An Optimization of Audio Classification and Segmentation using GASOM Algorithm

Dabbabi Karim, Cherif Adnen
Research Unity of Processing and Analysis of Electrical and Energetic Systems
Faculty of Sciences of Tunis, University Tunis El-Manar
2092 Tunis El-Manar, Tunis, Tunisia

Hajji Salah
School of Engineers of Tunis, 3000
University Tunis El-Manar, Tunis, Tunisia

***Abstract*—Now-a-days, multimedia content analysis occupies an important place in widely used applications. It may depend on audio segmentation which is one of the many other tools used in this area. In this paper, we present an optimized audio classification and segmentation algorithms that are used to segment a superimposed audio stream according to its content into 10 main audio types: speech, non-speech, silence, male speech, female speech, music, environmental sounds, and music genres, such as classic music, jazz, and electronic music. We have tested the KNN, SVM, and GASOM algorithms on two audio classification systems. In the first audio classification system, the audio stream is discriminated into speech no-speech, pure-speech/silence, male speech/female speech, and music/ environmental sounds. However, in the second audio classification system, the audio stream is segmented into music/speech, pure-speech/silence, male speech/female speech. For pure-speech/silence discrimination, it is performed in the two systems according to a rule-based classifier. Concerning the music segments in both systems, they are discriminated into different music genres using the decision tree as a classifier. Also, the first audio classification system has succeeded to achieve higher performances compared to the second one. Indeed, in the first system using the GASOM algorithm with leave-one-out validation technique, the average accuracy has reached 99.17% for the music/environmental sounds discrimination. Moreover, in both systems, the GASOM algorithm has always reached the best results of performances compared to KNN and SVM algorithms. Therefore, in the first system, the GASOM algorithm has been contributed to obtain an optimized consumption time compared to that one obtained using the two HMM and MLP methods.**

***Keywords*—Segmentation and classification audio; features extraction; features discrimination; GASOM algorithm**

I. INTRODUCTION

In order to facilitate and help the users to be more accurate and efficient in their research for multimedia contents on search engines, content-based indexing and retrieval technologies is a good way to help them to access directly to the required multimedia contents. Recently, the research in the multimedia content relies on the content-based audio retrieval and other relevant techniques such as the audio segmentation, the audio indexing, the audio browsing, and the audio annotation. Generally, there are many techniques to categorize the audio content into speech, music or other sounds, and there are different methods to process each type of them. Concerning the retrieve of speech and spoken documents, they are transformed into texts by automatic speech recognition systems. For the retrieve of music, an approximate string matching algorithm has been proposed in [1] to solve a string matching problem and to match strings of features, such as the rhythm, melody, and chord strings of musical objects in a music database. Also, besides speech and music, we can find general sounds that represent the major audio type. In some research, such sounds has been dedicated to the classification and in others, it has been used in more specific areas, such as the classification of piano [2] and ringing [3] sounds. Furthermore, in order to face the growing size of audio databases with a huge amount of audio data, an efficient organization and manipulation of data is required. For example, a discrimination of speech and non-speech segments with a high accuracy is required for such applications, such as the automatic transcription instance of broadcast news (BN), automatic speech and speaker recognition, recovery audio requests, and so forth. As the audio data contains alternating sections of different audio types, an automatic classification of its content into appropriate audio classes is a fundamental step in the processing of audio streams. Thus, this kind of separation is called audio content classification. Regarding the audio stream segmentation, it is often substantial with the classification process in the recovery system and they are together useful for many classification tasks. Moreover, the feature extraction process is a conditioning element for the overall classification performance as it includes three types of features which can be extracted from temporal, frequency, and coefficient domains. Concerning the time domain features, they include the Zero-Crossing Ratio (ZCR), the Silence Ratio (SR), the Root Mean Square (RMS), and so on. As for the frequency domain features, they contain the pitch, the bandwidth, the Spectral Centroid (SC), and so on. Also, the linear prediction coefficients (LPC) and the Mel-Frequency Cepstral Coefficients (MFCC) are widely exploited in automatic speech recognition and automatic classification of general sounds. Recently, the wavelet coefficients have attracted much attention of researchers thanks to its multi-resolution property and its better time-frequency resolution [4], [5]. Furthermore, a major change in the online service has been created by the excessive increase of multimedia data on the internet. Therefore, the audio information becomes an important part of most multimedia applications, especially music, which is the most common and popular example of online information. Thus, the segmentation and classification of audio streams according to their content is a useful means for analyzing

audio, video, and understanding content. However, performing this task requires an efficient and accurate technique. Such a technique is called audio segmentation which splits an audio stream into homogenous regions. Also, an emerging increase in digital data is caused by the advent of multimedia and network technology, which in turn begets a growing interest in multimedia content-based information retrieval. Indeed, the discrimination of audio signal according to its content is the fundamental step for its analysis and understanding. For audio segmentation and classification, it is considered as a pattern recognition problem and it includes two main stages: feature extraction and extracted-features-based classification [6]. Also, the categorization of audio content analysis applications can be performed in two parts: the first part is the discrimination of an audio stream into homogenous regions and the second part is the discrimination of a speech stream into segments of different speakers. In [7], [8], the discrimination of an audio stream into different audio types has been performed using Support Vector Machine (SVM) algorithm and K-Nearest Neighbor (KNN) algorithm. Moreover, the characterization of various audio content levels of a sound track has been carried out by frequency tracking in an audio indexing system proposed in [9]. This system has the specificity that it does not need any prior information. In [10], the authors have proposed a fuzzy approach that uses a hierarchical segmentation and classification according to automatic audio analysis. In [11], an extracted-features-based music and speech discrimination has been performed using a multi-dimensional Gaussian Maximum A posteriori (MAP) estimator, a Gaussian Mixture Model (GMM), a k-d tree-based spatial partitioning scheme, and a KNN classifier. Also, the change point detection is a process which splits the audio stream into homogenous and continuous temporal regions by searching for temporal boundaries. On the other hand, it has a problem which arises in the definition of homogeneity criteria. For this purpose, stream segmentation can be performed by calculating the Generalized Likelihood Ratio (GLR) statistics without prior knowledge of classes [12]. However, computing statistics using MFCC coefficients requires a large amount of data for training [12].

For a transcript of meetings and automatic camera tasks, the segmentation of the meeting of a group of persons according to their voices is required. Indeed, the segmentation of feature vectors has been carried out using Bayesian Information Criterion (BIC), which has required a large amount of data for training [13], [14]. Also, the Structures Support Vector Machine (SSVM) has been used by structured discriminator models for large-vocabulary speech recognition tasks and the determination of features has been performed by Hidden Markov Models (HMMs) [15], [16] and a Viterbi decoding [17]. The human auditory systems rely principally on perception, while audio retrieval systems are traditionally text-based, which is not sufficient to achieve perceptual similarity between two audio clips because it only elaborates the high-level audio content. Thus, a query technique has been used to solve this problem and it was a very different approach to audio classification. In [18], modeling of continuous probability distribution of audio characteristics has been performed by a Gaussian mixture model (GMM). Also, a MMI-supervised tree-based vector quantizer and a feed-forward neural network have been proposed in [15], [19], [20], [21] for the task of detecting speech and environmental sounds on a sound stream. Indeed, a Kernel Fisher discriminator-based regularized kernel has been used for an unsupervised change detection task [22], [23].

Speech is not only limited to be used as a mode of transmission words of messages, but it can be also used as a means of transmitting emotions, personality, etc. Indeed, in many speech applications, mainly in speech segmentation and speaker verification, words containing vowel regions have a vital importance. For this, dividing an audio stream into segments is possible by a vowel regions-based audio segmentation. In fact, the audio segmentation algorithms can be divided into three general categories: the first category includes the features extraction stage in which the time and frequency domain features are extracted, and then their classification is performed by a classifier in order to discriminate the different audio signals according to their content. For the second audio segmentation category, it includes the feature extraction statistics which are used for discrimination by a classifier. Thus, these types of features are called posterior probability-based features. In this category, the classifier requires a large amount of data for training in order to reach accurate results. Concerning the third category of audio segmentation algorithms, it requires the use of efficient discriminators, such as BIC, Gaussian Likelihood Ratio (GLR), and Hidden Markov Model (HMM). In fact, good results are given by these classifiers if a large amount of data for training is provided. Also, many applications have been performed using audio segmentation and classification. Among these applications we can find the content-based audio classification and retrieval which are most used in the entertainment industry, managing audio archives, use of commercial music, supervising, and so forth. Nowadays, millions of databases on the World Wide Web are presented for audio search and indexing, and for audio segmentation and classification. In the monitoring of broadcasts news programs, the audio classification has contributed to reach efficient and accurate navigation through the archives of broadcasts news. The analysis of superimposed speech is a complex problem, and consequently improved-performance systems are required. Also, the audio stream segmentation is a preprocessing step in many audio processing applications in which it has a significant impact on the speech recognition performance. For this, the proposed audio segmentation and classification algorithm must be optimized, efficient, and fast in order to be used in real-time multimedia applications. Indeed, the hybridization of Self-Organization-Map (SOM) algorithm with Genetic Algorithm (GA) (called GASOM algorithm) is such algorithm which meets these requirements. To deal with complex data characteristics, the GASOM algorithm allows avoiding weakness such as slow convergence time being always trapped in the local minima. Moreover, this algorithm requires less training data, and consequently a high accuracy and a reduced-consumption time can be achieved. Indeed, the weights of the SOM algorithm have been optimized using GA algorithm, which allows obtaining a better mapping quality of classification and labeling data. In this work, the input data in the first audio segmentation and classification system is segmented, and then classified into nine basic audio types: speech, silence, music, environmental sounds, speech male,

speech female, electronic music, classic music, and jazz music. Concerning the second audio segmentation and classification system, the input data is segmented, and then classified into eight basic audio types: speech, music, silence, speech male, speech female, electronic music, classic music, and jazz music. In this paper, we also exhibit possible solutions for classifying the audio stream using the two KNN and SVM classifiers. Furthermore, different descriptors have been proposed to face the audio variety and discriminate very well between the different audio types.

The remaining sections of this paper are organized as follows: in Section I, audio segmentation and classification steps, feature extraction process, classification approaches (KNN, SVM, and GASOM) are presented, and then discussed. In the next section, an exhibition of different evaluations used to assess the experimental tests. In last section, the experimental results are discussed.

II. Research Method

A. Pre-classification

At first, the audio signal has been segmented into 1-s frames by applying the growing-window technique with a sample rate of 16 KHz. Consequently, the DCT coefficients at each frame have been calculated by Fast Fourier transform (FFT).Indeed, these last steps form together the short-term Fourier transform (STFT) which is a category of short-term processing techniques. Thus, we have obtained a matrix of the STFT coefficients from which their magnitudes are calculated to form a resulting matrix that can be treated as an image. This image is called spectrogram of signal.

B. Audio Classification and Segmentation Step

A separated analysis of each widowed frame in the audio clip has been performed as a pre-classification step before the classification. After that, the normalized feature vectors have been extracted, and then the classification step has been performed by selecting one of the algorithms SVM, KNN, and GASOM. Concerning the classification of audio clip/frames into speech and non-speech segments, it has been performed using a SVM, KNN, or GASOM classifier. For the speech segments, they have been discriminated into silence and pure-speech segments according to a rule-based classifier as the speech signal contains mostly silence frames. After that, the pure-speech segments have been used by the SVM, KNN, or GASOM classifier in order to discriminate between male speech and female speech. Also, the SVM, KNN, or GASOM classifiers have been then used to classify the non-speech segments into musical and environmental sounds. At the end, music genre discrimination has been carried out by a decision tree using music segments. Fig. 1 illustrates the block diagram of the first proposed audio classification system. Indeed, the audio stream has been each time down sampled to 16000 KHz and the features {Zero-Crossing rate, short-time energy, spectrum flux, Mel-frequency cepstral coefficients, vector chroma, spectral centroid, harmonic ratio, energy of entropy, spectral energy, and periodicity analysis}have been extracted, and then classified. These features {Mel-frequency cepstral coefficients, spectral flux, zero-crossing rate, and short time energy} have been used by the selected classifier (KNN, SVM, or GASOM algorithm) to classify the audio stream into speech and non-speech segments. For the discrimination between silence and pure-speech segments, it has been performed by a rule-based classifier, and then the pure-speech segments have been discriminated into male speech or female speech using the KNN, SVM, or GASOM algorithm as a classifier and {harmonic ratio and frequency estimator} as features. Also, the discrimination of non-speech segments into music and environmental sounds has been performed by the KNN, SVM, or GASOM algorithm as a classifier and {spectrum flux and Mel-frequency cepstral coefficients} as features. Moreover, the features {the minimum of the sequence entropy values and the mean value of the spectral flux sequence} have been used by the decision tree as a classifier in order to discriminate between different musical genres.

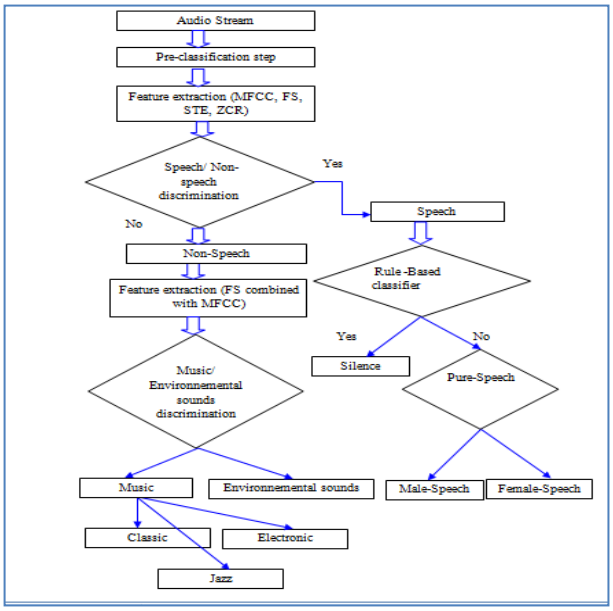

Fig. 1. Block scheme of the first audio classification and segmentation system.

C. Feature Extraction Step

At first, the audio signal has been divided into mid-term windows, and then the short-term processing technique has been applied for each segment. After that, the feature statistics have been calculated using feature sequences from each mid-term segment. Therefore, we obtain a set of statistics which represents each mid-term segment. In this work, the audio input has been divided into short-term windows and 23 audio features have been calculated per window. Thus, two mid-term statistics have been drawn per feature and a 46-dimensional vector has been obtained as output of the mid-term function. Also, the sizes of windows were 2 seconds and 0.05 seconds for mid-term and short-term processing, respectively. Moreover, the mid-term and short-term window steps were respectively set to 1 second and 0.025 seconds.

1) The Energy: The calculation of the short-term energy is given by the following expression:

$$E(i) = \sum_{n=1}^{W_L} |x_i(n)|^2, \; n = 1 \ldots W_L \qquad (1)$$

Where x_i and W_L are respectively the sequence of audio samples of the ith frames and the length of the frame. The normalization of the energy is usually performed in order to eliminate the dependence on the frame length. Thus, the expression of (1) becomes as follows:

$$E(i) = \frac{1}{W_L}\sum_{n=1}^{W_L}|x_i(n)|^2 \quad (2)$$

For the short-term energy variation, it is faster for speech frames than those of music because the speech signals contain weak phonemes and short periods of silence between words.

2) Zero Crossing-Rate (ZCR)

This feature is defined as a measure of the occurrences of signal changes from positive to negative or vice versa. Also, another more general definition is the amount of zero-crossing in the frame. Moreover, the ZCR feature is a good discriminator for a speech and music separation and it is higher for speech than to music as it contains more silent regions [24], [25]. Indeed, the ZCR feature is expressed as follows:

$$ZCR = \frac{1}{2(M-1)}\sum_{n=1}^{M-1}|sgn[x(n+1)] - sgn[x(n)]| \quad (3)$$

Where $x(n)$ and $sgn[.]$ represent respectively the discrete signal that is in the range of $n = 1, \dots, M$ and the sign function.

3) The Entropy of Energy

The interpretation of the measure of abrupt changes in the level-energy of an audio signal represents the short-term entropy of energy. Indeed, the calculation of this feature is carried out at first by dividing each short-term frame into k sub-frames of fixed duration. After that, the energy of each sub-frame j is calculated and divided by the total energy of the short-term frame ($E_{shortFrame}$)as in equation (1). Thus, the resulting sequence of sub-frame energy values e_j, j=1,…,K, is treated by a division operation (a standard procedure)as a sequence of probabilities such as in (4):

$$e_i = \frac{E_{subframe_j}}{E_{shortFrame_i}} \quad (4)$$

$$\text{Where} \qquad E_{shortFrame_i} = \sum_{k=1}^{K} E_{subframe_k} \quad (5)$$

At the end, the calculation of the entropy $H\,(i)$ of a sequence $e_i$is carried out according to the following equation:

$$H(i) = -\sum_{j=1}^{k} e_j\, log_2(e_i) \quad (6)$$

4) The Spectral Centroid and Spread:

The two simple measures of the spectral position and shape are carried out by the spectral centroid and the spectral spread. For the spectral centroid, it is defined as the center of 'gravity' of the spectrum. Indeed, the value of the spectral centroid C_i of the i^{th} audio frameis givenby the following expression:

$$C_i = \frac{\sum_{k=1}^{W_{f_L}} kX_i(k)}{\sum_{k=1}^{W_{f_L}} X_i(k)} \quad (7)$$

Concerning the second central moment of the spectrum, which is the spectral spread, it can be calculated by taking the derivation of the spectrum from the spectral centroid according to the following equation:

$$S_i = \sqrt{\frac{\sum_{k=1}^{W_{fL}}(k-C_i)^2 X_i(k)}{\sum_{k=1}^{W_{fL}} X_i(k)}} \quad (8)$$

5) The Spectral Entropy (SE)

The calculation of the spectral entropy is similar to that one of the entropy of energy with a difference that this latter is performed in the frequency domain [26]. Indeed, the spectrum of the short-term frame is at first divided into L sub-bands (bins), and then the energy E_f of the fth sub-band, $f = 0, \dots, , L-1$, is normalized by the total spectral energy, which is

$$n_f = \frac{E_f}{\sum_{f=0}^{L-1} E_f}, f = 0, \dots, L-1.$$

At the end, the entropy of the normalized spectral energy $n_f$is carried out according to the following equation:

$$H = -\sum_{f=0}^{L-1} n_f\,.\,log_2(n_f) \quad (9)$$

In [27], [28], an efficient discrimination between speech and music has been performed by the variant of the spectral entropy called chromatic entropy.

6) The Spectral Flux (SF)

The measure of the spectral change between two successive frames is performed by spectral flux which is calculated as the squared difference between the normalized magnitudes of the spectra of two successive short-term windows such as:

$$Fl_{(i,i-1)} = \sum_{k=1}^{W_{f_L}}(EN_i\,(k) - EN_{i-1}(k))^2 \quad (10)$$

$$\text{Where,} \qquad EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{W_{f_L}} X_i(l)} \quad (11)$$

$\mathrm{EN_i(k)}$ is defined as the kth normalized DTF coefficient at the ith frame.

7) The Spectral Rolloff

The frequency below which a certain percentage (usually around 90%) of the magnitude distribution of the spectrum is concentrated, is defined as a spectral rolloff. Each time that the $mth\ DFT$coefficient corresponds to the spectral rolloff of the ith frame, the expression satisfying this condition is given by the following equation:

$$\sum_{k=1}^{m} X_i(k) = C\sum_{k=1}^{W_{f_L}} X_i(k) \quad (12)$$

Where C is the adopted percentage. Also, the normalization of the spectral rolloff frequency is usually performed by dividing it with$\mathrm{w_{f_L}}$so that it takes values between 0 and 1.

8) MFCC Coefficients

This feature represents the cepstral representation of the signal where the distribution of frequency bands is carried out according to the Mel-scale instead of the linearly spaced approach. Let $\widetilde{O_k}$ the power at the output of the kth frame

filter, the resulting MFCC coefficients are expressed by the following equation:

$$c_m = \sum_{k=1}^{L} (log\ \widetilde{O_k}) \cos\left[m\left(k - \frac{1}{2}\right)\frac{\pi}{L}\right], m = 1, \dots, L. \quad (13)$$

Furthermore, the MFCC coefficients are defined according to (13) as the coefficients of the discrete cosine transform of Mel-scaled log-power spectrum. Also, the MFCC coefficients have been used in many audio analysis applications, such as speaker clustering [29], music genre classification [30], and speech recognition [31].

9) The Chroma Vector

The chroma vector is defined as the 12-element representation of the spectral energy [32]. Moreover, this descriptor has been widely applied in music-related applications [33]-[36]. Indeed, the computation of the chroma vector is performed by grouping the DFT coefficients of a short-term window into 12 bins: one of the 12 equal-tempered pitch classes of Western-type music is represented by each bin. Therefore, the mean of the log-magnitudes of respective DFT coefficients is produced by each bin such as:

$$v_k = \sum_{n \in S_k} \frac{X_i(n)}{N_k}, k \in 0, \dots, 11. \quad (14)$$

Where S_k and N_k represent respectively a subset of frequencies that correspond to the DFT coefficients and the cardinality of S_k.

10)Periodicity Estimation and Harmonic Ratio

In general, we can categorize the audio signals into a periodic (noise-like) and quasi-periodic. Despite the fact that some signals have a periodic behavior, it is so hard to find the same periods for two signals. Concerning the voiced signals and the majority of music signals, they are included in the category of quasi-periodic signals. For the estimation of the fundamental frequency, it is carried out according to the autocorrelation function, which calculates the correlation between the shifted signal and the original one [37]. After that the fundamental period which exhibits the maximum autocorrelation is chosen to be the lag. Indeed, the correlation $R_i(m)$ can be defined as the correlation of the i^{th} frame with itself at time-lag m such as:

$$R_i(m) = \sum_{n=1}^{W_L} x_i(n) x_i(n-m) \quad (15)$$

Therefore, the calculation of the normalized autocorrelation function for thei^{th} frame is given by the following equation:

$$\Gamma_i(m) = \frac{R_i(m)}{\sqrt{\sum_{n=1}^{W_L} x_{i(n)^2} \sum_{n=1}^{W_L} x_{i(n-m)^2}}} \quad (16)$$

Where W_L is the number of samples per frame and m is the time-lag.

Also, the harmonic ratio is defined as the maximum value of $\Gamma_i\ and\ it$ is determinate by the following equation:

$$HR_i = \max_{T_{min} \le m \le T_{max}} \{\Gamma_i(m)\} \quad (17)$$

Where T_{min} and T_{max} are the allowable values of the fundamental period.

Therefore, the position of the occurrence of the maximum value of Γ_i is used to determinate the selected fundamental frequency as follows:

$$\Gamma_0^i = \arg\max_{T_{min} \le m \le T_{max}} \{\Gamma_i(m)\} \quad (18)$$

III. Classification Approaches

We have designed two audio classification systems: in the first one, the SVM/KNN/GASOM classifiers are at first applied to classify segments into speech/non-speech segments, and then the non-speech segments are used for music/environmental sounds discrimination using the SVM, KNN or GASOM algorithm as a classifier. After that, the music segments are used by the decision tree classifier to discriminate between the different music genres. For the features of speech segments, they are discriminated by a rule-based classifier into pure-speech and silence, and then the SVM, KNN or GASOM algorithm, is also used to discriminate between the pure-speech segments into male speech and female speech. Concerning the second audio classification system, a speech and music discrimination is at first performed using the KNN, SVM or GASOM algorithm as a classifier, and then the music segments are classified into different music genres using the decision tree classifier. For the speech segments, they are used by a rule-based classifier to discriminate between the silence and pure-speech segments. After that, the pure-speech segments are used to discriminate between male speech and female speech using KNN, SVM or GASOM algorithm as a classifier.

A. Super Vector Machine (SVM) Algorithm

The learning of an optimized separation hyper plan for given positive and negative examples is performed by the Super Vector Machine (SVM) [38], [39]. Indeed, this classifier minimizes the probability of misclassifying unseen patterns for a fixed data that has an unknown probability distribution. Thus, the SVM allows obtaining an optimized performance on training data, and consequently the structural risks are minimized. In fact, this characteristic makes the difference between SVM and other traditional pattern recognition techniques in term of optimization. Also, we distingue two types of SVM: linear and kernel-based non-linear. The complication of the distribution of features in the audio data causes areas of overlap between the different classes and there is no possibility to separate them linearly. Such a situation can be manipulated by a kernel support vector machine. Moreover, the kernel has been used by SVM in order to create an optimal separation hyper plane [40], [41]. Indeed, the kernel function implicitly maps the input vectors to a high-dimensionality feature space in which they are linearly separable. Among the most well-known and used functions of kernel, we can mention: polynomial, function-based Gaussian radial, and a multilayer perception. In fact, the kernel-based Gaussian radial has empirically shown its high performance compared to other

types of kernel. For this, we have used it in our proposed models. Furthermore, the expression of the kernel-based Gaussian radial is given as follows:

$$k(x,y) = \exp\left(-\frac{x-y^2}{2\sigma^2}\right) \quad (19)$$

Where, σ is the width of the Gaussian function.

B. K-Nearest Neighbor (KNN) Algorithm

The KNN classifier is a non-parametric classifier which works as follows: for each input vector to be classified, a search is started in order to find the location of the k nearest training examples, and then the class which has the largest members in this location is assigned to the input. Indeed, the measure of the neighborhood is performed using the Euclidian distance. Also, the domination of certain features due to their range of values during the calculation of the Euclidian distance, requires the use of the linear method (20) as a remedy of this issue by normalizing the jth feature, $j = 1, \dots, L$, to zero mean and the standard deviation to 1:

$$\widehat{\gamma_i} = \frac{\gamma_i(j) - \mu(j)}{\sigma(j)}, i = 1, \dots, M, j = 1, \dots, L. \quad (20)$$

Where, $\mu(j)$is the mean value of the jth feature, $\sigma(j)$ is the respective standard deviation, L is the dimensionality of the feature space, and M is the number of training samples.

C. Self-Organized Mapping (SOM) Algorithm

The neural network map SOM was inspired from biology by Teuvo Kohonen. It is assimilated as many elementary processors represented by the neurons which are connected to each other in order to exchange information. In fact, the parallel and massive work of many formal neurons offers them the capacity for learning and deciding in the recognition task [42], [43]. In general the activation function is non-linear and it differs from an application to another. Moreover, the neural weights in the vicinity of the activated neuron (winner neuron) are updated by the learning rules, which make them close to the input vector:

$$\Delta w_i = \gamma h_{iv}(x(t) - w_i) \quad (21)$$

Where γ is the learning ratio and h_{iv} is the neighborhood function which relies on the distance between the units i and v on the map.

Furthermore, the map SOM network can be a universal tool of representation and recognition by virtue of its non-linear activation function. Thus, this algorithm can be applied in an unsupervised manner and it can be used for the recognition of voluminous input data.

D. GASOM Algorithm

To avoid the degradation of the diversity of genetic population in early generations, the SOM algorithm in order is used to maintain it thanks to its observed approximation property. Also, in order to increase the space research towards an optimal solution and avoid premature convergence, the Genetic Algorithm (GA) was hybridized to the SOM algorithm. This suggested algorithm allows the introduction of feature vectors into the SOM map in order to perform learning and testing operations. Indeed, there is an activation of a single neuron of the SOM map at each iteration, and consequently an appointing of the best matching unit (BMU). Among other neurons of the map, the best representative of the data inputs at this iteration is called the winning neuron. Also, every time we obtain a BMU neuron via the training iterations, which is special to each input and we will get an individual (a chromosome) assigned to this input for the reconstruction of population to be treated by the Genetic Algorithm (GA). Indeed, the representation of each chromosome is performed by a matrix of criteria which corresponds to the matrix of criteria for each neuron of a SOM map type during the iterations of learning or test [44]. After that the equation of changes and the update of the vectors of weights determinate the new chromosomes forming the new population for the next generation. Moreover, the modification of the update equation for the training of SOM map is performed by adding new coefficients according to the fitness values of the chromosomes of the current population. Furthermore, the ability of an input data is completely simulated by the weight of neuron as it is the largest organelle in the unit. Therefore, the diversification of population in the SOM topology has a huge effect on the evolution of the results of data recognition of the weights of units in the evolutionary process. Indeed, the explanatory diagram of the GASOM hybridization is shown in Fig. 2.

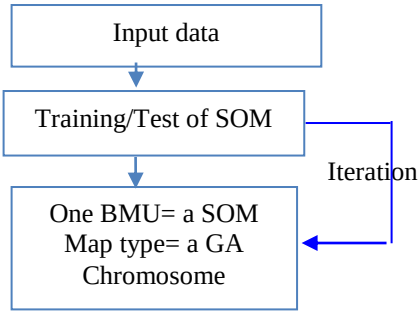

Fig. 2. Explanatory diagram of the GASOM hybridization.

E. Discrimination Steps for First Audio Classification System

1) Speech and Non-Speech Discrimination

This discrimination has been performed by the KNN, SVM or GASOM classifiers which have been applied with MFCC coefficients, SF, ZCR, and STE. Concerning the training databases, they were used to generate speech and non-speech code books.

2) Speech and Silence Discrimination

The detection of silence was performed according to features STE and ZCR by using 1-s window. For the classification, it has been performed by a rule-based classifier: each time when STE and ZCR exceed the predefined threshold value, then they were classified as pure-speech frame, otherwise they were classified as silence frame.

a) Male and Female Speech Discrimination

We describe in this sub-section a voice-based gender identification approach which can be used for the annotation of multimedia content-based indexing. Typically, the range of values of the fundamental frequency for a male speaker is quite

narrow (between 80-200 Hz) and large for a female speaker (150-350Hz). The gender identification system proposed in this work is based on n general audio classifier and it consists of three main steps: In the first step, the features {harmonic ratio and the periodicity estimation} are extracted and normalized (statistics). After that the different segments are clustered using GASOM, KNN or SVM algorithm as a classifier. In this work, we have used the correlation-based pitch estimation feature since it relies considerably on the speech quality.

After the segmentation of the signal, each window obtained of duration T is modeled by a vector composed of two fundamental frequencies in ascending order (low and high frequency) representing the Harmonic Ratio (HR) in that frame. To avoid the incorrect peak selection caused by the existence of sub-harmonics in the spectrum and to look for a single peak representing exactly the sum of the harmonics and sub-harmonics, the sufficiently strong sub-harmonics are examined to see if they can be considered as a pitch candidate or not. Indeed, if the estimated HR in each frame exceeds the HR_thresold value (0.4), then the sub_HR is considered as an f0 candidate, otherwise the harmonic is favored. Therefore, we obtain two matrixes containing the f0 and HR candidates for each frame. After that, the values of the averages and variances of HR are calculated in each frame, and then normalized by their respective maximum so that the classifier captures the relation between the peak in the spectrum and other frequency bands. For the test stage, we have used 50 pairs of voice samples. While, 25 pairs of voice samples has been used to train the gender speech classifier in the training stage. Moreover, each sample is regarded containing a single speaker and the T window used in this stage is a training of basic units, and it is similar to that used in the test stage.

3) Discrimination of Music and Environmental Sounds

This discrimination was performed using non-speech segments. Also, the FS feature was combined with MFCC coefficients and they are used as descriptors for this discrimination. Moreover, one of the algorithms KNN, SVM and GASOM was used as a classifier in this stage. Experiences have proved that the SF feature for music is lower than that for environmental sounds.

a) Discrimination of Music Genres

We have used the long-term feature for each segment of music such as the minimum entropy values and the average SF values of the sequences to discriminate the different musical genres. Also, the decision tree was used as a classifier since it is self-exploratory and easy to interpret. It has to mention here that the long-term feature for classic music has higher values compared to those for electronic music and this can be explained by the smoother energy changes (high-entropy) in the classic music and, these long-term feature values cannot be reached by the Jazz music. Also, we have tried the spectral Rolloff descriptor besides the entropy and the spectral flux, and we have found out that these latter were the best for this kind of discrimination.

F. Discrimination Steps for Second Audio Classification System

1) Music and Speech Discrimination

The statistic values (mean) of the sequences of spectral flux segments were used to discriminate between music and speech. Furthermore, the values obtained for the spectral flux were higher for speech than for music due to the fast alternation of local spectral changes between the speech phonemes. Moreover, we have tried the flux centroid and the chroma vectors as descriptors for this kind of discrimination, and the best discrimination result has been also reached by the spectral flux. Also, one of the algorithms SVM, KNN, and GASOM was used each time as a classifier in this discrimination.

2) Speech and Silence Discrimination, Male and Female

3) Speech Discrimination, and Discrimination of Music Genres

These discriminations have been performed in the same way as those of the first audio classification system.

The two audio classification systems are given in Fig. 3 and 4.

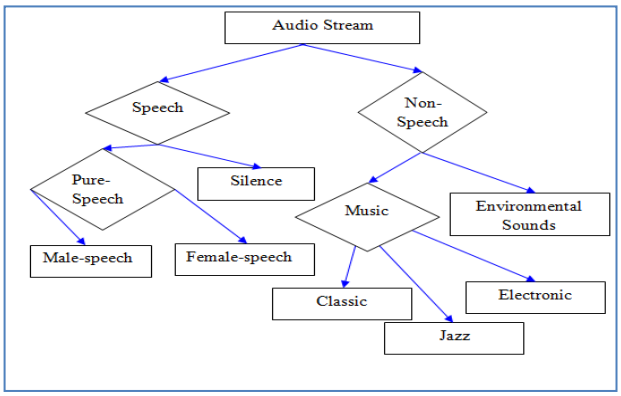

Fig. 3. First audio classification system.

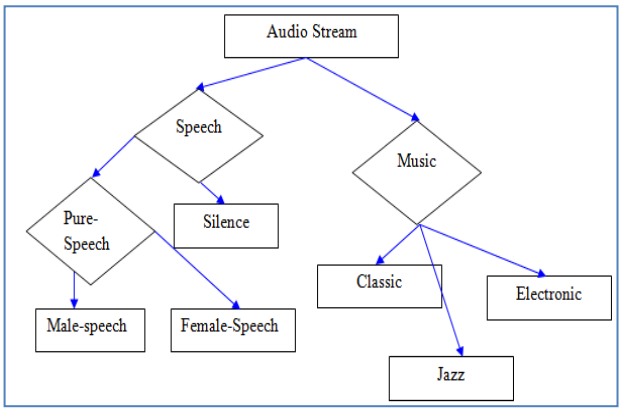

Fig. 4. Second audio classification system.

IV. EVALUATIONS

A. Measures of Performance

To know the type of errors during the training and testing phases, we have used the CM confusion matrix, which is a $N_c \times N_c$ matrix whose rows and columns refer to the true and predicted class labels, respectively, of the dataset. Indeed, the confusion matrix is expressed as follows:

$$CM_n(i,j) = \frac{CM(i,j)}{\sum_{n=1}^{N_c} CM(i,n)} \quad (22)$$

Where, $CM(i,j)$ is the number of samples of class i, which are assigned to class j by the adopted classification method. Also, we have used the *overall accuracy* (*Acc)* which is defined as the ratio of the samples of dataset that have been correctly classified. Indeed, this evaluation criterion has the following expression:

$$Acc = \frac{\sum_{m=1}^{N_c} CM(m,m)}{\sum_{m=1}^{N_c}\sum_{n=1}^{N_c} CM(m,n)} \quad (23)$$

Moreover, in order to describe how well the classification algorithm performs on each class, we define two class-specific measures: the first measure is the *class Recall*,Re(i), which is expressed as the proportion of data with true class label i that are correctly assigned to class i*:*

$$Re(i) = \frac{CM(i,i)}{\sum_{m=1}^{N_c} CM(i,m)} \quad (24)$$

Where $\sum_{m=1}^{N_c} CM(i,m)$ is the total number of samples that are recognized to belong to class i. Concerning the second measure, it is the *class precision, pr(i)*, which is defined as the ratio of samples that are correctly classified to class i with taking into account the total number of samples that are classified to that class.

$$Pr(i) = \frac{CM(i,i)}{\sum_{m=1}^{N_c} CM(m,i)} \quad (25)$$

Where, $\sum_{m=1}^{N_c} CM(m,i)$ is the total number of samples that are classified to class i.

For the *F_1-measure*, it is defined as the harmonic mean values of precision and recall, such as:

$$F_1(i) = \frac{2Re(i)\Pr(i)}{Re(i)+\Pr(i)} \quad (26)$$

B. Validation Methods

To generalize the performance of classifiers outside the training dataset, we have applied in this work two validation approaches:

1) Leave-One-Out Approach

It can be defined as a variation of k-fold cross-validation which splits randomly the dataset into non-overlapping k subset of equal size. Also, this technique is an exhaustive validation technique which is known by producing very reliable validation results.

2) Repeated-Hold-Out Approach

This approach allows refining and repeating k-times the Hold-out approach which splits the dataset into non-overlapping subsets: one for the test and the other for the training. Thus, the division of the dataset into two subsets is performed randomly at each iteration.

V. Results and Analysis

The first audio database used for the evaluation of our algorithms contains many audio types such as speech, music, environmental sounds, others1, others2, others3, which are extracted from different audio events. For the others1 type, it includes low-energy environmental sounds, such as wind, rain, silence, background sound, etc. Concerning the others2 type, it includes environmental sounds with abrupt changes in signal energy such as the sound of thunder, a door closing, an object breaking, etc. While, the others3 type contains high- energy sounds, non-abrupt environmental sounds, such as machine sounds. Also, the audio data in this data set are provided as 4-second chunks at two sampling rates (48 kHz and 16 kHz) with 48 kHz and 16 kHz for respectively the data in stereo and mono. Indeed, the 16 kHz recordings were obtained by down sampling the right-hand channel of the 48 kHz recordings. Thus, each audio file corresponds to a single chunk [45]. Moreover, we have used another data set containing sounds of different music genres, which are extracted from film soundtracks and music effects. Indeed, this dataset consists of 1000 audio tracks each 30 seconds long and it contains 10 genres whose each one is represented by 100 tracks. Furthermore, the tracks are all 22050Hz Mono 16-bit audio files in .wav format [46]. More details about this dataset can be found in [46]. In fact, we have used 2/3 of the dataset for training and 1/3 for testing different classifiers. In this work, we have used KNN, SVM, and GASOM algorithms as classifiers to test our models. We can note from Ttable I that for speech/non-speech discrimination, all algorithms have reached good classification results. Also, for speech/silence discrimination, all algorithms have reached the best classification result which is 100%. Moreover, for male/female speech discrimination, there is a little confusion between the two genres and the best classification value (98.8%) has been reached by GASOM algorithm with the leave-one-out validation technique. Good classification results have been also reached by the GASOM algorithm for music/environmental sounds discrimination in which it has reached the best value (99.4%). In the discrimination of music genres, the best results were 96.4% for classic music, 100% for jazz music, and 94.6% for electronic music, which were all obtained using a decision tree and a GASOM algorithm as classifiers in all previous levels of the audio discrimination process. Also, we can mention from Table I that all algorithms give good classification results in the speech/non-speech, speech/silence, and male/female speech discriminations. Moreover, the SVM algorithm has exceeded the KNN algorithm and it was competitive to GASOM algorithm in all audio discrimination types. Furthermore, the best discrimination results for all discrimination types have been achieved with all algorithms using leave-one-out as a validation technique. For the repeated-hold-out technique, the discrimination results have been always under those obtained with the leave-one-out validation technique.

From Table II, we can show a slight difference between GASOM algorithm and other algorithms in the classification results for the speech/music discrimination. Indeed, the percentage of speech which was recognized as speech is 97.85% for GASOM algorithm with the leave-one-out validation technique against 92.7% and 97.7%, respectively for the KNN and SVM algorithms. In speech/music discrimination, we have also tested the centroid flux and chroma vector, but the best result has been obtained by the spectral flux as it is recorded in Table II. For the silence/speech discrimination, the best results (100%) have been obtained by all algorithms like in the first proposed system. Concerning the

male/female speech discrimination, the best result (95.7%) has been obtained using the GASOM algorithm as a classifier and leave-one-out as a validation technique. Also, this algorithm has proved its dominance by contributing to reach the best classification result using the decision tree as a classifier for the discrimination of music genres in which this classifier has reached the best value (94.2%) for the classic music. For the jazz music, 93.5% was the best classification result achieved by the decision tree as a classifier in the phase of discrimination of musical genres and the KNN algorithm as a classifier in all previous levels of the audio discrimination process. Furthermore, the best classification result for the electronic music (93.3%) has been reached by the decision tree as a classifier in the discrimination of different music genres and the KNN and SVM algorithms as classifiers in all previous levels of the audio discrimination process. Like in the first proposed system, the leave-one-out validation technique in this second audio classification system has mostly reached the best discrimination results compared to the repeated-hold-out validation technique.

Now, we can summarize the efficiency of the two proposed systems by comparing the performance results. From Tables III and IV, we can note that the first audio classification system has proved its success as it has reached the best performance results using different classification algorithms in all levels of the audio discrimination process by comparison to the second audio classification system. Also, the GASOM algorithm has reached the best F1-measure average for the music/environmental sounds discrimination with the leave-one out validation technique. For the male/female speech discrimination in the second audio classification system, the F1-mesure average has reached the best value (94.99%) using GASOM algorithm as a classifier and repeated hold-out as a validation technique. However, it has reached 98.04% in the first audio classification system using the same algorithm and leave-one out as a validation technique. Furthermore, for the discrimination of musical genres, the F1-measure average in the first audio classification system has reached the best value (97.04%) using the decision tree as a classifier and the GASOM algorithm as a classifier (with the leave-one-out validation technique) in all previous levels of the audio discrimination process. However, it has only reached 93.22% in the second audio classification system using the same algorithm and the same validation technique. We can note also that the performance results (for the discrimination of male/female speech and musical genres) were better for the first audio classification system as it contains more stages of audio discrimination. Thus, these discrimination stages have contributed to pure the audio segments from one level of audio discrimination to another until the discrimination of musical genres. For this, the results for discrimination of musical genres in the first audio classification system were better than in the second one.

TABLE I. CONFUSION MATRIX FOR DIFFERENT AUDIO CLASSIFICATION STEPS USING DIFFERENT ALGORITHMS IN THE FIRST AUDIO CLASSIFICATION SYSTEM

Confusion Matrix for Different Audio Classification Steps Using KNN Algorithm								
Leave-One-Out (best K=11)			*leave-one-out (best K=3)*			*Leave-One-Out (best K=3)*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	98.10	1.90
Non-Speech	7.10	92.90	**Silence**	0.00	100	**Male-Speech**	5.80	94.20
Repeated-Hold-Out (best K=7)			*Repeated-Hold-Out (best K=7)*			*Repeated-Hold-Out (best K=3)*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	97.60	2.40
Non-Speech	7.10	92.90	**Silence**	0.00	100	**Male-Speech**	6.00	94.00
Leave-One-Out (best K=3)				*Leave-One-Out (best K=3)*				
Music	95.80	4.20				**Classic**	92.50	7.50	0.00
Environmental Sounds	6.50	93.50				**Jazz**	0.00	100	0.00
Repeated-Hold-Out (best K=3)						**Electronic**	3.40	2.60	94.00
Music	94.60	5.40		*Repeated-Hold-Out (best K=3)*				
Environmental Sounds	6.10	93.90				**Classic**	89.50	9.70	0.80
						Jazz	0.40	98.30	1.30
						Electronic	3.40	2.60	94.00
Confusion Matrix for Different Audio Classification Steps Using SVM Algorithm								
Leave-One-Out			*Leave-One-Out*			*Leave-One-Out*		
Speech	98.1	1.9	**Speech**	100	0.00	**Female-Speech**	98.7	1.3
Non-Speech	5.4	94.6	**Silence**	0.00	100	**Male-Speech**	3.8	96.2
Repeated-Hold-Out			*Repeated-Hold-Out*			*Repeated-Hold-Out*		
Speech	97.1	2.9	**Speech**	100	0.00	**Female-Speech**	97.6	2.4
Non-Speech	7.5	96.9	**Silence**	0.00	100	**Female-Speech**	6.00	94.00
Leave-One-Out					*Leave-One-Out*			

Music	97.9	2.1			**Classic**	93.1	6.9	0.00
Environmental Sounds	2.6	97.4			**Jazz**	0.00	100	0.00
Repeated-Hold-Out					**Electronic**	2.40	1.00	96.6
Music	97.1	2.9				repeated-hold-out		
Environmental Sounds	2.8	97.2			**Classic**	89.5	9.7	0.8
					Jazz	0.4	98.3	1.3
					Electronic	3.4	2.6	94.0
Confusion Matrix for Different Audio Classification Steps Using GASOM Algorithm								
Leave-One-Out			*Leave-One-Out*			*Leave-One-Out*		
Speech	98.3	1.70	**Speech**	100	0.00	**Female-Speech**	98.80	1.20
Non-Speech	4.4	95.60	**Silence**	0.00	100	**Male-Speech**	2.80	97.20
Repeated-Hold-Out			*Repeated-Hold-Out)*			*Repeated-Hold-Out*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	97.60	2.40
Non-Speech	7.50	92.50	**Silence**	0.00	100	**Female-Speech**	2.90	97.10
Leave-One-Out					*Leave-One-Out*			
Music	**99.40**	**0.60**			**Classic**	96.4	3.6	0.00
Environmental Sounds	1.05	98.95			**Jazz**	00.00	100	0.00
Repeated-Hold-Out					**Electronic**	4.40	1.00	94.60
Music	97.80	2.20			Repeated-Hold-Out			
Environmental Sounds	2.40	97.60			**Classic**	91.50	7.70	0.80
					Jazz	0.40	98.8	0.80
					Electronic	3.40	2.60	94.00

TABLE II. CONFUSION MATRIX FOR DIFFERENT AUDIO CLASSIFICATION STEPS USING DIFFERENT ALGORITHMS IN THE SECOND AUDIO CLASSIFICATION SYSTEM

Confusion Matrix for Different Audio Classification Steps Using KNN Algorithm								
Leave-One-Out (best K=13)			*leave-one-out (best K=13)*			*Leave-One-Out (best K=13)*		
Speech	92.70	7.30	**Speech**	100	0.00	**Female-Speech**	93.10	6.90
Music	9.25	90.75	**Silence**	0.00	100	**Male-Speech**	5.80	94.20
Repeated-Hold-Out (best K=15)			*Repeated-Hold-Out (best K=15)*			*Repeated-Hold-Out (best K=15)*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	92.60	7.40
Music	7.50	92.50	**Silence**	0.00	100	**Male-Speech**	6.20	93.80
Leave-One-Out (best K=13)					*Repeated-Hold-Out (best K=15)*			
Classic	92.50	7.50	0.80			**Classic**	91.50	7.70	0.80
Jazz	6.50	93.50	0.00			**Jazz**	4.40	90.70	4.90
Electronic	4.20	2.50	93.30			**Electronic**	4.40	2.60	93.00
Confusion Matrix for Different Audio Classification Steps Using SVM Algorithm								
Leave-One-Out			*Leave-One-Out*			*Leave-One-Out*		
Speech	97.70	2.30	**Speech**	100	0.00	**Female-Speech**	94.80	5.20
Music	3.60	96.40	**Silence**	0.00	100	**Male-Speech**	5.80	94.20
Repeated-Hold-Out			*Repeated-Hold-Out*			*Repeated-Hold-Out*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	94.60	5.40
Music	4.10	95.90	**Silence**	0.00	100	**Female-Speech**	6.05	93.95
Leave-One-Out					*Repeated-Hold-Out*			
Classic	93.20	6.80	0.00		**Classic**	92.10	7.10	0.80

Jazz	7.50	97.4	0.00		**Jazz**	4.10	90.70	5.20
Electronic	4.20	2.50	93.30		**Electronic**	4.80	2.60	92.60
Confusion Matrix for Different Audio Classification Steps Using GASOM Algorithm								
Leave-One-Out			*Leave-One-Out*			*Leave-One-Out*		
Speech	97.85	2.15	**Speech**	100	0.00	**Female-Speech**	94.80	5.20
Music	3.40	96.60	**Silence**	0.00	100	**Male-Speech**	5.80	94.20
Repeated-Hold-Out			*Repeated-Hold-Out)*			*Repeated-Hold-Out*		
Speech	97.30	2.70	**Speech**	100	0.00	**Female-Speech**	95.70	4.30
Music	4.00	96.00	**Silence**	0.00	100	**Female-Speech**	5.80	94.20
Leave-One-Out					*Repeated-Hold-Out*			
Classic	94.20	5.80	0.00		**Classic**	92.60	7.00	0.40
Jazz	6.90	98.95	0.00		**Jazz**	3.20	91.40	5.40
Electronic	5.25	2.50	92.25		**Jazz**	4.80	2.60	92.60

TABLE III. DIFFERENT PERFORMANCE RESULTS OBTAINED USING DIFFERENT ALGORITHMS FOR THE FIRST AUDIO CLASSIFICATION SYSTEM

The Performance Results Using KNN Algorithm					
Classification Type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Non-Speech	**Repeated-Hold-Out**	95.00	95.10	95.00	95.04
	Leave-One-Out	95.00	95.10	95.00	95.04
Speech -Silence	**Repeated-Hold-Out**	100	100	100	100
	Leave-One-Out	100	100	100	100
Female and Male Speech	**Repeated-Hold-Out**	95.80	95.90	95.80	95.84
	Leave-One-Out	96.15	96.25	96.15	96.19
Music and Environmental Sounds	**Repeated-Hold-Out**	97.15	97.25	97.15	97.19
	Leave-One-Out	94.65	94.75	94.65	94.69
Classic, Jazz and Electronic Music	**Repeated-Hold-Out**	90.10	91.2	94.1	92.62
	Leave-One-Out	95.26	95.36	95.26	95.30
The Performance Results Using SVM Algorithm					
Classification Type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Non-Speech	**repeated-hold-out**	97.00	97.10	97.00	97.04
	leave-one-out	96.35	96.45	96.35	96.39
Speech -Silence	**repeated-hold-out**	100	100	100	100
	leave-one-out	100	100	100	100
Female and Male Speech	**repeated-hold-out**	95.80	95.90	95.80	95.84
	leave-one-out	97.45	97.55	97.45	97.49
Music and Environmental Sounds	**repeated-hold-out**	97.65	97.75	97.65	97.69
	leave-one-out	96.56	96.66	96.56	97.11
Classic, Jazz and Electronic Music	**repeated-hold-out**	93.93	94.00	93.93	93.96
	leave-one-out	96.56	96.70	96.56	96.63
The Performance Results Using GASOM Algorithm					
Classification Type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Music	**Repeated-Hold-Out**	96.35	96.45	96.35	96.39

	Leave-One-Out	96.95	97.00	96.95	96.97
Speech –Silence	**Repeated-Hold-Out**	100	100	100	100
	Leave-One-Out	100	100	100	100
Female and Male Speech	**Repeated-Hold-Out**	97.35	97.45	97.35	97.39
	Leave-One-Out	98.00	98.10	98.00	98.04
Music and Environmental Sounds	**Repeated-Hold-Out**	97.70	97.80	97.79	97.74
	Leave-One-Out	99.17	99.27	99.17	99.21
Classic, Jazz and Electronic Music	**Repeated-Hold-Out**	94.76	94.86	94.76	94.80
	Leave-One-Out	97.00	97.10	97.00	97.04

TABLE IV. THE OBTAINED PERFORMANCES USING DIFFERENT ALGORITHMS FOR THE SECOND AUDIO CLASSIFICATION SYSTEM

The Obtained Performance Using KNN Algorithm					
Classification type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Music	**repeated-hold-out**	91.22	91.32	91.22	91.26
	leave-one-out	97.05	97.15	97.05	97.09
Speech -Silence	**repeated-hold-out**	100	100	100	100
	leave-one-out	100	100	100	100
Female and Male Speech	**repeated-hold-out**	93.20	93.30	93.20	93.24
	leave-one-out	93.95	94.05	93.95	93.99
Classic, Jazz and Electronic Music	**repeated-hold-out**	90.1	91.2	94.1	92.62
	leave-one-out	93.10	93.20	93.10	93.14
The Obtained Performance Using SVM Algorithm					
Classification type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Music	**Repeated-Hold-Out**	96.50	96.60	96.50	96.54
	Leave-One-Out	97.05	97.15	97.05	97.09
Speech -Silence	**Repeated-Hold-Out**	100	100	100	100
	Leave-One-Out	100	100	100	100
Speech Female and Speech Male	**Repeated-Hold-Out**	94.27	94.37	94.27	94.31
	Leave-One-Out	94.50	94.60	94.50	94.54
Classic, Jazz and Electronic Music	**Repeated-Hold-Out**	91.86	91.96	91.86	91.90
	Leave-One-Out	93.00	93.10	93.00	93.04
The Obtained Performance Using GASOM Algorithm					

Classification type	Validation Method	Overall accuracy	Average Precision	Average Recall	Average F1 measure
Speech-Music	**Repeated-Hold-Out**	96.65	96.75	96.65	96.69
	Leave-One-Out	97.22	97.32	97.22	97.26
Speech -Silence	**Repeated-Hold-Out**	100	100	100	100
	Leave-One-Out	100	100	100	100
Speech Female and Speech Male	**Repeated-Hold-Out**	94.95	95.05	94.95	94.99
	Leave-One-Out	94.50	94.60	94.50	94.54
Classic, Jazz and Electronic Music	**Repeated-Hold-Out**	92.20	92.30	92.20	92.24
	Leave-One-Out	93.18	93.28	93.18	93.22

To make a comparison, the first audio classification system has been developed using Hidden Markov Model (HMM) and Multilayer Perceptron *(*MLP) classifier. For the MLP classifier, it is a Multilayer Perceptron Feed Forward Fully Connected Neural Network (MLPFFFCNN) with a sigmoid activation function. Indeed, it is a neural network with 3 hidden layers with 4 neurons for each one and a number of output units equals to the number of classes. Concerning the training of this classifier, it has been carried out using back propagation algorithm and the stopping criterion has been set according to the Mean Square Error (MSE) when it reaches the zero value. For the second classifier, it is a background HMM with 4 states in order to represent the sequences of observation vectors. Moreover, a refinement stage has been added using a Viterbi decoding as a resegmentation stage in order to refine the segmentation results. The Esperance-Maximization (EM) algorithm has been also used in order to learn the parameters of HMM. As it is shown in Table V, the HMM classifier has succeeded to achieve good results in terms of measured performances by comparison to MLP algorithm. Indeed, it has reached the best F1-measure averages in all levels of the first audio classification system. However, these results remain competitive to those obtained with GASOM algorithm which has succeeded to reach the best results as it was mentioned above.

Furthermore, the GASOM algorithm has outperformed the HMM and MLP algorithms in terms of time consumption for which it has reached the best results in all audio classification levels as it shown in Table VI. Thus, this speed of processing makes this algorithm so desired in real-time applications.

TABLE V. DIFFERENT PERFORMANCE RESULTS OBTAINED USING THE MLP AND HMM ALGORITHMS FOR THE FIRST AUDIO CLASSIFICATION SYSTEM

Classification type	Classifier	Overall accuracy	Average Precision	Average Recall	Average F1 measure
Speech-Non-Speech	*MLP*	95.50	95.60	95.50	95.54
	HMM	96.00	96.10	96.00	**96.04**
Speech – Silence	*MLP*	100	100	100	100
	HMM	100	100	100	**100**
Female speech/Male Speech	*MLP*	95.8	95.9	95.8	95.84
	HMM	97.00	97.10	97.00	**97.04**
Music /Environmen-tal Sound	*MLP*	96.75	96.85	96.75	96.79
	HMM	97.30	97.40	97.30	**97.34**
Classic, Jazz and Electronic Music	*MLP*	93.80	93.90	93.80	93.84
	HMM	95.50	95.60	95.50	**95.54**

TABLE VI. THE OBTAINED TIME CONSUMPTION IN ALL AUDIO CLASSIFICATION LEVELS USING MLP, HMM AND GASOM ALGORITHMS

Audio classification step Algorithm	Speech-Non Speech	Speech - Silence	Female speech/ Male Speech	Music /Environ--mental Sounds	Classic, Jazz and Electronic Music
MLP	1.6	1.0	1.9	1.1	1.8
HMM	1.2	0.9	1.1	0.8	1.4
GASOM	0.5	0.6	0.7	0.5	1.0

VI. Conclusion and Future Work

Two audio classification systems have been proposed in this work in which an audio stream is discriminated into homogenous regions and classified into basic audio types such as speech, non-speech, silence, music, environmental sounds and so on. The principle goal is to exploit audio segmentation algorithms which can be integrated in multimedia content analysis applications and audio recognition systems. Indeed, three algorithms have been used for the two audio classification and segmentation systems. For the first system, the audio stream has been discriminated into speech/non-speech, pure-speech/silence, male/female speech, environmental sounds/music, music genres: classic, jazz, and electronic music. Concerning the second system, the audio stream has been segmented into speech/music, pure-speech/silence, male/female speech, music genres: classic, jazz, and electronic music. For the discrimination of musical genres and pure-speech/silence, the decision tree and a rule-based classifier are respectively used as classifiers. While, in the retaining levels of two audio classification systems, one of the algorithms KNN, SVM, and GASOM has been used as a classifier. Experimental results have shown that the GASOM algorithm is so efficient for most audio discrimination types in terms of accuracy and time consumption. Thus, this advantage plus the no-requirement of much training data makes this algorithm very useful for real-time multimedia applications. In future work, the two proposed systems can be exploited to perform many applications, such as the automatic speech recognition, the human-computer interaction systems, the speaker tracking, and so on. Also, the GASOM algorithm can be combined with k-means algorithm in order to access more data and achieve better performances.

References

[1] Chih-Chin Liu, Jia-Lien Hsu, Chen ALP (1999) An approximate string matching algorithm for content-based music data retrieval, IEEE IntConf Multimedia Comp Syst 1: 451–456.

[2] Kosugi N, Nishihara Y, Kon'ya S, Yamamuro M, Kushima K (1999) Music retrieval by humming-using similarity retrieval over high dimensional feature vector space, IEEE Pacific Rim ConfCommun, Comp Signal Processing 404–407.

[3] Kataoka M, Kinouchi M, Hagiwara M (1998) Music information retrieval system using complex-valued recurrent neural networks, IEEE IntConfSyst, Man, and Cybernetics, 5: 4290–4295.

[4] Guohui Li, Khokhar AA (2000) Content-based indexing and retrieval of audio data using wavelets, ICME 2000 2: 885–888.

[5] Subramanya SR, Youssef A (1998) Wavelet-based indexing of audio data in audio/multimedia databases, ProcInt Workshop on Multi-Media Database Management Sys 46–53.

[6] I.McLoughlin,*AppliedSpeechandAudioProcessing:WithMATLAB Examples*, Nanyang Technological University, CambridgeUniversity Press, 2009.

[7] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speechand Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[8] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, 2003.

[9] M. L. Coz, J. Pinquier, R. Andre-Obrecht, and J. Mauclair, "Audio indexing including frequency tracking of simultaneous multiple sources in speech andmusic," in *Proceedings of the 11th InternationalWorkshop on Content-Based Multimedia Indexing(CBMI '13)*, pp. 23–28, IEEE, Veszprem, Hungary, June 2013.

[10] S. Kiranyaz, A. F. Qureshi, and M. Gabbouj, "A generic audio classification and segmentation approach for multimedia indexing and retrieval," *IEEE Transactions on Audio, Speech andLanguage Processing*, vol. 14, no. 3, pp. 1062–1081, 2006.

[11] Scheirer E, Slaney M (1997) Construction and evaluation of a robust multifeature speech/music discriminator, ICASSP-97 2: 1331–1334.

[12] A. Dessein and A. Cont, "An information-geometric approach to real-time audio segmentation," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 331–334, 2013.

[13] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi, "Audio segmentation and speaker localization in meeting videos," in *Proceedings of the 18th International Conference on PatternRecognition (ICPR '06)*, vol. 2, pp. 1150–1153, IEEE, Hong Kong, August 2006.

[14] J. Huang, Y. Dong, J. Liu, C. Dong, and H.Wang, "Sports audio segmentation and classification," in *Proceedings of the IEEEInternational Conference on Network Infrastructure and DigitalContent (IC-NIDC '09)*, pp. 379–383, IEEE, Beijing, China, November 2009.

[15] J. Hennebert, M. Hasler, and H. Dedieu, "Neural networks in speech recognition," in *Proceedings of the 6th MicrocomputerSchool of Neural Networks, Theory and Applications (Micro-Computer '94)*, pp. 23–40, Prague, Czech Republic, 1994.

[16] P. Nguyen, G. Heigold, and G. Zweig, "Speech recognition with flat direct models," *IEEE Journal on Selected Topics in SignalProcessing*, vol. 4, no. 6, pp. 994–1006, 2010.

[17] S.-X. Zhang and M. J. F. Gales, "Structured SVMs for automatic speech recognition," *IEEE Transactions on Audio, Speech andLanguage Processing*, vol. 21, no. 3, pp. 544–555, 2013.

[18] P. Hu, W. Liu, W. Jiang, and Z. Yang, "Latent topic model for audio retrieval," *Pattern Recognition*, vol. 47,no. 3, pp. 1138–1143, 2014.

[19] Birkenes, T. Matsui, K. Tanabe, S. M. Siniscalchi, T. A. Myrvoll, andM. H. Johnsen, "Penalized logistic regression with HMM log-likelihood regressors for speech recognition," *IEEETransactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1440–1454, 2010.

[20] Z. Ping, T. Li-Zhen, and X. Dong-Feng, "Speech recognition algorithm of parallel subband HMM based on wavelet analysis and neural network," *Information Technology Journal*, vol. 8, no. 5, pp. 796–800, 2009.

[21] V. R. V. Krishnan and P. BabuAnto, "Features of wavelet packet decomposition and discrete wavelet transform for Malayalam speech recognition," *International Journal of Recent Trends inEngineering*, vol. 1, no. 2, pp. 93–96, 2009.

[22] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Capp´e, "A regularized kernel-based approach to unsupervised audio segmentation," in *Proceedings of the IEEE International Conferenceon Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 1665–1668, IEEE, Taipei, Taiwan, April 2009.

[23] T. Giannakopoulos and S. Petridis, "Detection and clustering of musical audio parts using Fisher linear semi-discriminant analysis," in *Proceedings of the IEEE 20th European SignalProcessing Conference (EUSIPCO '12)*, pp. 1289–1293, IEEE, Bucharest, Romania, August 2012.

[24] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, 2003.

[25] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-basedmusic information retrieval: current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[26] Hemant Misra, ShajithIkbal, HervéBourlard, HynekHermansky, Spectral entropy based feature for robust ASR, in: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'04, vol. 1, IEEE, 2004, pp. I–193.

[27] A. Pikrakis, T. Giannakopoulos, S. Theodoridis, A computationally efficient speech/music discriminator for radio recordings, in: International Conference on Music Information Retrieval and Related Activities, ISMIR06, 2006.

[28] A. Pikrakis,T. Giannakopoulos, S.Theodoridis,A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks, IEEETransactions on Multimedia 10 (5) (2008) 846–857.

[29] TheodorosGiannakopoulos,SergiosPetridis,Unsupervised speaker clustering in a linear discriminant subspace,in:Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications, ICMLA '10, 2010, pp. 1005–1009.

[30] G. Tzanetakis,P. Cook,Musical genre classification of audio signals, IEEETransactions on Speech and Audio Processing 10 (5) (2002) 293–302.

[31] David Pearce, Hans günterHirsch,The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, in: ISCA ITRWASR2000, 2000, pp. 29–32., Ericsson Eurolab Deutschland Gmbh.

[32] GregoryH.Wakefield,Mathematical representation of joint time-chroma distributions, in: SPIE's International Symposium on Optical Science,Engineering, and Instrumentation, International Society for Optics and Photonics, 1999, pp. 637–645.

[33] Mark A. Bartsch, Gregory H.Wakefield,Audiothumbnailing of popular music using chroma-based representations, IEEETransactions on Multimedia 7 (1) (2005) 96–104.

[34] Mark A. Bartsch, Gregory H.Wakefield,To catch a chorus: using chroma-based representations for audio thumbnailing, in: 2001 IEEEWorkshop on the Applications of Signal Processing to Audio and Acoustics, IEEE, 2001, pp. 15–18.

[35] Meinard Müller, Frank Kurth, Michael Clausen, Audio matching via chroma-based statistical features, in: Proceedings of ISMIR, London, GB, 2005, pp. 288–295.

[36] A. Pikrakis,T. Giannakopoulos, S.Theodoridis,A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks, IEEETransactions on Multimedia 10 (5) (2008) 846–857.

[37] Lawrence R. Rabiner, RonaldW. Schafer, Introduction to digital speech processing, Foundations andTrends in Signal Processing,Now Publishers Inc, 2007.

[38] J.Weston and C. Watkins, "Support vector machines for multiclass pattern recognition," in *Proceedings of the 7th EuropeanSymposium on Artificial Neural Networks (ESANN '99)*, vol. 99, 1999.

[39] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on NeuralNetworks*, vol. 13, no. 2, pp. 415–425, 2002.

[40] K.-B.Duan and S. S. Keerthi, "Which is the bestmulticlass SVM method? An empirical study," in *Multiple Classifier Systems*, vol. 3541 of *Lecture Notes in Computer Science*, pp. 278–285, Springer, Berlin, Germany, 2005.

[41] P. Clarkson and P. J. Moreno, "On the use of support vector machines for phonetic classification," in *Proceedings of theIEEE International Conference on Acoustics, Speech, and SignalProcessing (ICASSP '99)*, vol. 2, pp. 585–588, IEEE, Phoenix, Ariz, USA, March 1999.

[42] Marie Cottrell et al : 'Cartes auto-organisées pour l'analyse exploratoire de données et la visualisation', Université Paris 1, 2003.

[43] Gaussier E, et al: 'A hierarchical model for clustering and categorising documents', Advances in Information Retrieval,Proceedings of the 24th BCS-IRSG European Colloquium on IR Research ECIR-02, Glasgow. Lecture Notes in Computer Science 2291, p.229-247, Springer, 2002.

[44] Mohamed Salah Salhi, NajetArous, NoureddineEllouze: 'A suitable model of evolutionary SOM for phonemes recognition', journal IRECOS Napoly-Italy, (Indexé COMPENDEX – Elsevier– Copernicus). Laboratoire LSTS- ENIT_Tunis, septembre 2009.

[45] http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio.

[46] George Tzanetakis, Gtzan genre collection. http://marsyas.info/download/data_sets.