3 CLASSES SEGMENTATION FOR ANALYSIS OF FOOTBALL AUDIO SEQUENCES

S. Lefèvre, B. Maillard, N. Vincent*

Laboratoire d'Informatique
E3i / Université de Tours
64 avenue Portalis
37200 Tours - FRANCE
{lefevre,vincent}@univ-tours.fr

Abstract: In this paper we are dealing with segmentation of audio data in order to analyse football audio/video sequences. Audio data is divided into short sequences (typically with duration of one or half a second) which will be classified into several classes (speaker, crowd and referee whistle). Every sequence can then be further analysed depending on the class it belongs to. In order to segment audio data, several methods are presented. First simple techniques are reviewed for segmentation in two classes. From the limitations of these approaches, a method based on cepstral analysis is detailled. Next we present two more complex methods dealing with 3 classes segmentation. The first one is based on hidden Markov models whereas the second one is a combination of a C-Mean classifier and multidimensional hidden Markov models.

1. INTRODUCTION

Nowadays multimedia data represents a huge quantity of information. So manual indexing of these data is not possible anymore and systems are needed to perform automatic content-based multimedia indexing [1]. Most of the time multimedia indexing results from analysis of image sequences. Several authors have proposed to take also into account audio information [2, 3].

We are concerned with football video sequences indexing. Analysis of the audio track of a video sequence can help to detect predefined events (as the modification of the score). In order to perform audio analysis, we have chosen to first classify audio segments into speaker, crowd or referee whistle. Contrary to global segmentation as in [3], the segmentation is here performed on a local basis (sequence duration is less than one second) as in the work from Kermit and Eide [4]. Audio segments are then analysed depending on the class they belong to. In this paper we focus on the problem of segmentation of audio data into several classes.

First we will present two simple techniques used for 2 classes segmentation. These are respectively based on analysis of the frequency and of the amplitude of the audio signal. We will then show the limitations of these simple approaches and present a method based on the cepstral analysis. In a second part we will deal with 3 classes segmentation which can be achieved using more complex approaches. We will present two methods, which are respectively based on classical hidden Markov models and on a new combination of a C-Mean classifier and multidimensional hidden Markov models.

2. 2 CLASSES SEGMENTATION

As far as the signal that we want to recognize is concerned, frequency or amplitude of the signal are 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 detailled.

2.1. Simple approaches

Before proposing a complex method for audio segmentation, we have to check and show the limitations of simple approaches. We present here two methods for segmentation of audio data in 2 classes. These methods are respectively based on analysis of frequency and amplitude of the audio signal to segment audio sequences into whistle/non whistle and speaker/crowd.

In order to detect audio segments with referee whistle, we start from the assumption the sound produced by a whistle is composed of two or three frequencies belonging to interval $[3700, 4300]$ Hz. An example of a spectogram 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 spectogram 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 misdetection of the speaker voice as the referee whistle. Figure 1 (b) containing the spectogram of a speaker audio segment illustrates this problem.

Besides audio segmentation into 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 av-

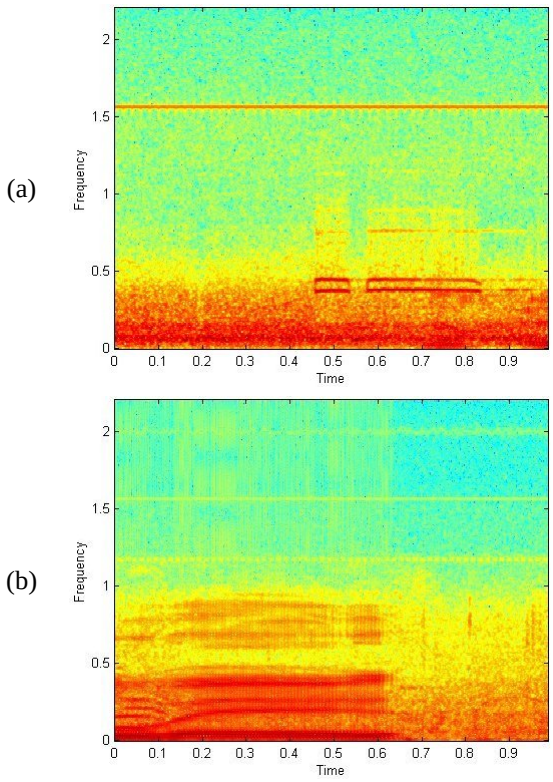

Fig. 1. Spectograms of audio signals corresponding to whistle referee (a) and speaker voice (b).

erage amplitude is not equal for both classes. So in a single way, it is possible to segment audio data into speaker and crowd depending on the average of 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 adaptative threshold is not necessary and a fixed threshold is used instead. A learning strategy is used to determine this threshold. Results are presented in table 1 where recall and precision are defined for a given class as the ratio of correctly classified sequences by 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 the quality of the method is not sufficient to segment correctly our audio data into speaker or crowd.

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 broadcast relative to several football games.

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.

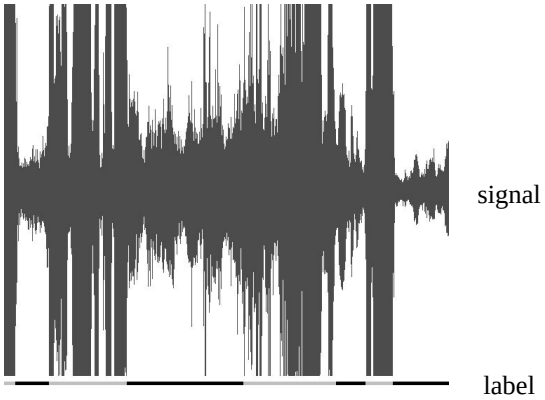

Fig. 2. Audio signal containing segments corresponding either to speaker (grey label) either to crowd (black label).

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

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

2.2. 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. A cepstrogram is a 3-D graphical representation of the audio signal based on the cepstrum computation. Figures 3 and 4 show respectively 2-D projections from cepstrograms of audio segments produced by crowd and speaker.

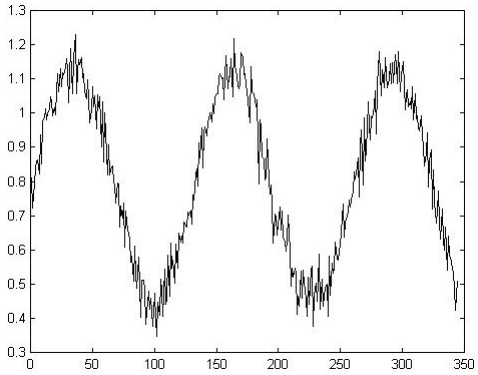

Fig. 3. 2-D projection from the cepstrogram of an audio signal corresponding to crowd.

From these figures we can understand the crowd is represented by a sinusoidal curve contrary to the speaker voice. The proposed segmentation method is based on this fact. First, curve frequency and phase are estimated. Then we compute an euclidean distance between theorical sinusoidal curve and observed signal. If the distance

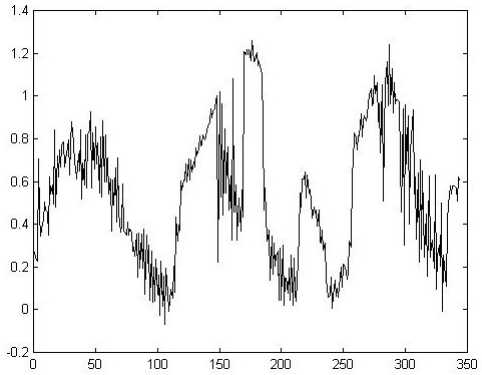

Fig. 4. 2-D projection from the cepstrogram of an audio signal corresponding to speaker voice.

obtained is below a threshold, audio segment is classified into crowd. Otherwise it is classified into speaker voice. 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.

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

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

We can see the quality of this method is higher than the simple approach based on amplitude analysis. However the quality rate has to be further increased, which can be done by combining several methods. This will be presented in the next section where we deal with 3 classes segmentation.

3. 3 CLASSES SEGMENTATION

In the previous section several methods for segmentation of audio data in two classes were presented. It has been shown that quality of these methods was not good enough to be used alone. So in this section we deal with more complex methods dedicated to segmentation of audio sequences in 3 classes. The first method is based on hidden Markov models whereas the second one is an original combination of a C-Mean classifier and multidimensional hidden Markov models.

3.1. Hidden Markov models

Hidden Markov models (HMM) are one of the most used tools in speech analysis and processing. A good introduction to these models can be found in [5]. 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. Learning and recognition are respectively performed using the well-known Baum-Welch [6] 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 consist of a set of features successfully explored in [7]. 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 (ESRB1-3). 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 these 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 will be shifted of 512 samples. Table 3 shows results of 3 classes segmentation. Quality rates are better than with simple approaches reviewed in previous section.

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

Table 3. Results of 3 classes segmentation based on hidden Markov models.

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 C-Mean classifier. The method which also includes a cepstral analysis will be described in next section.

3.2. Combination of a C-Mean classifier with multidimensional hidden Markov models

In order to solve problems where observation data is multidimensional, it is possible to use an extension of HMM called multidimensional hidden Markov models (noted M-HMM) [9]. This model is particularly adapted to the segmentation of audio data using several features. The method proposed here consists of two successive steps based on a C-Mean classifier and on multidimensional hidden Markov models.

The C-Mean method is used here to classify every audio segment into several classes so that the variation within the classes is less important. The classification is based on 5 features (NSR, SPR, NPR, VDR, 4ME) related to an audio segment. Then within every class and for each of the 3 possible states (whistle, crowd, speaker) a M-HMM is built. The observation data for the M-HMM is composed of 7 features for every frame : VSTD, ZSTD,

PSTD, FC, FB, ERSB, and a feature representing the cepstrum. Diagram of the proposed approach is given in figure 5.

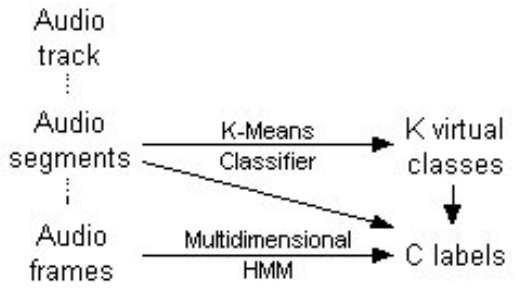

Fig. 5. Diagram of the proposed approach combining a C-Mean classifier and multidimensional HMM.

The use of a C-Mean classifier in the M-HMM creation process allows to increase the quality of the M-HMM built. Tests have been performed to determine the optimal number of classes in the C-Mean classifier. Classification in 3 classes gives the best results. So here we need $3 \times 3 = 9$ M-HMM contrary to the previous method where only 3 HMM where necessary.

Every audio segment involved in the learning process will be classified using the C-Mean algorithm. Depending on the result of this classification and on the nature of the audio segment (whistle, crowd or speaker), it will be used in learning process for the appropriate M-HMM.

In order to analyse an audio segment we first classify it using the C-Mean algorithm. Then we process the Forward algorithm on the 3 M-HMM corresponding to the selected class. We finally label the audio segment into the class whistle, crowd or speaker for which the Forward algorithm gives the highest probability. Results given in table 4 show the accuracy of this method.

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

Table 4. Results of 3 classes segmentation based on combination of a C-Mean classifier and multidimensional hidden Markov models.

4. CONCLUSION

In this paper we have been dealing with segmentation of audio sequences characterized by a short duration (typically between 0.5 and 1 second). From the limitations of several simple approaches reviewed in a first part, some more complex methods were presented. The first one is based on hidden Markov model whereas the second one is a combination of a C-Mean classifier and multidimensional hidden Markov models involving a feature linked with cepstral analysis. Some tests have been performed to show that the method based on an original combination of the C-Mean classifier and multidimensional hidden Markov models outperforms the other methods.

Future work includes tests on other audio features (as those reviewed in [10]) in order to increase the recognition rates. An implementation of the last algorithm on a multiprocessor workstation is also considered to obtain a real time process. Finally the proposed method will be integrated in a football event recognition system and be used as a preprocessing step for audio data analysis.

REFERENCES

[1] S.W. Smoliar and H. Zhang, "Content-based video indexing and retrieval," *IEEE Multimedia*, vol. 1, no. 2, pp. 62–72, Summer 1994.

[2] J.S. Boreczky and L.D. Wilcox, "A hidden markov model framework for video segmentation using audio and image features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seatlle, WA, May 1998, vol. 6, pp. 3741–3744.

[3] Y. Wang, Z. Liu, and J.C. Huang, "Multimedia content analysis," *IEEE Signal Processing Magazine*, pp. 12–36, November 2000.

[4] M. Kermit and A.J. Eide, "Audio signal identification via pattern capture and template matching," *Pattern Recognition Letters*, vol. 21, pp. 269–275, 2000.

[5] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

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

[7] 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*, vol. 20, no. 1/2, pp. 61–79, October 1998.

[8] E.L. Bocchieri and J.G. Wilpon, "Discriminative feature selection for speech recognition," *Computer Speech & Language*, vol. 7, no. 3, pp. 229–246, July 1993.

[9] J. Yang, Y. Xu, and C.S. Chen, "Hidden markov model approach to skill learning and its application to telerobotics," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 5, pp. 621–631, 1994.

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