speech-music classification nor for Music-Music segmentation (such as Rock and Jazz). This result is confirmed in Figures 11 and 12. In fact, the higher values of ZCR (from 0.38 seconds to 0.42 seconds of Figure 11) indicate the presence of noise or unvoiced speech segments. However, the low values signify the localisation of a voiced speech sound or a silence (from 0.1 seconds to 0.37 seconds and from 0.42 seconds to 0.6 seconds of Figure 11).

In this case, the short time energy can separate the silence from the Voiced speech frames. In music, it is more difficult to deduce the speech from a noise or speech because the values are very closer and similar, that is why this parameter must be associated with other acoustic parameters.

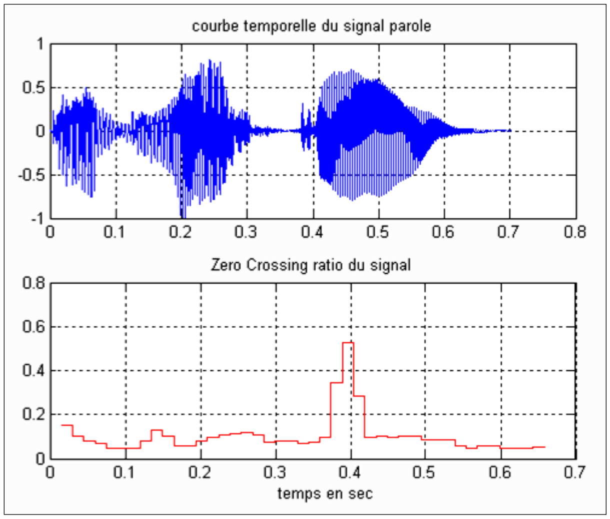


Figure 11. ZCR of a speech signal.

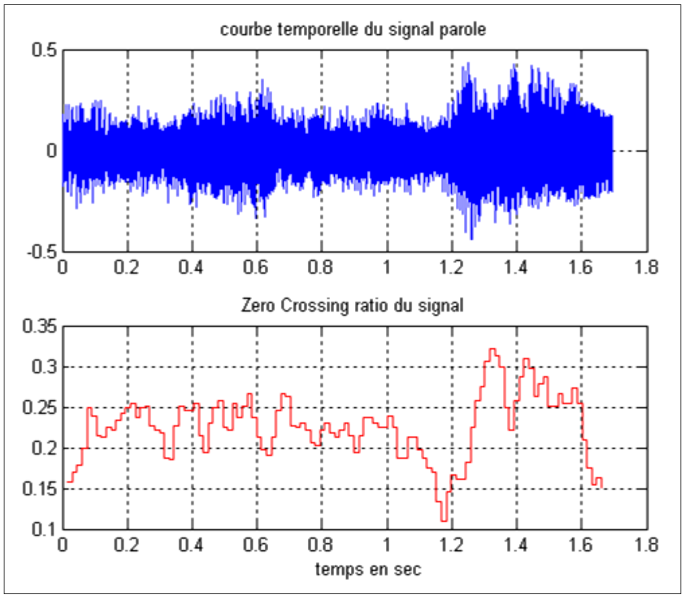


Figure 12. ZCR of a Music signal.

Entropy

The application of the concept of the entropy for the problems of detection of speech or music is based on the fact that the spectrum is better organized for S-M segments than for noise segments. The entropy indicates the quantity of information in a message. The more it is raised, the more there is information. Therefore, compression tries to transmit a number of bits slightly higher than the entropy if it is compression without loss of information. If not, the entropy of the signal will be decreased.

For a signal X={$x_{1,}$ ....,x_N } with a probability distribution {$p_{1,}$ ....,p_N }, we define the entropy H as[12]:

$$\mathrm{H}\,[X] = -\sum_{i=1}^{N} p_i \ln p_i. \tag{3}$$

We can observe in Figure 13 that for speech, we obtain high values of entropy (about 0.92 between 0.4 seconds and 1.3 seconds), however for music of Figure 14, the values are variable.

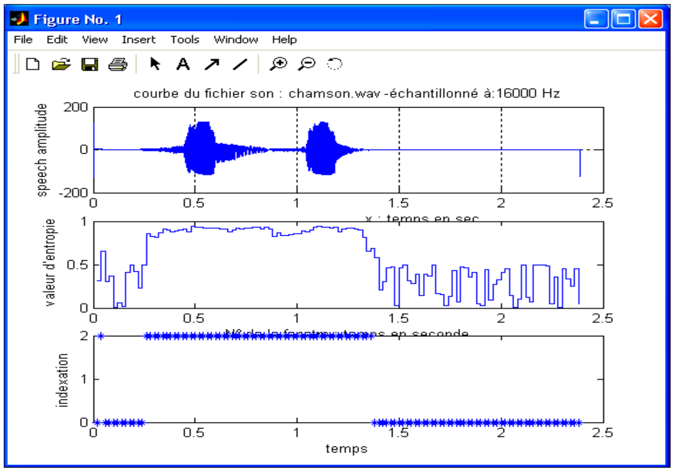


Figure 13. Entropy evolution of a speech signal.

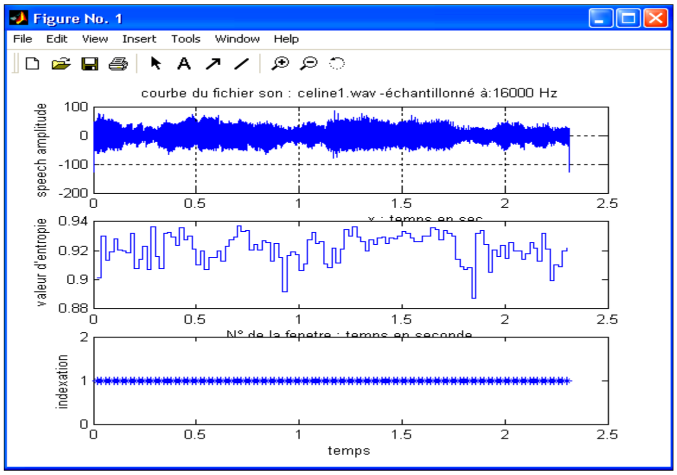


Figure 14. Entropy of a Music signal.

Spectral centroid

The spectral centroid is the frequential gravity centre of a DSP (Power Spectral density). It is expressed by (4) as[7]:

$$C(i) = \frac{\sum_{n=1}^{N} w_n . S_i(w_n)}{\sum_{n=1}^{N} S_i(w_n)} \tag{4}$$

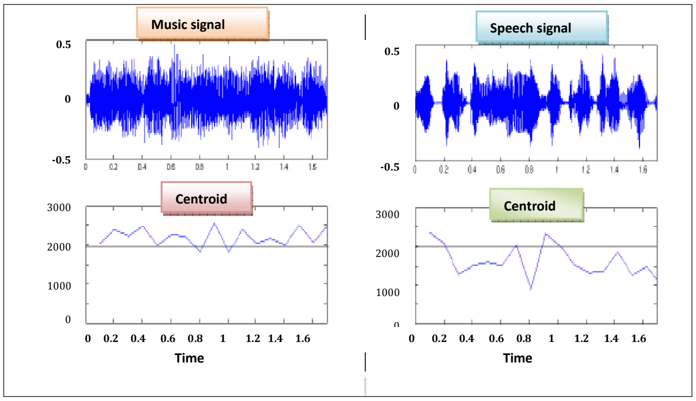


Figure 15. Spectral Centroid of Music and speech.

$S_i(w_n)$ corresponds to the spectral component of ith frame at the frequency (w_n) and *N* is the number of samples by frame. It is higher for the music than for speech sounds. In general 6 octaves are necessary to describe the music and 3 octaves are enough for the speech. The variation of spectral centroid is also significant: an important variation characterizes Voiced/Unvoiced voiced alternation. The centroid of Figure 15 for a music signal is high to a threshold of 2000 contrarily for a speech sound.

Spectral Flatness

VAD can be computed by using short term features such as Spectral Flatness (SF) and Short-term Energy[13].

Spectral Flatness feature is calculated using the following equation:

SF in db =10 Log_{10} (Gm/Am) (5)

Where (Am) and (Gm) are respectively, arithmetic and geometric means of the speech spectrum.

The Long-term spectral flatness measure LSFM feature is computed using the spectra of the last R frames of the input signal x(n).

$$L_x(m) = \sum_k \log_{10} \frac{\mathrm{GM}(m, \omega_k)}{\mathrm{AM}(m, \omega_k)}, \qquad (6)$$

$$GM(m, w_k) = \sqrt{\prod_{n-m-R+1}^{m} S(n, w_k)} \qquad (7)$$

$$AM(m, w_k) = \frac{1}{R} \sum_{n-m-R+1}^{m} S(n, w_k) \qquad (8)$$

3 Training and recognition

The algorithm contains two mains stages: The first one is the training of the audio sequences by using Baum-Walch and K-means algorithms, and the second stage is the recognition and the classification decision by using Viterbi decoding algorithm. Four classification techniques are used in audio indexing tasks[14]:

- Gaussian Mixture Models (GMM)
- Support Vector Machines (SVM)
- Artificial Neural networks (ANN)
- Hidden markov Models (HMM).

Figure 16 represents the principle of the training-recognition-classification procedure. The training step uses a database or a codebook constituted of audio parameters, yet the recognition stage uses feature extraction and HMM or GMM modeling in order to code every segment. Finally, the classifier uses the codebook and the real time model in order to identify the audio track by using viterbi decoder.

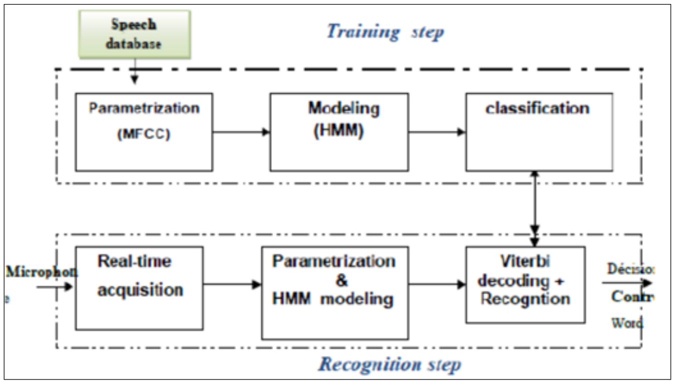
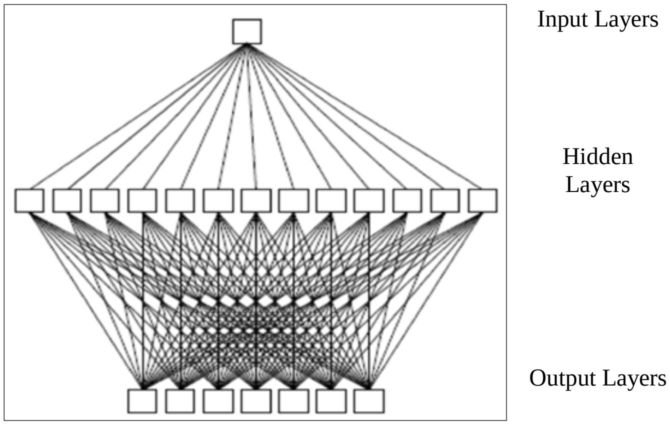


Figure 16. Training-recognition and classifier method.

3.1 ANN classifier

The Artificial Neural Network can be divided into input layers, hidden layers and output layers. The input layers are constituted of MFCC, H, C and SF as defined in the last paragraph.



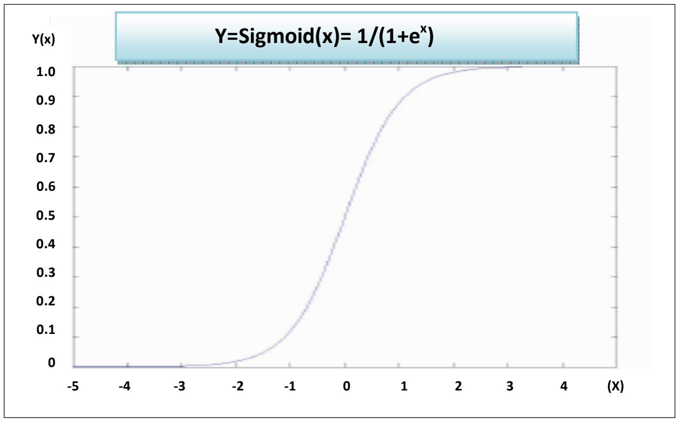


Figure 17. ANN classifier.

The architecture of the proposed neural networks is composed of three layers, 24 input layers, a hidden layer containing 220 neurons (sigmoid is applied as activation function) and 3 output layers containing the discrimination S, P and N.

3.2 Support vector machines

SVM is a binary classification method of supervised training. It was introduced by Vapnik in 1995. This method is based on the existence of a linear classifier called hyperplane which separates two classes. SVM are initially planned to solve classification problems in two classes, but there are now multi-class generalizations.

The original vector space is transformed into a Hilbert space with a function called kernel K. The hyperplane is computed using training vectors. The optimal hyperplane is orthogonal to the shortest line connecting the convex hull of the two classes, and intersects it half-way[15]. For two classes of sample data, the purpose of SVM is to find a classifier that will separate the data and maximize the distance between these two classes. The classifier is a linear classifier called hyperplane whose function is given by the following equation

$f(x) = b + w.\,x$ (9)

The closest data to the hyper plane are called support vectors.

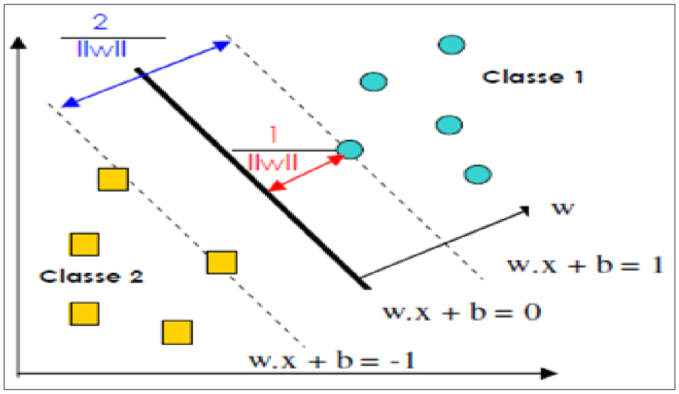


Figure 18. SVM classifier principle.

The distance between data and the hyper plan HP is:

$$D = \frac{f(z)/\|w\|}{1/\|w\|} = f(z) \qquad (10)$$

Where the HP equation is:

f(z) = (∑ai yi (xi z)) +b) (11)

Hence, the distance from centroïd to HP class is D':

$$D' = \frac{f(z)}{f(\bar{X})} D / f(x) \qquad (12)$$

with:

(x) is the centroid of trained data relative to class C

(Pc) is the ratio between the number of recognized segments of C class and the number of the segments in the same class.

Finally, the probability of the SVM output is defined as:

$$P(c/z) = \frac{P_c}{1+\exp(1-D')} \qquad (13)$$

3.3 HMM

Hidden Markov Models are useful for data statistical modeling and classification. The implementation of a system of recognition based on hidden Markov models HMM, has three phases:

-Describe a network whose topology reflects the sentences, vocabulary words or basic units

-Make the training mode settings: λ = (π, A, B)

-Carry out the actual recognition occurrence by calculating the maximum likelihood[16].

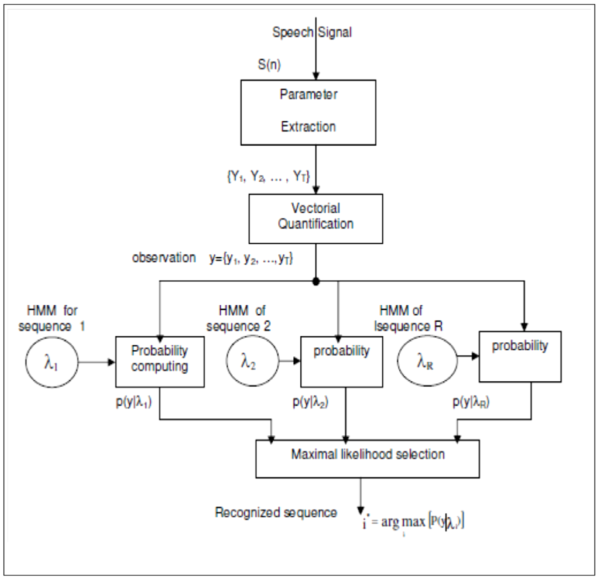


Figure 19. HMM classifier algorithm.

4 Simulation results

4.1 Audio database

The performances of a S-M-N segmentation depends of the selected database which is constituted of segments from audio programs containing music songs with Arabic speech under several noisy environments ARDB2006. Indeed, in applications like the automatic transcription of multimedia documents, TV video program transcription, it is necessary to avoid activating a recognition system “with great vocabulary” on a portion of signal corresponding only to music or songs. In the same way, it is important to locate the alternation of segments of speech, music or silence because this corresponds to a certain structure of the audio document.

4.2 Automatic segmentation and speech-music discrimination

Figure 20 gives an example of an automatic speech indexing. In this example the acquired audio signal is a speech sound. The computing of short time energy and ZCR has allowed us to detect the voiced speech zones (index=1 in Figure 20) and the unvoiced or the silence segments (index=0 in Figure 20).

Yet, in Figure 21, we have mixed music of portion of the singer “Abdelhalim Hafidh” with a speech segment of unknown speakers. The simulation results have given a good and accurate indexing and discrimination between the two segments. This result is very interesting because it can dissociate Music from speech independent of the speaker.

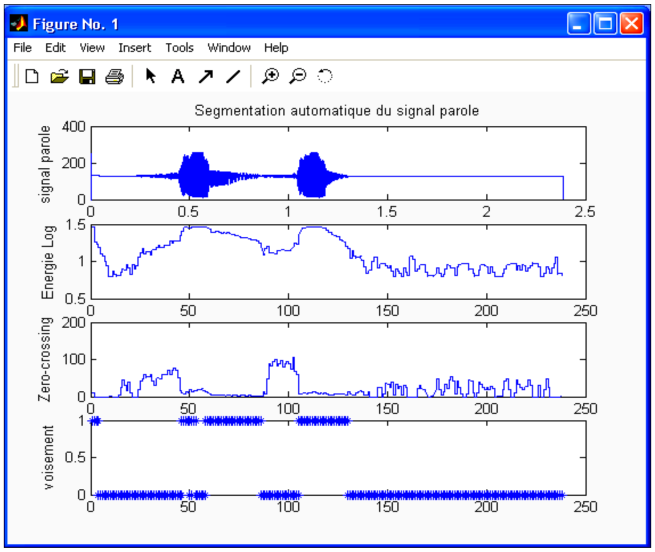


Figure 20. Automatic segmentation of an audio signal.

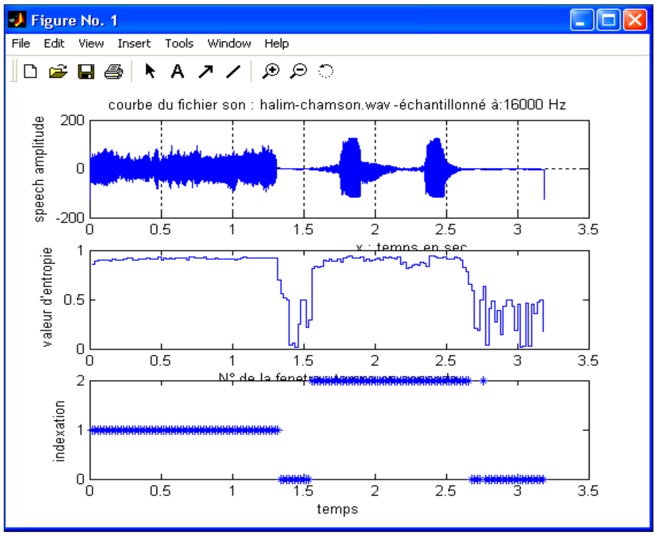


Figure 21. Speech-Music discrimination.

For another example, in the case of Figure 22, we used the entropy variance VH as an additional discrimination parameter. In this case VH=2% for the music segments (from 0 to 1.3 seconds), yet for the speech zone (from 1.5 seconds to 2.6 seconds), VH=8%. The same conclusion can be observed in Figure 23. The lower part of the curve gives the indexing decision, where the Music segment is coded by 1 and the Speech segment by 2.

The interest of this method is its robustness to recording environment and the sound types of documents. A high variance VH reflects the presence of speech. The detection of silence is done on the basis of calculation of the entropy compared to a threshold. The results obtained by the development of our interface for various audio signals are as follows:

-The automatic indexing value “2” of Figures 22 and 23 indicates speech zones,

-the value ”1” corresponds to the music,

-the “0” value indicates the silence.

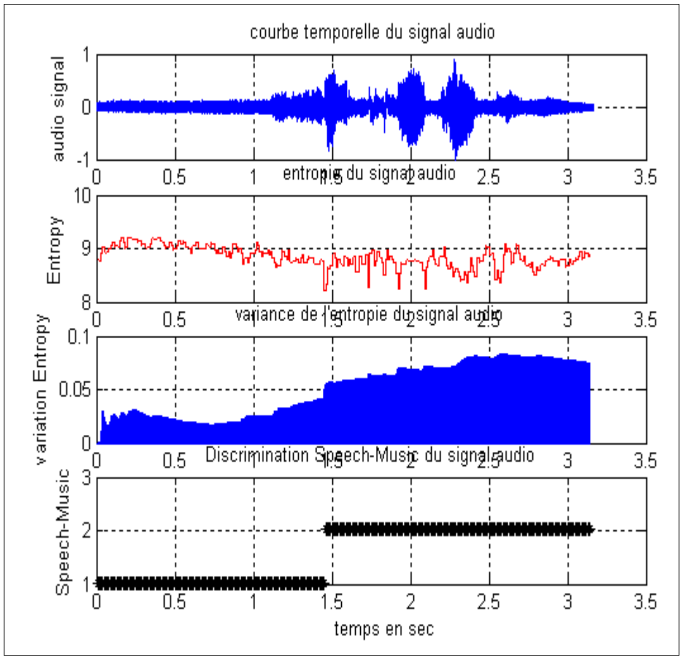


Figure 22. Speech-Music discrimination of S+M.

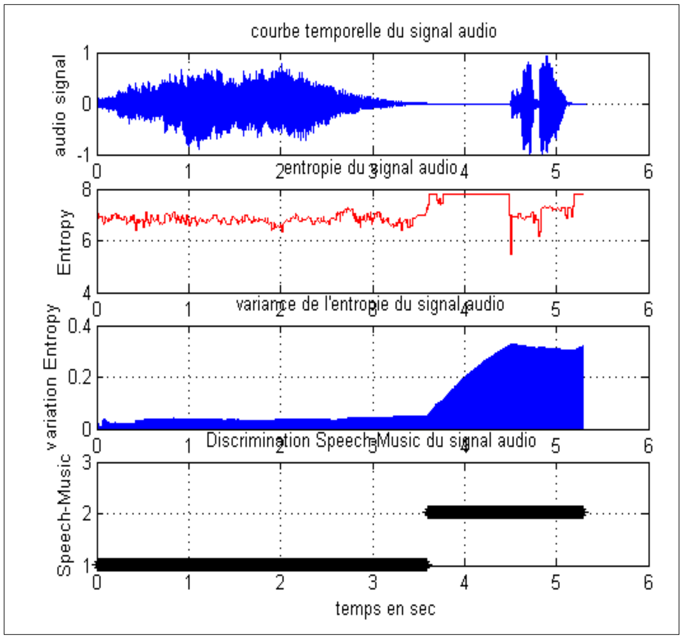


Figure 23. Speech-Music discrimination of S+M.

4.3 Training-recognition results

We have applied the features extraction with the algorithm of figure 16 for HMM, SVM and ANN classifiers. The results of Tables 1 and 2 indicate the obtained values of classification ratio. In fact, the classification rate is a validation factor for discrimination results and is defined by the ratio between the samples (frames) appropriately identified by the classifier in relation to the total number of samples. For this, a comparison is made for each audio processed file between the actual real content of the audio clip and the segmented and indexed (if the frame is correctly indexed and classified it will take a value=1 if not 0). Thus, if we have in the file 1000 frames and we identify 650 correct frames (true classification S, M, N), then the recognition rate will be 650/1000=65%.

Table 1. Classification ratio CR (with as input: MFFC+additional statistical audio parameters H, C, SF, VH).

Classifier	Speech	Music	Noise	Others
HMM/GMM	99%	97%	99%	74%
SVM	96%	92%	97%	66%
ANN	98%	93%	98%	70%

The HMM/GMM method has given classification results a little better than the ANN ones because the parameterization and the feature extraction uses statistical parameters such as entropy which is better adapted to the HMM classifier whose model is itself statistical and relies on the calculation of probabilities.

Table 2. Classification ratio CR (with as input: Only MFFC).

Classifier	Speech	Music	Noise	Others
HMM/GMM	89%	83%	90%	55%

Reading Table 1 can lead to a misunderstanding: In fact, “noise” refers to the identified noise portions of the audio file and the recognition rate. This not only means that the system is noisy but our application sets up parts of this noise that can be applause or engine operation effects. This can be considered as a contribution of this work because sometimes these frames of noise can be main information in crime and investigation applications. For example, in Table 1, CR classification ratio (CR) in red 99% and blue are not the same; if not, there will be no interest of the work since with or without noise, we obtained the same result. But, they mean that speech frames are correctly identified at 99% and for the noise, they are correctly identified at 99%.

Table 2 shows that the good choice of the feature extraction step affects the classification accuracy. Indeed, the hybridization of the MFCC coefficients with statistical coefficients (entropy, centroid, Spectral Flatness,..), increases the recognition rate by 10%. Also, the obtained results show a slight superiority of the HMM/GMM classifier of the order of 5% to 8% compared to ANN and SVM, but this is valid only with large audio database.

4.4 Comparative study

Table 3 gives a comparison with previous studies essentially from 2000 to today. Most of the SMN (Speech-Music-Noise) discrimination results show rates ranging from 88% to 96% with SVM, ANN or HMM methods. The difference between them lies in the feature extraction part essentially based on the MFCC coefficients. We note that researches that associated with these (MFCC) other statistical or energetic parameters gave slightly better results of about 5%. Our hybrid method has this advantage and has resulted in overall classification rates of 98% even in a noisy context.

Table 3. S/M Classification results (CR) of others researches.

Research study	**Classification method**	**CR for Speech**	**CR for music**	**Global CR**
Ajmera, 2002[19]	HMM	92	97	95
Munoz, 2007[18]	ANN	95	96	95.4
Ruiz, 2010[17]	ANN	95	98	96.9
Lei & Hua, 2011[21]	SVM	95	94	94,6
Wei-Ho Tsai, 2014[22]	GMM	93.9	95	94.3
M.Velayatipour M.Mosleh[20],	SVM	94	88	90.5
M.Mosleh, 2014[20]	GA Genetic Algorithm	94	97,5	96,4
Lars Ericsson, 2015	MLER, PLDA	95	98	97.3
L. Bouafif, 2017	HMM/GMM	99	97	98.2

5 Conclusion

Automatic content analysis of an audio document is a process that extracts its semantic meaning. This involves segmenting the document into units of meaning, classification between predefined categories, indexing to make it available for search, navigation and eventually for future transcription**.**

In this study, we developed an automatic indexing system of audio signals and a speech-music discrimination interface. We succeeded to implement this procedure under Matlab environment. This system is based on audio parameters extraction and statistical estimation. This double approach makes possible to define the concept of decomposition of an audio signal in frames of speech, music, noise and silence. The idea is to follow the dynamics of the audio signal by a classical features (MFCC+Δ) and additional statistical parameters such as the entropy, the centroid, the spectral Flatness. In addition, we used automatic classifier HMM/GMM, SVM and ANN algorithm for accurate discrimination and indexing into, music, speech, noise, silence and others segments. The future step of this work is its implementation and evaluation on a hardware interface such as a DSP or FPGA platform.

References

1. Nguyen, T., Sun, H., Zhao, S., Khine, S., Tran, H. D., Ma, T. L. N., Ma, B., Chng, E. S., Li, H., “The speaker diarization systems,” Proceedings of the RT’09, NIST Rich Transcription Workshop, Florida, USA, Vol. 14, pp. 1740-1766, 2009.
2. Pinquier, J., Arias, A., André-Obrecht, R., “Audio classification by search of primary components,” International workshop on Image, Video and Audio Retrieval and Mining, Québec, Canada, October 2004.
3. AGUILAR, J., “Méthodes à vecteurs de support et Indexation sonore. IRIT Laboratory Research report document,” Institut de Recherche en Informatique de Toulouse, France, 2004
4. Mirrezaie, S. M., Ahadi, S. M., “Speaker Diarization in a Multi-Speaker Environment Using Particle Swarm Optimization and Mutual Information,” IEEE International Conference on Multimedia and Expo, ICME, 2008.
5. Moattar, M. H., Homayounpour, M. M., “A review on speaker diarization systems and approaches: Review Article,” Speech Communication, Vol. 54, No. 10, pp. 1065-1103, 2012.
6. Harb, H., Chen, L. M., Yves, J., “Segmentation du son en se basant sur la distance de KulBack-Leibler,” pp. 63-68 CORESA’01, Dijon, 2001.
7. Ezzaidi, H., “Discrimination parole/musique et étude de modèles pour un système d’identification du locuteur dans le contexte de conférences téléphoniques,” PHD, Quebec, 2002.
8. Boite, R., “Traitement de la parole,” Presses Polytechniques Romandes, Collection, 1999.
9. Harb, H., “Classification du signal sonore en vue d’une indexation par le contenu des documents multimédias,” PHD Thesis, Ecole Doctorale, Lyon, France, 2001.
10. Ma, Y. N., Akinori, N., “Efficient voice activity detection algorithm using long-term spectral flatness measure,” EURASIP Journal on Audio, Speech, and Music Processing, Vol. 21, 2013.
11. Wu, Z., Cao, Z., “Improved MFCC-based feature for robust speaker identification,” Tsinghua Science & Technology, Vol. 10, No. 2, pp. 158-161, 2005.
12. Seck, M., Bimbot, F., Zugaj, D., Delyon, B., “Two-class signal segmentation for speech/music detection in audio tracks,” *Proceedings of EUROSPEECH’99*, September 1999.
13. Scheirer, E., Slaney, M., “Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. ICASSP’97,” Munich, Vol. II, pp. 1331-1334, 1997.
14. Lu, L., Zhang, H. J., Jiang, H., “Content analysis for audio classification and segmentation,” IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 7, October 2002.
15. Anibal, J., Pinquier, J., André-Obrecht, R., “Evaluation of classification techniques for audio,” Proceedings of the 13th European IEEE Signal Processing Conference, Antalya, Spain, 2005.
16. Nagaraja, B. G., Jayanna, H. S., “Feature extraction and modelling techniques for multilingual speaker recognition: A review,” International Journal Signal and Imaging Systems Engineering, Vol. 9, No. 2, 2016.
17. Reyes, N. R., Vera Candeas, N., García Galán, S., Muñoz, J. E., “Two-stage cascaded classification approach based on genetic fuzzy learning for speech/music discrimination,” Engineering Applications of Artificial Intelligence 23, pp. 151-159, 2010.
18. Munoz-Exposito, Garcia-Gala, J. E., Ruiz-Reyes, S., N. Vera-Candeas, P., “Adaptive network-based fuzzy inference system vs. other classification algorithms for warped LPC-based speech/music discrimination,” Engineering Applications of Artificial Intelligence 20, Elzevier, pp. 783-793, 2007.
19. Ajmera, J., McCowan, I., Bourlard, H., “Robust HMM-based speech/music segmentation,” IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. I-297-I-300, 2002.
20. Masoumeh Velayatipour, A., Mohammad Mosleh, B., “A review on speech-music discrimination methods,” International Journal of Computer Science & Network Solutions February, Vol. 2, No. 2, 2014.
21. Lei, X., Fu, Z. H., Feng, W., Luo, Y., “Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news,” Multimedia Systems, pp. 101-112, 2011.
22. Tsai, W. H., Ma, C. H., “Speech and singing discrimination for audio data indexing. ” 2014 IEEE International Congress on Big Data, ”Anchorage, Alaska, USA, 2014.

The authors can be reached at: b2lamia@yahoo.fr.