A two level strategy for audio segmentation

Sébastien Lefèvre, Nicole Vincent

A Two Level Strategy for Audio Segmentation

Sébastien Lefèvre [a,*]

[a] *LSIIT – Université Louis Pasteur (Strasbourg I)*
Parc d'Innovation, Bd Brant, BP 10413, 67412 Illkirch Cedex, France

Nicole Vincent [b]

[b] *CRIP5 – Université René Descartes (Paris V)*
45 rue des Saints Pères, 75270 Paris Cedex 06, France

Abstract

In this paper we are dealing with audio segmentation. The audio tracks are sampled in short sequences which are classified into several classes. Every sequence can then be further analysed depending on the class it belongs to. We first describe simple techniques for segmentation in two or three classes. These methods rely on amplitude, spectral or cepstral analysis, and classical hidden markov models. From the limitations of these approaches, we propose a two level segmentation process. The segmentation is performed by computing several features for each audio sequence. These features are computed either on a complete audio segment or on a frame (set of samples) which is a subset of the audio segment. The proposed approach for microsegmentation of audio data consists of a combination of a K-Mean classifier at the segment level and of a Multidimensional Hidden Markov Model system using the frame decomposition of the signal. A first classification is obtained using the K-Mean classifier and segment-based features. Then final result comes from the use of Multidimensional Hidden Markov Models and frame-based features involving temporary results. Multidimensional Hidden Markov Models are an extension of classical Hidden Markov Models dedicated to multicomponent data. They are particularly adapted to our case where each audio segment can be characterized by several features of different natures. We illustrate our methods in the context of analysis of football audio tracks.

Key words: Audio segmentation, Cepstral Analysis, K-Mean, Multidimensional Hidden Markov Models, Multilevel Analysis

* Corresponding author.
Phone: +33 390244570. Fax: +33 390244455. E-mail: lefevre@lsiit.u-strasbg.fr.
Postal: LSIIT, Parc d'Innovation, Bd Brant, BP 10413, 67412 Illkirch Cedex, France.
Email addresses: lefevre@lsiit.u-strasbg.fr (Sébastien Lefèvre), nicole.vincent@math-info.univ-paris5.fr (Nicole Vincent).

1 Introduction

Analysis and classification of audio data happen to be important tasks in many applications, such as speech recognition or content-based indexing or retrieval. Several mathematical tools are frequently used in this research field, as for example Neural Networks or Hidden Markov Models (HMM). However most of the HMM-based approaches proposed in the literature are dedicated to speech recognition. The aim of the method presented in this paper is not a speech recognition task, but rather a segmentation of the audio data into different clusters identified by a label. In our application we are using the audio track associated with a football video report. Three elements are to be identified, the referee whistle, the crowd noise and the speaker voice. Such a process can bring further information in the global interpretation task of the video sequence. Besides more specific analysis can be later performed on each of the labeled parts of the audio sequence. For instance, it is possible to look for occurrences of predefined words in the speaker voice data. The analysis of crowd noise can also give an important information. Finally, the knowledge available on football game rules can be used to search for some specific actions (goal scored, match end).

The most important difficulty to be solved in such a problem lies in the variability of the characteristics of each class to be extracted. This is the point that motivated the choices in our method. Two observation levels are considered, the segment level and the frame level. In a first step the whole segments are clustered in order to deal with more similar parts. These clusters have nothing to do with the final information we want to extract from the data. No semantic label can be assigned to them. They enable to adapt the process to the shared characteristics within each cluster. The second step relies on the use of original HMM models.

In the first part we will review how audio data is processed in other systems. Then we will justify the need for a more complex method by experiments using simple methods. The studied approaches are related to analysis of the frequency and the amplitude of the audio signal, but also to cepstral analysis and classical hidden markov models. Then we will present an original segmentation method which is a new combination of a K-Mean classifier and Multidimensional Hidden Markov Models. We will recall the mathematical tools we are to use in our method before we will detail both steps of our audio segmentation method. Finally results on football broadcast tracks will be presented and commented.

2 Audio data analysis

When an application is concerned, a first step in the methodology is to determine whether the analysis has to be performed on a global or a local basis. In the first case, the goal would be to classify complete audio sequences, as in the approach by Wang *et al.* [1] who classify TV audio tracks. It is also possible to classify short audio segments (typically less than one second long) in order to detect events in audio sequences, as in the work from Kermit and Eide [2]. Event detection can even be performed in real time [3].

Segmentation and classification of audio data have been studied by many researchers. They can be seen as pattern recognition problems where two issues have to be solved: choice of the classifier and selection of audio features. Li *et al.* [4] studied a total of 143 features to determine their discrimination capability. Pfeiffer *et al.* describe in [5] basic features used in audio analysis. Wold *et al.* [6] analyse and compare audio features for content-based audio indexing purpose. Li [7] performs experiments to compare various classification methods and feature sets. Bocchieri and Wilpon [8] discuss the influence of the feature number and the need for feature selection. When dealing with compressed audio tracks, it is also possible to compute some specific features, as in the work from Tzanetakis and Cook [9] with MPEG audio.

Several researchers have proposed to use HMM to perform audio analysis. Kimber and Wilcox [10] create a HMM for each speaker or acoustic class. Learning and recognition are respectively performed using the well-known Baum-Welch and Viterbi algorithms. Battle and Cano [11] propose to use Competitive HMM instead of traditional HMM in case of unsupervised training. Finally Hirsch [12] uses an adaptative HMM architecture in order to deal with audio signal from telecommunications.

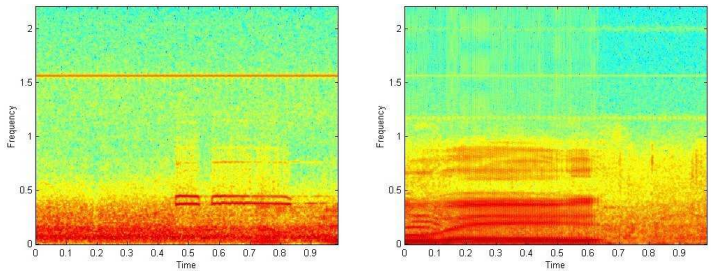
3 Some basic segmentation strategies

As far as the signal that we want to recognize is concerned, frequency or amplitude of the signal are possible discriminating elements. In this section we will see how to use them to extract whistle sound and crowd. We will also show the limitations of this kind of approaches. Then a more complex but higher quality method based on a cepstral analysis will be detailed. In all these approaches, speaker and whistle can be discriminated from a spectral analysis whereas crowd and speaker are differing form an amplitude or a cepstral point of view. When dealing with segmentation of more than two classes, more sophisticated methods should be used, and we propose in this case a HMM based-method.

3.1 Spectral and amplitude analysis

Before proposing a complex method for audio segmentation, we have to check and show the limitations of simple approaches. Here we present two methods for segmentation of audio data in 2 classes. These methods are respectively based on frequency and amplitude analysis of the audio signal. We consider two different segmentations. On the one hand whistle/non whistle and on the other hand detection of speaker and crowd.

In order to detect audio segments with referee whistle, we start from the assumption that the sound produced by a whistle is composed of two or three frequencies belonging to the interval $[3700, 4300]$ Hz. An example of a spectrogram representing a segment with whistle is shown in figure 1 (a). We can clearly see the horizontal lines representing the frequencies of the whistle sound. The segmentation into whistle sound / non whistle sound is performed through three successive steps. The spectrogram can be thresholded in order to keep only most significant values. Then for each frequency the amplitude associated with the frequency in the spectrogram is computed. Only frequencies with an amplitude higher than a predefined threshold are considered. Finally the audio segment is classified into whistle sound if the number of resulting lines is higher or equal to two. The main limitation of this method comes from the fact that the whistle sound frequencies can be a subset of the speaker voice frequencies. It results in the confusion between the speaker voice and the referee whistle. Figure 1 (b) containing the spectrogram of a speaker audio segment illustrates this problem.



(a) (b)

Fig. 1. Spectrograms of audio signals corresponding to whistle referee (a) and speaker voice (b).

Besides audio segmentation with labels speaker/crowd can be based on signal amplitude analysis. Audio signal containing some segments classified into speaker and some others into crowd is presented in figure 2. We can see that the average amplitude is not equal for both classes. So in a simple way, it is possible to segment audio data into speaker or crowd depending on the average of the signal amplitude. If the average amplitude is higher than a fixed threshold the audio segment is classified into speaker voice. Otherwise it is classified into crowd noise. We consider properties of audio data are almost constant all along the sequence. So an adap-

tative threshold is not necessary and a fixed threshold is used instead. A learning strategy is used to determine this threshold. As audio tracks to be studied may be issued from different broadcast sources, so with properties of high variability, the learning should be done, on line, by relying on a corpus made from the first seconds of the audio track. Results are presented in table 1 where recall and precision are defined for a given class as the ratio between the correctly classified sequences and respectively the total number of sequences belonging to this class and the total number of sequences classified in this class. We can conclude from these results that the quality of the method is not sufficient to segment correctly our audio data into speaker or crowd.

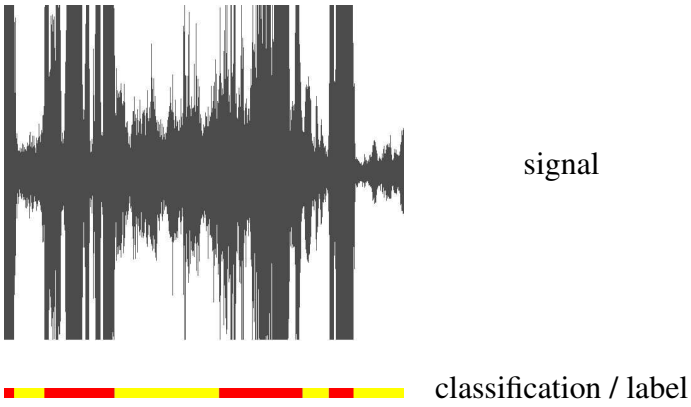


Fig. 2. Audio signal (top) containing segments classified (bottom) either into speaker (red) or crowd (yellow).

Quality rates presented in tables included in this paper were obtained by comparing results using the described segmentation methods and a ground truth which has been obtained by manual scoring of audio data by several users. Only audio segments for which a consensus between all users has been obtained are considered. Finally, audio sequences were belonging to different broadcasts relative to several football games.

Table 1
Results of 2 class segmentation based on the signal amplitude analysis.

Class	Recall	Precision
Whistle	77 %	50 %
Speaker	62 %	84 %

The two simple previous approaches presented here do not allow to correctly segment audio data into 2 different classes. In order to classify sequences into speaker or crowd, it is possible to use more complex features as cepstral analysis. This will be decribed in next section.

3.2 Cepstral analysis

Cepstrum is a tool widely used in speech analysis and recognition. It is defined as a combination of three successive steps: Fourier transform, logarithm, and inverse Fourier transform. It allows to determine the speech fundamental frequency and to separate excitation signal and pure speech signal. As a spectrogram represents the spectrum of a signal, it is possible to use a cepstrogram which is a 3-D graphical representation of the audio signal based on the cepstrum computation. Figure 3 shows the 2-D projections from two cepstrograms of audio segments produced respectively by crowd and speaker.

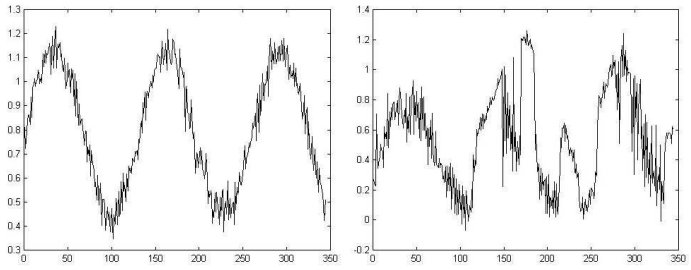


Fig. 3. 2-D projection from the cepstrogram of an audio signal corresponding to crowd (left) and to speaker voice (right).

From these figures we can understand the crowd is represented by a more or less sinusoidal curve contrary to the speaker voice. Two reasons can be found to explain this phenomenon: the sound produced by the crowd can be considered as noise, or it can be amplified by the stadium (echo phenomenon). So it is possible to segment these two kinds of sounds based on this principle. First, we regress the curve obtained with a sinusoidal curve with same frequency and phase. Then we compute an euclidean distance between theoretical sinusoidal curve and observed signal. If the distance obtained is below a threshold, audio segment is classified into crowd. Otherwise it is classified into speaker voice class. As in the method based on amplitude analysis, threshold can be given from a supervised learning procedure. Results of this method are shown in table 2.

Table 2
Results of 2 class segmentation based on cepstral analysis.

Class	Recall	Precision
Crowd	72 %	96 %
Speaker	98 %	86 %

We can see the quality of this method is higher than the simple approaches based on spectral and amplitude analysis. However the quality rates are not satisfying yet and the method described here cannot deal with more than two different classes. A process to build a decision tree with selection of discriminating features could be organized, but would have to be achieved afresh when a new class is added to

the study. So the use of Hidden Markov Models can be suited to characterize the samples from each class in order to later perform a classification of new samples.

3.3 3 class segmentation using HMM

The methods described previously were based on a single feature and were unable to ensure satisfying results. In order to combine these different features, it is possible to use Hidden Markov Models.

Hidden Markov Models (HMM) are one of the most often used tools in speech analysis and processing. A good introduction to these models can be found in [13]. We present here a method based on this classical tool applied to segmentation of audio data into three classes which are referee whistle, crowd, and speaker voice.

The segmentation method proposed here is based on ergodic HMM rather than left to right HMM. Learning and recognition are respectively performed using the well-known Baum-Welch [14] and Forward algorithms. We define three HMM, one for each class: referee whistle, crowd, and speaker voice. Each segment will be classified into the class with the highest score.

Observation data consists of a set of features successfully explored in [15]. For each audio segment with a duration of 1 second, we compute 11 features: non-silence ratio (NSR), volume standard deviation (VSTD), standard deviation of zero crossing rate (ZSTD), volume dynamic range (VDR), standard deviation of pitch period (PSTD), smooth pitch ratio (SPR), non-pitch ratio (NPR), frequency centroid (FC), frequency bandwidth (FB), 4 Hz modulation energy (4ME), and energy ratio of subband 1-3 (ERSB1-3). When dealing with several features in audio processing, it is necessary to determine which features provide the greatest contribution to the recognition performance and to select the most efficient features [8]. So we performed a Principal Component Analysis on these 11 features. As a result no feature was rejected because of its lack of contribution.

In order to analyse 1 second long audio segments, we divide them into frames containing 1024 samples. Two successive frames are shifted of 512 samples. Table 3 shows results of a 3 class segmentation. Quality rates are better than with simple approaches reviewed in previous section.

In order to improve the classification quality, we propose to use multidimensional hidden Markov models instead of standard HMM and to combine them with a K-Mean classifier. The method which also includes a cepstral analysis will be described in next section.

Table 3
Results of 3 class segmentation based on Hidden Markov Models.

Class	Recall	Precision
Whistle	88 %	88 %
Crowd	61 %	87 %
Speaker	77 %	90 %

4 A two step strategy

From the previous experiments, we have noticed that the main difficulty in audio segmentation comes form the variability of feature within each class. So we have decided to consider two levels of observation, the segment level and the frame level. First the complete segments are classified in order to obtain classes with more similar features values. However these classes have no link with the final information we want to extract from the data. Their goal is only to let the process adapt to the features shared inside each class. The second step is then based on the results obtained and on the more efficient learning of Multidimensional HMM because of a more homogeneous population.

4.1 K-Mean Classifier

K-Mean classifier is a tool widely used in pattern recognition [16]. This classifier considers a number K of clusters which is determined *a priori*. The resulting partition is defined by the location of the K cluster centers and is obtained by minimising the average distortions (computed from an Euclidean distance) of all data points belonging to the K clusters. We briefly recall here the successive steps of this algorithm.

First we set randomly the K initial cluster centers. Then every point of the dataset is assigned to the closest cluster, resulting in a new data partition. This is performed by computing an Euclidean distance between the given point and every cluster center. Next the cluster centers are computed again based on the current partition. These two last steps are repeated until convergence (*i.e.* no data point changes its association with a cluster).

4.2 Multidimensional Hidden Markov Models

Hidden Markov Models are dedicated to statistical modelling of a process that varies in time. So they are particularly adapted to audio analysis [13]. However, when dealing with multidimensional observation data, it is possible to use an ex-

tension of HMM called Multidimensional Hidden Markov Models [17,18]. In this section we will briefly recall main HMM concepts and present Multidimensional HMM.

4.2.1 Hidden Markov Models

An Hidden Markov Model can be associated with a set of random variables representing states of a discrete stochastic process composed of an hidden and of an observable part. It is characterized by a set $\mathrm{S} = \{\mathrm{S}_1, \ldots, \mathrm{S}_\mathrm{N}\}$ of hidden states of the HMM, a set $\mathrm{V} = \{\mathrm{V}_1, \ldots, \mathrm{V}_\mathrm{M}\}$ of symbols which can be generated by the HMM, a probability distribution matrix B of symbol generation, a probability distribution matrix A of transitions between states and a probability distribution vector Π of the initial state. An HMM can then be modelled by the triple $\lambda = \{\mathrm{A}, \mathrm{B}, \Pi\}$.

The segmentation method proposed in this paper is based on ergodic HMM. Learning and recognition will be respectively performed using Baum-Welch [14] and Forward [13] algorithms. We will use one HMM for each class of the audio segments.

4.2.2 Multidimensional HMM

Multidimensional HMM are particularly useful when dealing with data composed of several independent components. Indeed the HMM will generate R different symbols at a given time t and not only one anymore. The difference with using observation vectors consists of the fact that vectors components can have different natures. Here the observation describes the evolution of R processes instead of the evolution of only one process represented in a R-D space with noise.

The HMM architecture is then modified. The model contains only one A state transition matrix but R matrices B, one for each of the simultaneously observable process. The HMM architecture is also characterized by the number R of processes linked with the Multidimensional HMM, the set $\mathrm{V}^\mathrm{r} = \{\mathrm{V}^\mathrm{r}_1, \ldots, \mathrm{V}^\mathrm{r}_{\mathrm{M}^\mathrm{r}}\}$ of symbols linked to the process P_r, the probability distribution matrix B_r of generation of symbols linked to process P_r, the set $\mathrm{V} = \{\mathrm{V}^1, \ldots, \mathrm{V}^\mathrm{R}\}$ of dictionaries of symbols V^r linked to each process, and finally the set $\mathrm{B} = \{\mathrm{B}_1, \ldots, \mathrm{B}_\mathrm{R}\}$ of probability distribution matrices.

The method we propose in this paper involves Baum-Welch and Forward algorithms. More precisely, we use modified algorithms in order to deal with Multidimensional HMM.

These tools are used in the process detailed in next section.

4.3 Main description of the process

As any other method, the recognition method we have elaborated is relying on a set of features. We think it is important to get information at different observation scales as the type of information is not the same. But indeed it may be difficult to manage these levels only with one HMM model. So we have decided to consider two observation levels in the audio track. The signal or audio track (noted AT) is divided into segments (noted AS) that are themselves divided into frames (noted AF). Then features concern either the segment level or the frame level. At the segment level a K-Mean classifier will allow to cluster the segments into K virtual classes. Within classes the variability of the signal with respect to the features used is decreased. These classes have no real link with the predefined C labels we want to associate with each segment. Then we specialise the classifiers to only one of the virtual class to label the segments. For this task frame features are used within $(C \times K)$ Multidimensional HMM. Diagram of the proposed approach is given in figure 4.

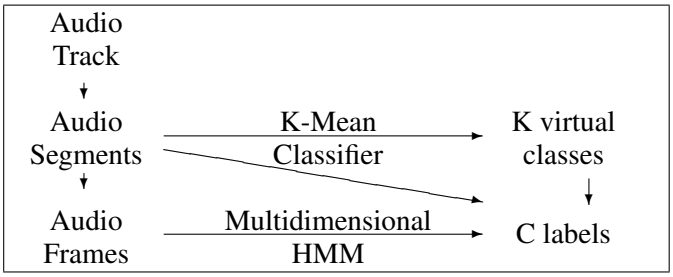


Fig. 4. Diagram of the two step strategy.

Now we are to precise the features we have made use of as well as the learning procedure and the recognition one.

4.4 Audio features

We perform audio segmentation using a set of 12 features. Some are taken from [15] and have been recalled previously: non-silence ratio (NSR), volume standard deviation (VSTD), standard deviation of zero crossing rate (ZSTD), volume dynamic range (VDR), standard deviation of pitch period (PSTD), smooth pitch ratio (SPR), non-pitch ratio (NPR), frequency centroid (FC), frequency bandwidth (FB), 4 Hz modulation energy (4ME), and energy ratio of subband 1-3 (ERSB1-3). We also use a feature representing the cepstrum (CBF: Cepstrum-Based Feature). Among these features, five are related to an audio segment (NSR, SPR, NPR, VDR, 4ME) and will be used in a first segmentation step based on the K-Mean classifier. Other features (VSTD, ZSTD, PSTD, FC, FB, ERSB, CBF) are related to a frame and will be analysed through Multidimensional HMM.

4.5 Learning step

The learning is performed in a supervised way and is based on three successive steps which will be described below: feature computation, K-Mean classification, and HMM creation. Let K and C respectively represent the temporary and final numbers of classes in our segmentation.

In order to analyse an audio segment AS, we first have to compute its related features. In a preprocessing step the audio segments are divided into N shifted frames noted AF_1 to AF_N.

The K-Mean classifier can then be applied using only the segment-based features in order to obtain a first classification into K clusters or virtual classes (figure 5a). The parameter K has to be set *a priori* and has an influence on the number of Multidimensional HMM to be built then. Indeed the use of a K-Mean classifier in the Multidimensional HMM creation process allows to increase the quality of the labelling process.

Once the audio segments have been classified into one of the K clusters, they will be involved in the learning process of the appropriate Multidimensional HMM. This learning process is achieved using only data in one of the K clusters, and for each of the K clusters, C HMM are elaborated, as shown in figure 5b where p audio segments $AS^{j}_{V_i}$ are used to create HMM^{i}_{1} to HMM^{i}_{c} linked to virtual class AS_{V_i}. The learning algorithm used is the multidimensional Baum-Welch algorithm and of course uses the frame based features that allow to describe the segments by an observation whose length is the number N of frames. The proposed method needs $K \times C$ Multidimensional HMM in order to perform segmentation in C classes.

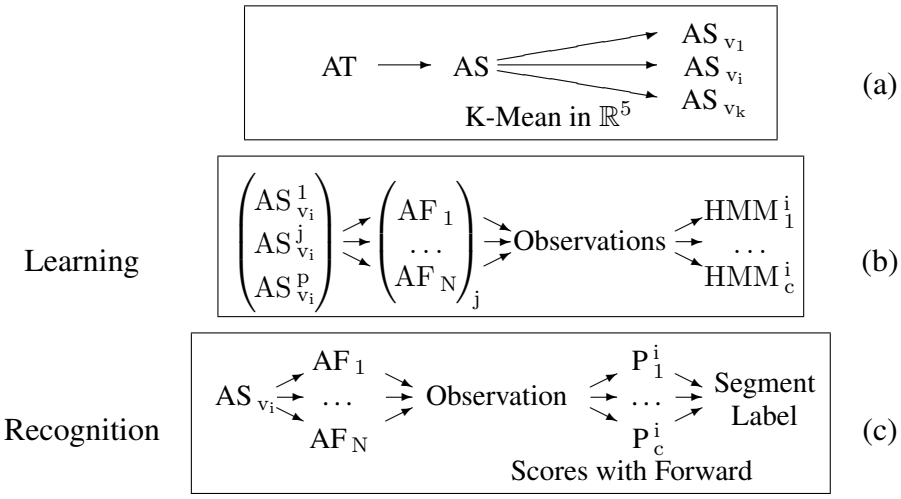


Fig. 5. Description of successive steps: (a) classification into virtual classes followed by (b) the learning phase or (c) the recognition phase.

In order to avoid the Baum-Welch algorithm (as any gradient-like optimisation technique) being trapped by local extrema, we do not use it directly but rather make use of the GHOSP algorithm [18] which aims to find optimal HMM involving Baum-Welch algorithm as a step of a global evolutionary optimisation process (since genetic algorithms are known for their robustness). GHOSP (Genetic Hybrid Optimization & Search of Parameters) is able to automatically find the optimal HMM content, i.e. the triple $\lambda = \{\mathrm{A}, \mathrm{B}, \Pi\}$ but also the number N of hidden states, whatever the kind of HMM is considered (here we use ergodic multidimensional HMM). It relies on both Baum-Welch algorithm (for HMM optimisation) and Forward algorithm (as an evaluation criterion). For more details on this algorithm, the reader can refer to [18].

4.6 Recognition step

The goal of the recognition step is to classify every audio segment into one of the C classes. This step involves similar processing as learning procedure. First audio segment features are computed. Then each segment is classified using K-Mean classifier in one of the K classes (see figure 5a). We then process the Forward algorithm on the C Multidimensional HMM $\mathrm{HMM}_1^{\mathrm{i}}$ to $\mathrm{HMM}_{\mathrm{c}}^{\mathrm{i}}$ linked with the selected class $\mathrm{AS}_{\mathrm{V_i}}$. The audio segment is finally labeled into the class for which the Forward algorithm gives the highest score $\mathrm{P}_{\mathrm{max}}^{\mathrm{i}}$, as shown in figure 5c.

4.7 Results

The method proposed here is a general approach which can perform segmentation of any kind of audio track. We have also applied it to football audio broadcast tracks. The goal was to classify every audio segment into one of the three following classes ($\mathrm{C} = 3$): referee whistle, crowd, and speaker voice. Once the classification of an audio segment has been obtained, it is then possible to analyse it with an adequate processing (*e.g.* speech recognition is performed only on "speaker voice" audio segments).

Duration of audio segments analysed in our application is equal to one second. Frame-based features are computed by dividing the audio segment into frames containing 1024 samples. Two successive frames are shifted of 512 samples in order to keep continuity property. We have made use of K-Mean classifier involving 3 clusters ($\mathrm{K} = 3$) so the proposed method needs 9 Multidimensional HMM.

The trials have been achieved on a total of 616 audio segments (21 for referee whistle, 148 for crowd, and 447 for speaker voice) extracted from different videos. Recognition rates are given in table 4. In comparison with traditional HMM-based

approaches as the one presented previously, we can notice that the recall rate is 10 to 15 % higher for similar precision rate.

Table 4
Results of a 3 class segmentation using the two step approach.

Class	K-Mean + M-HMM		HMM (table 3)	
	Recall	Precision	Recall	Precision
Whistle	95 %	86 %	88 %	88 %
Crowd	75 %	86 %	61 %	87 %
Speaker	95 %	90 %	77 %	90 %

5 Conclusion

In this paper, we have tackled the problem of audio data microsegmentation, *i.e.* segmentation of audio sequences characterized by a short duration (typically between 0.5 and 1 second). We have first presented some simple approaches based on spectrum or amplitude analysis, cepstrum analysis and classical hidden Markov models. Their limitations justify the need for more complex techniques. So we propose to involve a K-Mean classifier before using Multidimensional Hidden Markov Models. The use of the K-Mean classifier helps us to get recognition of better quality whereas the use of Multidimensional HMM allows to deal with data composed of several independent features. The features have been computed at different scale levels to ensure complementarity of the analysis. Results have shown that this method outperforms classical HMM-based approach.

Future work includes tests on other audio features and other classifiers (especially unsupervised algorithms) in order to confirm the improvement of recognition rates. An implementation of the method on a multiprocessor workstation is also considered to obtain a real time process. The proposed method will be integrated in a football event recognition system as a preprocessing step for audio data analysis. Finally other applications will be developed.

References

[1] Y. Wang, Z. Liu, J. Huang, Multimedia content analysis using both audio and visual clues, IEEE Signal Processing Magazine 17 (6) (2000) 12–36.

[2] M. Kermit, A. Eide, Audio signal identification via pattern capture and template matching, Pattern Recognition Letters 21 (2000) 269–275.

[3] T. Zhang, C. Kuo, Heuristic approach for generic audio data segmentation and annotation, in: ACM International Conference on Multimedia, Vol. 1, Orlando, USA, 1999, pp. 67–76.

[4] D. Li, I. Sethi, N. Dimitrova, T. McGee, Classification of general audio data for content-based retrieval, Pattern Recognition Letters 22 (2001) 533–544.

[5] S. Pfeiffer, S. Fischer, W. Effelsberg, Automatic audio content analysis, in: ACM International Conference on Multimedia, Boston, USA, 1996, pp. 21–30.

[6] E. Wold, T. Blum, D. Keislar, J. Wheaton, Classification, search, and retrieval of audio, in: CRC Handbook of Multimedia Computing, CRC Press, 1999.

[7] S. Li, Content-based classification and retrieval of audio using the nearest feature line method, IEEE Transactions on Speech and Audio Processing 8 (5) (2000) 619–625.

[8] E. Bocchieri, J. Wilpon, Discriminative feature selection for speech recognition, Computer Speech & Language 7 (3) (1993) 229–246.

[9] G. Tzanetakis, P. Cook, Sound analysis using mpeg compressed audio, in: International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000.

[10] D. Kimber, L. Wilcox, Acoustic segmentation for audio browsers, in: Interface Conference, Sydney, Australia, 1996.

[11] E. Battle, P. Cano, Automatic segmentation for music classification using competitive hidden markov models, in: International Symposium on Music Information Retrieval, Plymouth, USA, 2000.

[12] H. Hirsch, Hmm adaptation for applications in telecommunication, Speech Communication 34 (2001) 127–139.

[13] L. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[14] L. Baum, J. Eagon, An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology, Bulletin of American Society 73 (1967) 360–363.

[15] Z. Liu, Y. Wang, T. Chen, Audio feature extraction and analysis for scene segmentation and classification, Journal of VLSI Signal Processing for Signal, Image, and Video Technology 20 (1/2) (1998) 61–79.

[16] K. Fukunaga, Statistical Pattern Recognition, Academic Press, 1990.

[17] J. Yang, Y. Xu, C. Chen, Hidden markov model approach to skill learning and its application to telerobotics, IEEE Transactions on Robotics and Automation 10 (5) (1994) 621–631.

[18] T. Brouard, Algorithmes hybrides d'apprentissage de chaînes de markov cachées : Conception et applications à la reconnaissance de formes (in french), Phd dissertation, Universiy of Tours, France (January 1999).

Biography of the authors

Sébastien Lefèvre received in 1999 the M.S and Eng. degrees from the University of Technology of Compiègne in Computering Engineering, and in 2002 the Ph.D. degree from the University of Tours in Computer Sciences. He is currently an Assistant Professor in the Department of Computer Sciences and the LSIIT, University Louis Pasteur, Strasbourg. From 1999 to 2002 he was with AtosOrigin as a Research and Development Engineer. In 2003 he was with the Polytechnical School of the University of Tours as an Assistant Professor. His research interests are in image and video processing, multimedia analysis and indexing, and mathematical morphology.

After studying in Ecole Normale Supérieure graduation in Mathematics, Nicole Vincent received a Ph.D. in Computer Sciences in 1988 from Insa de Lyon. She became full Professor in University of Tours in 1996 and moved Professor in University Paris Descartes since 2003. There she heads the Center of Research in Computer Science (CRIP5) and the team Systèmes Intelligents de Perception (SIP). She is specialised in pattern recognition, signal and image processing and video analysis.