MULTIPLE EVENTS TRACKING IN SOUND TRACKS

M. Betser, G. Gravier

IRISA (INRIA & CNRS) / METISS
Campus Universitaire de Beaulieu, Rennes
{mbetser, ggravier}@irisa.fr

ABSTRACT

Detecting and tracking broad sound classes in audio documents is an important step toward structuration. In the case of complex audio scenes, such as TV broadcast sound tracks, one problem is that several audio events may occur simultaneously. In this paper, we propose a two-step approach to detect superimposed events. The first step is a blind segmentation step, followed by an event detection step on each segment. In order to better evaluate the quality of the system, new performance measures have been introduced, more suited to the superimposed events detection task. We also extend the two-step approach with an equivalent Viterbi-based event detection approach.

1. INTRODUCTION

Two main approaches can be considered to detect and track events in an audio document. The first consists in segmenting the audio document into acoustically homogeneous segments, eventually clustered, before performing a detection step. The latter consists in labelling the various segments or clusters with the acoustic event(s) they correspond to or, alternately, in finding out the segments or clusters containing a particular event [1]. In this two step approach, segmentation is usually based on an information criterion, and classification typically makes use of statistical models whose parameters are estimated off-hand on a training corpus.

Another approach consists in using statistical models to jointly segment and classify the audio track. This last approach have been extensively used for speaker tracking tasks using Gaussian mixture models (GMM) [2].

In a previous work [3], it has been highlighted that most systems generally assume that no audio events occur simultaneously. However, most audio documents, and especially sports ones, cannot be described as a succession of isolated events. In order to completely describe those documents, we therefore need to be able to detect simultaneous audio events. Several methods were previously proposed in [3] to detect simultaneous events.

This paper extends our previous work and focuses on the two-step approach. We propose a detection step splitting the detection problem into independent class *vs* no-class problems within a maximum a posteriori (MAP) approach. Based on this event tracking or detection framework, we propose new performance measures which, in our sense, are better suited when simultaneous events are present. Experiments are carried out on a tennis video corpus. In particular, we investigate the influence of the segmentation step and approximations of the event prior probabilities. Finally, we extend the detection framework for joint segmentation and detection using hidden Markov models (HMM).

2. A MAXIMUM A POSTERIORI APPROACH FOR EVENT TRACKING

The aim of this work[1] is the detection, in an audio document, of sound events which may occur simultaneously. We focus on a two step detection system, which first computes a prior segmentation into homogeneous segments (homogeneous means same classes presents on all the segment), with an eventual clustering step in order to gather non-adjacent similar segments, and then classifies these segments, using GMMs.

In a previous work, several methods were proposed, the best one being the Viterbi approach, where models for superimposed events were computed from the isolated event ones by concatenation (concatenation method). A first version of a two-step method was also tested, but yield no improvements. The segmentation step used a Bayesian information criterion (BIC), with no clustering step. The detection step was achieved with a two hypothesis statistical test between the *present* and *not present* hypothesis, for each sound class. Class models were trained on segments where only the event considered is present while “anti-class” models were trained on segments that do not contain the event considered.

We describe here under the three steps, namely segmentation, clustering and detection, of our new approach.

[1] This work was carried out in the framework of the Domus Videum project founded by the French RNRT.

Segmentation and clustering

As previously, segmentation is based on a BIC information criterion. Results are presented on a three step algorithm: first there is a rough change detection with a growing window; then each change is adjusted more precisely; last, wrong changes are suppressed.

A two-step algorithm, using a BIC-based change indicator has also been tested, but as results were very similar, we only report on the three step approach.

The clustering algorithm used is a hierarchical clustering as described in [4]. Each cluster is initialized with a segment. At each step, a distance matrix between clusters is computed using a Kullback-Liebler (KL) similarity measure, considering all the possible merges. Two segments can be merged if their merging globally increases a BIC criterion. Therefore, the second step is to find the closest clusters verifying this criterion and to merge them.

When computing the KL similarities and the BIC criterion, as in the segmentation step, a single Gaussian is used to model the clusters.

Classification

One way to see the event detection problem is to consider each event as a source, which is emitting or not. The sound track can then be represented as a mixture of these sources, and each segment can be associated with a state represented by a random variable $X = \{X_i, i = 1..d\}$, where X_i equals one if event i is present or zero otherwise, and d is the number of events considered.

The MAP criterion is used to determine the most likely state $\hat{x}$ among all the possible states $x = \{x_i = (0,1), i = 1..d\}$,

$$\hat{x} = \arg\max_x P(X = x|y) \tag{1}$$

where y represents the data for one segment. The problem is to evaluate the observation likelihood $P(y|x)$ for each possible state, as well as the corresponding prior probabilities $P(x)$. If more than two events are present, it becomes impossible to compute all these models, because of the lack of training data. We therefore propose to hypothesize statistical independence of each event conditionally to data. The likelihood of an observation y can be computed for a state x as:

$$\begin{aligned} P(X = x|y) &= \prod_i P(X_i = x_i|y) \\ &\sim \prod_i P(y|X_i = x_i).P(X_i = x_i) \end{aligned} \tag{2}$$

In this case only two models, $P(y|X_i = 1)$ and $P(y|X_i = 0)$, are needed. It might be useful to note that in this formalism, $P(y|X_i = 1)$ represents all the data where the class i is present and is different from $P(y|X_i = 1, X_{k\neq i} = 0)$ which represents data where the class is alone, a fact which has been verified experimentally. The MAP criterion can be rewritten as:

$$\hat{x} = \arg\max_x \prod_{i,x_i=1} \frac{P(X_i = 1)}{P(X_i = 0)}.\frac{P(y|X_i = 1)}{P(y|X_i = 0)} \tag{3}$$

Equivalently, we can make independent decisions for each event by comparing the likelihood ratio $l_i = P(y|X_i = 1)/P(y|X_i = 0)$ to the threshold given by $\beta_i = P(X_i = 0)/P(X_i = 1)$. Clearly, if $l_i < \beta_i$, state $\hat{x}_i$ should be set to 0 in order to maximize (3).

In a previous work, a unique threshold had been used for all the classes, which corresponds to the case where the sources have the same prior probabilities. We experimented several approximations of the prior probabilities as described in section 4.1.

3. PERFORMANCE MEASURE

The point of view of this work is a *detection* point of view, where false alarm and false acceptance errors have an equal cost.

Let $T_i(1,1)$ be the total duration where event i was correctly detected, and $T_i(0,0)$ the total duration where the absence of event i was correctly detected. Similarly, $T_i(1,0)$ and $T_i(0,1)$ are the total time of false alarms, and miss detection respectively. Finally, let $T_i(1)$ and $T_i(0)$ denote the total amount of time in the reference where event i is respectively present or absent. Using the above notations, the classical performance measure figures used in the detection problems can be extended to multiple event tracking. We define the three following measures:

$$\begin{aligned} \%\text{corr} &= \frac{\sum_i T_i(1,1) + T_i(0,0)}{d.T} \\ \%\text{FA} &= \frac{\sum_i T_i(0,1)}{\sum_i T_i(1)} \\ \%\text{FR} &= \frac{\sum_i T_i(1,0)}{\sum_i T_i(0)} \end{aligned}$$

%corr is the average classification rate across all the events considered. It is easily verified that:

$$\%\text{corr} + \frac{\sum_i T_i(1)}{d.T}.\%\text{FA} + \frac{\sum_i T_i(0)}{d.T}.\%\text{FR} = 1$$

As our purpose is to detect superimposed events, we need also an indicator of the quality of multiple event detection. The multi-label segment recognition rate, %mcorr, is defined as the ratio between the duration of correctly recognized multiple label segments and the total duration of multiple segments in the ref. segmentation.

	prior1	*prior2*	*prior3*	*Ref.*	*Davis*	*Vit. 2*	*Vit. 1*
% corr	89	89	88	93	92	89	88
% FA	22	16	26	13	13	25	20
% FR	7	9	6	5	7	6	9
% mcorr	54	34	54	65	63	54	32

Tab. 1. The first five columns are results of the binary hypothesis testing method: the three kind of prior tested in section 2, an experience made on the reference segmentation and a test made only on the end of the coupe Davis match. 'Vit. 2' stands for the 2 states HMMs method described in section 5, and 'Vit. 1' recalls results for a Viterbi using model concatenation.

4. EXPERIMENTS

Experiments were carried out on a tennis video sound track corpus, composed of three tennis documents. The training corpus contains an entire game (Rolland Garros) plus the two first sets of a Davis cup game. The last set and three additional sets from Bercy tournament composes the test corpus. Our goal is to structure the sound tracks using four sound classes: *speech*, *music*, *applause* and *tennis*. The *tennis* class corresponds to tennis noises such as ball hits, player screams, etc.

A 64 component GMM with diagonal covariance matrices was estimated for each class. Cepstral coefficients plus first and second order derivatives are used to represent the signal. Results are summarized in table 1.

4.1. Approximation of the prior probabilities

In this first experiment, we compare several approximations of the prior probabilities, $P(X_i = 0)$ and $P(X_i = 1)$. These probabilities can be estimated on the training corpus (prior1). Assuming the sources have the same prior probabilities, an optimal threshold can be determined on the training corpus (prior2). One problem with these two approaches is that decisions are made independently for each event thus making it difficult to penalize states where many events are present. We therefore tried a prior giving in (1), an equal probability to global states with zero, one or two classes presents, and a zero probability to the others, which is approximately verified on the tennis corpus, but also on many other sound documents (prior3).

Best results are of course obtained for the first kind of priors, other priors slightly degrading the performances. However, when exact prior class probabilities are not available, approximations can be used without a dramatic drop of performance. It can be noticed that the three priors lead to different kind of errors. Depending on the type of error which need to be minimized, a different prior could be used.

Compared to our baseline Viterbi system (Vit. 1), the proposed approach leads to better results, especially in detecting superimposed events. If the decoding is made with same acoustic conditions as in the training corpus (Davis), performances are improved greatly. This suggest that a model adaptation on the test data could be useful.

4.2. Influence of the segmentation

The theoretical BIC threshold is 1, but as we can see in figure 1, this leads to large segments, larger than they should be (the average length in the reference is 4.99s). When decreasing the threshold, more segments are found. The segmentation leading to the best results has approximately the same number of segments as in the reference segmentation, with a BIC threshold of 0.6. This value has been retained for all the other experiments. Yet, there is still a drop in the performances compared to a perfect segmentation. This means that some of the changes are misplaced and that some events we want to detect are ignored. The missing events are still not detected for low values of BIC threshold.

The same problem occurs with the clustering algorithm. No improvement is noted after the clustering step: on one hand, the class estimation is globally better for clusters than for individual segments, but on the other hand, some segments with multiple sound classes will be placed in wrong clusters especially when there is a strong dominant class (typically when there is commentator speech over tennis).

The single Gaussian model used to compare segments, is a too rough approximation in the case of superimposed sound events. The classification rates obtained for a perfect segmentation (Ref.) suggest that well trained class models can better distinguish superimposed events, and it is therefore reasonable to think that integrating them in the segmentation process will help.

5. TWO STATES HMMS

One way to introduce class models during the segmentation phase is to use a Viterbi algorithm. In this algorithm segmentation and classification phases are mixed. In order to keep the class *vs* no-class representation which is the central point of this paper, we propose a Viterbi approach using two state models for each class.

Considering a HMM with all the possible global states (cf. figure 2(a)), the Viterbi algorithm ensures the result-

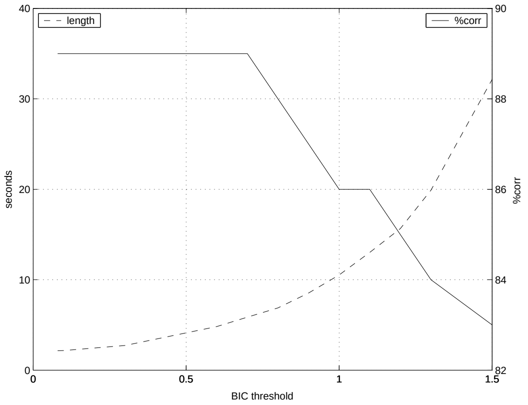

Fig. 1. The first curve (plain) is the %corr measure and the second curve (dots) is the average length of segments, both as functions of the BIC threshold.

ing path will be optimal in the maximum likelihood sense. If we assume the classes are independent, as in the second section, the likelihoods of a global state can be expressed as a product of the class and no-class likelihoods as in (2). From this, it can be easily shown that using this HMM is equivalent to using d smaller HMMs, one for each class, with two states (cf. figure 2(b)).

In summary, using a Viterbi algorithm on d class *vs* no-class HMMs, and concatenating their resulting segmentations leads to an optimal global segmentation under the hypothesis made in 2.

Experiments

The class *vs* no-class model used in the two state HMMs, are the same as the ones used in the hypothesis tests and the transition probabilities are estimated on the training corpus.

The results obtained for two states HMMs (Vit. 2), are approximately the same as those obtained with hypothesis testing. This shows that using the class *vs* no-class model independently to do the classification is effectively an interesting alternative to the blind segmentation. The two states HMM results are not better, because the Viterbi algorithm has also structural hypothesis which hurts the segmentation. The classification of frame depends only on the preceding one (order 1 Markovian Hypothesis), so it imperfectly represents the notion of segment. To avoid the HMMs to change state too often, we have used a classical technique which consists in penalizing the transition between two different states. Consequently, there will be inaccuracies on segment transitions and short segments will be absorbed by nearby bigger segments.

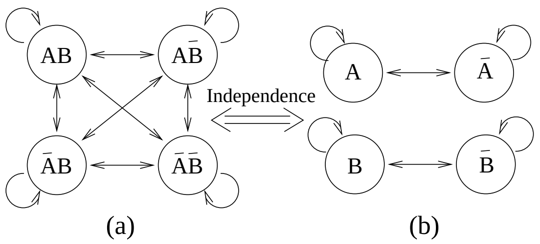

Fig. 2. For two independent sound classes A and B, equivalence between a HMM with all possible global states (a), and two binary HMMs (b)

6. CONCLUSION

In this paper, we have presented a generic system to extract audio information from video sound tracks, based on a two step decoding and able to detect simultaneous sound events. The segmentation step was performed with a BIC-based segmentation algorithm and the classification step using a MAP approach. We have shown that this system can yield good results compared with the best Viterbi-based system tested in a preceding paper, and that it could be improved by using a model-based segmentation instead of a “blind” BIC segmentation. Following this remark, a Viterbi segmentation/classification on two states HMMs has been proposed. Results were comparable to the binary hypothesis testing, even if the Viterbi has its own segmentation limitations. In a near future, we will investigate in order to improve the quality of the segmentation, and to improve blind adaptation of models to new data.

7. REFERENCES

[1] M. Cettolo. Segmentation, classification and clustering of an Italian broadcast news corpus. *Content-based Multimedia Information Access*, 2000.

[2] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, and S.J. Young. Segment generation and clustering in the HTK Broadcast news transcription system. *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[3] G. Gravier M. Betser and R. Gribonval. Extraction of information from video sound tracks - can we detect simultaneous events? In *International Workshop on Content Based Multimedia*, 2003.

[4] Bowen Zhou and John Hansen. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. In *Intl. Conf. Speech and Language Processing*, volume 3, pages 714–717, 2000.