A Two Level Classifier Process for Audio Segmentation

S. Lefèvre[1,2] B. Maillard[1] N. Vincent[1]

[1] Laboratoire d'Informatique
E3i / Université de Tours
64 avenue Portalis - 37200 Tours - FRANCE

[2] AtosOrigin
19 rue de la Vallée Maillard
BP1311 - 41013 Blois Cedex - FRANCE

E-mail: {lefevre,vincent}@univ-tours.fr

Abstract

We are dealing in this paper with audio segmentation. We propose a two level segmentation process that enables the audio tracks to be sampled in short sequences which are classified into several classes. 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-Means 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-Means 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 Model dedicated to multicomponents data. They are particularly adapted in our case where each audio segment can be characterized by several features of different nature.

1. Introduction

Analysis and classification of audio data is an important task in many applications, such as speech recognition or content-based indexing. 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 litterature 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 in 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 enough information in the global interpretation task of the video sequence. Besides more specific analysis can be later performed on each of the labeled part of the audio sequence.

The most important difficulty to solve such a problem is 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. 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 the mathematical tools we are to use in our method will be presented before we detail both step of our audio segmentation method. Finally results on football broadcast tracks will be presented and commented.

2. Audio data processing

Segmentation and classification of audio data has been studied by many researchers. It can be seen as a pattern recognition problem where two issues have to be solved: choice of the classifier and selection of audio features. Li et al [8] studied a total of 143 features to determine their discrimination capability. Pfeiffer et al describe in [11] basic features used in audio analysis. Wold et al [15] analyse and compare audio features for content-based audio indexing purpose. Li [9] performs experiments to compare various classification methods and feature sets. Bocchieri and Wilpon [3] 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 [13] with MPEG audio.

Once the classifier and the features have been selected,

we have to determine if the analysis of audio sequences will be performed on a global or a local basis. In the first case, the goal is to classify complete audio sequences, as in the approach by Wang et al [14] to 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 [6]. Event detection can even be performed in real time [17].

Several researchers have proposed to use HMM to perform audio analysis. Kimber and Wilcox [7] 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 [1] propose to use Competitive HMM instead of traditional HMM in case of unsupervised training. Finally Hirsch [5] uses an adaptative HMM architecture in order to deal with audio signal from telecommunications.

In the method we propose we are using classical tools. We recall them in next session, specially HMM for which we use an original model that enables to consider each component of the data in a some what independent way.

3. Theoretical Tools

3.1. K-Means Classifier

K-Means classifier is a tool widely used in pattern recognition [4]. 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).

3.2. 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 [12]. However, when dealing with multidimensional observation data, it is possible to use an extension of HMM called Multidimensional Hidden Markov Models [16]. In this section we will briefly recall main HMM concepts and present Multidimensional HMM.

Hidden Markov Models

An Hidden Markov Model is 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, a probability distribution vector Π of the initial state. An HMM can then be modelled by the triplet $\lambda = \{\mathrm{A}, \mathrm{B}, \Pi\}$.

The segmentation method proposed in this paper is based on ergodic HMM. Learning and recognition will be respectively performed using Baum-Welch [2] and Forward [12] algorithms. We will use one HMM for each class of the audio segmentation.

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 in 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 an A state transition matrix but R matrixes B, one for each of the simultaneously observable process. The HMM architecure 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 dictionnaries 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 matrixes.

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 we detail in next section.

4. Combination of both approaches

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 in a 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). So features concern either the segment level or the frame level. At the segment level a K-Means 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 classifiers are used on each virtual class to label the segments. For this task frame features are used within C $\times$ K HMM. Diagram of the proposed approach is given in figure 1.

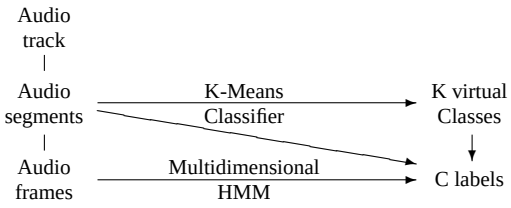

Figure 1. Diagram of the proposed two level approach.

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

4.1. Audio features

We perform audio segmentation using a set of 12 features. Some are taken from [10]: 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 (ESRB1-3). 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-Means classifier. Other features (VSTD, ZSTD, PSTD, FC, FB, ERSB, CBF) are related to a frame and will be analysed through Multidimensional HMM.

4.2. 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-Means classification, and HMM creation. Let K and C respectively represent the temporary and final number 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 several shifted frames noted AF_1 to AF_N.

The K-Means 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 2a). 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-Means 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, it 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 2b 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 whoose length is the number of frames. The proposed method needs K $\times$ C Multidimensional HMM in order to perform segmentation in C classes.

4.3. 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-Means classifier in one of the K classes (see figure 2a). We then process the Forward algorithm on the C Multidimensional HMM HMM^i_1 to HMM^i_c linked with the selected class AS_{V_i}. The audio segment is finally labeled into the class for which the Forward algorithm gives the highest probability P^i_{max}, as shown in figure 2c.

5. Results

The method proposed in this paper has been applied on football audio broadcast tracks. The goal was to classify every audio segment into one of the three following classes (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 or half a 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-Means classifier involving 3 clusters ($K = 3$) 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 1. In comparison with traditional HMM-based approaches, the recall rate is 10 to 15 % higher for similar precision rate.

Class	Recall	Precision
Whistle	95 %	86 %
Crowd	75 %	86 %
Speaker	95 %	90 %

Table 1. Results of 3 classes segmentation.

6. Conclusion

In this paper, we proposed a method for microsegmentation of audio sequences. We combine a K-Means classifier with Hidden Markov Models in order to analyse audio segment using several audio features based either on segment or frame. The use of the K-Means classifier helps us to get recognition of better quality whereas the use of Multidimensional HMM allows to deal with data composed of several independant features. 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.

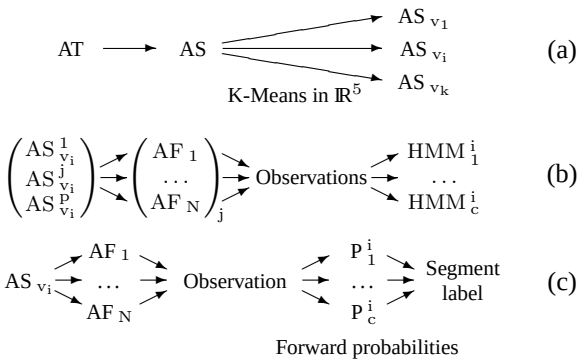

Figure 2. Description of successive steps: (a) classification into virtual classes followed by either (b) learning or (c) recognition.

References

[1] E. Battle and P. Cano. Automatic segmentation for music classification using competitive hidden markov models. In *Int. Symp. on Music Information Retrieval*, Plymouth, MA, Oct. 2000.

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

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

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

[5] H. Hirsch. Hmm adapatation for applications in telecommunication. *Speech Communication*, 34:127–139, 2001.

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

[7] D. Kimber and L. Wilcox. Acoustic segmentation for audio browsers. In *Interface Conf.*, Sydney, Australia, Jul. 1996.

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

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

[10] Z. Liu, Y. Wang, and 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):61–79, Oct. 1998.

[11] S. Pfeiffer, S. Fischer, and W. Effelsberg. Automatic audio content analysis. In *ACM Int. Conf. on Multimedia*, Boston, MA, Nov. 1996.

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

[13] G. Tzanetakis and P. Cook. Sound analysis using mpeg compressed audio. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, Jun. 2000.

[14] Y. Wang, Z. Liu, and J. Huang. Multimedia content analysis. *IEEE Signal Processing Mag.*, pages 12–36, Nov. 2000.

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

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

[17] T. Zhang and C. Kuo. Heuristic approach for generic audio data segmentation and annotation. In *ACM Int. Conf. on Multimedia*, volume 1, pages 67–76, 1999.