EXTRACTION OF INFORMATION FROM VIDEO SOUND TRACKS - CAN WE DETECT SIMULTANEOUS EVENTS?

M. Betser, G. Gravier, R. Gribonval

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

ABSTRACT

Detecting and tracking broad sound classes in audio documents is an important step toward their structuration. In the case of complex audio scenes, such as the sound track of a TV broadcast, one problem is that several classes of sound maybe present simultaneously. It is therefore important to detect such superimposed events. Most methods would necessitate to estimate a model for each combination of sound classes that is to be detected, which is intractable in practice since it requires a lot of manual labelling. In this paper, we propose and compare several approaches to detect simultaneous events using only the models of the base classes we are interested in. Two main approaches are compared: model combination and binary hypothesis tests. The results show that the best results are obtained with the model combination approach.

1. INTRODUCTION

Detecting and tracking sound events in an audio document is a key step in any audio indexing system. In particular, the detection of broad sound classes, such as speech, music, noise or a particular speaker, in an audio document can provide valuable information on the structure of the document for further processing. A typical example is the processing of broadcast news audio documents where the parts of the document containing speech have to be detected before being transcribed by an automatic speech recognition system [7, 3]. Automatic audiovisual document processing, such as video abstracting or indexing, is also an emerging application where detecting and tracking broad sound classes in the video sound track is needed in order to help understanding the structure of the document [5, 8].

Two main approaches have been considered so far to detect and track sound events in an audio document. Both of them respects the standard indexing architecture shown in figure 1 . The first approach consists in segmenting the audio document into acoustically homogeneous segments which are then clustered to group together non adjacent similar segments. These two steps usually rely on the use of an information criterion (see, *e.g.*, [7]). Finally, the last step consists in labelling the various clusters obtained with the acoustic event they correspond to or, alternately, in finding out which cluster(s) correspond(s) to the acoustic events of interest. This is usually done using some statistical model of the sound classes to be detected.

The second approach consists in using directly such models, thus assuming that the classes are known a priori. The models, whose parameters are estimated on some training data are then used in a detection and segmentation task. Such approaches have been extensively used for speaker tracking tasks using speaker models based on Gaussian mixture models (GMM) [1, 6].

In the framework of video sound track indexing, one problem comes from the fact that it is common to have several sound classes simultaneously. For example, there may be music superimposed with speech or speech along with some other meaningful event such as applauses. Detecting such simultaneous events is therefore crucial. Note that in the broadcast news framework, detecting simultaneous events is also of interest as it may be used to trigger some specific processing such as source separation to separate speech from music.

In the approaches previously described, detecting simultaneous events requires a model for each possible class combination. This is irrealistic since it would require a huge amount of training material manually transcribed. Furthermore, introducing new sound classes in a system would be very costly since it would require to train new models for the combined classes.

In this paper, we therefore propose and compare several techniques to detect simultaneous events in video sound tracks using only models representing a single class such as speech, music or applause. Two different approaches are studied. The first one is based on Gaussian mixture models (GMM) where multi-class models are obtained by combining single class models. The second approach implements binary hypothesis statistical tests to detect whether an event is present or not. These different approaches are presented in the next section while experimental results, made on two

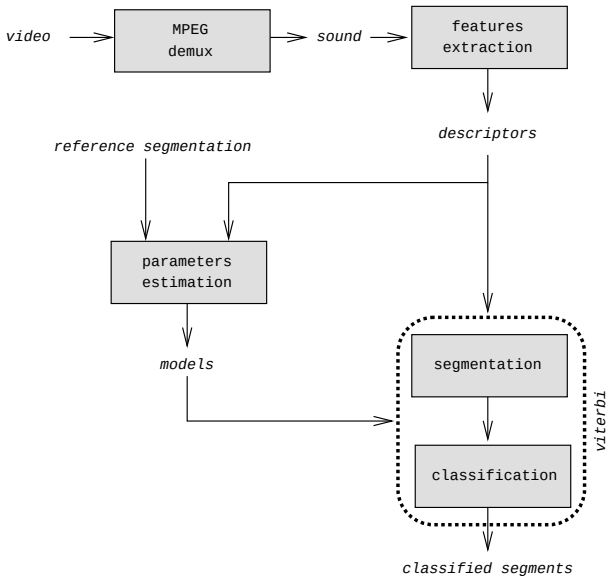

Fig. 1. Architecture of the system. The steps 'segmentation' and 'classification' can be combined into one if we use a viterbi decoding system.

different corpora, are presented and discussed in section 3.

2. METHODS

In this section, we review the theoretical aspects of the different segmentation methods studied.

2.1. Viterbi based system

The baseline system consists of a continuous density ergodic HMM where each state represents an audio class C_i. The state densities $p(x_t|C_i)$ are mixture of Gaussians whose parameters are estimated using features x_t belonging to class C_i. The segmentation is classically done by finding out the best state sequence $\hat{q}$, *i.e.*

$$\hat{q} = \arg\max_q \sum_t \ln p(x_t|q_t) + \ln p(q_t|q_{t-1}) + \gamma\delta(q_t - q_{t-1})$$

where $p(q_t|q_{t-1})$ is the transition probability and γ is a class insertion penalty factor which controls the average segment length, high values of γ being used to avoid short segments.

Clearly, it is not possible to detect simultaneous events with this approach unless there exists a model for the superimposed events in which case one can add a state to the HMM (see section below). Another way to extend this algorithm to detect simultaneous events consists in combining the N-best paths rather than considering the single best path $\hat{q}$. The path combination consists in classifying a frame x_t into all the classes found at time t in the N-best paths. This method can be extended to the use of a lattice of class hypotheses rather than the N-best paths.

2.2. Model combination

As mentioned previously, one way to detect multiple classes using the Viterbi approach is to add a state for each combination of classes. In this study, we limit ourselves to the combination of two and three classes. There are two reasons for this choice. First using combinations of more than three classes would lead to too complex HMMs. For example, if we describe a sound track with five classes using combinations of three classes it would lead to a HMM with already $C_5^1 + C_5^2 + C_5^3 = 25$ states. Second it was observed that the number of segments which exhibits more than two simultaneous events is very low. The problem is therefore to find out the state conditional densities $p(x_t|C_i, C_j)$ (also denoted p_{ij} for sake of simplicity) for the combined classes. The trivial approach which consists in estimating the parameters from training data is not feasible in practice since it would require a large amount of training data labeled manually. One solution consists in estimating the multiple class density by combining the two or three single class densities $p_i(x_t)$ and $p_j(x_t)$. Two different approaches to model combination were tested.

The first approach assumes, that when combining N classes, all these classes are present on the segment but not simultaneously. Therefore, assuming an equal prior probability for both classes, the density function can be written as

$$p_{i_1..i_N}(x_t) = \frac{1}{N}\left(\sum_k p_{i_k}(x_t)\right) \quad .$$

At the model level, the density $p(x_t|C_i, C_j)$ is a GMM which corresponds to the *concatenation* of the GMMs corresponding to the two classes C_i and C_j with a renormalization of the weights.

The second approach consists in combining the models assuming the features are additive. If we assume that the feature vectors are additive, *i.e.* the feature vectors corresponding to a mixture of the classes C_i and C_j is given by

$$x_t = x_t^{(i)} + x_t^{(j)} \quad ,$$

where $x_t^{(i)}$ is the feature for class C_i at *t*, then the probability density function of x_t is the *convolution* of the class density functions, given by

$$\begin{aligned} p_{ij}(x_t) &= \int p_i(x_t - y)\ p_j(y)\ dy \\ &= (p_i * p_j)(x_t) \quad . \end{aligned}$$

Since the single class densities are GMMs with K components g_{ik} and weights w_{ik} ($k = 1, \ldots, K$), we have

$$p_{ij}(x_t) = \sum_{k,l} w_{ik} w_{jl} (g_{ik} * g_{jl})(x_t)$$

where $g_{ik} * g_{jl}$ is also a Gaussian density whose parameters are given by

$$\mu_{ij,kl} = \mu_{ik} + \mu_{jl}$$
$$\sigma^2_{ij,kl} = \sigma^2_{ik} + \sigma^2_{jl} \quad .$$

In this last equation, μ_{ik} and σ_{ik} are the mean and variance respectively of $g_i k$. Using this approach, combination of three classes would lead to model with K^3 components, assuming each gaussian has K components, which requires far too much ressources and computation time. This is why we restricted the study to the combination of two classes.

The assumption of additivity of the feature vectors is clearly not verified by the cepstral coefficients but holds if the features considered are the energies (no module, no log) at the output of a filter-bank.

The concatenation and convolution methods yield GMMs with $2K$ (two classes concatenation), $3K$ (three classes concatenation) and K^2 (two classes convolution) components respectively, assuming the single class GMMs have K components. These models being more complex, they tend to always give better likelihoods than the single class models and are therefore privileged in the Viterbi decoding. To avoid this unbalance, it is possible to replace the likelihood by an information criterion which penalizes the likelihood according to the model complexity [2]. In this work, we used a penalty derived from the Akaike information criterion and given by

$$\alpha p_{\text{akaike}} = \frac{\alpha}{2} \left(K.(2D + 1) - 1 \right) \qquad (1)$$

where K is the number of Gaussians components in the model and D the dimension of the data. The weight α controls the contribution of the model complexity to the state conditional score.

2.3. Binary hypothesis tests

Finally, binary hypothesis statistical tests can also be used to detect simultaneous audio events using only models corresponding to single class events. This approach consists in obtaining a prior segmentation of the input signal, and to implement a binary test on each segment and for each class to determine whether the class is present in this segment or not. We have used two different segmentation methods: a Viterbi segmentation and a Bayesian information criterion (BIC) based segmentation. The first method uses a standard Viterbi system, but without considering the classification. The BIC criterion is a likelihood criterion penalized by the model complexity, *i.e.* the number of parameters of the model. For the model M with m parameters, the BIC criterion has the form:

$$BIC(M) = \log L(X, M) - \lambda \frac{m}{2} \log N_X \quad ,$$

where N_X is the number of acoustic vectors, X represents the data we want to model, $L(X, M)$ is the likelihood function for model M, and λ is a penality weight we have to adjust. Considering the sequence $X = x_1, ..., x_{N_X}$, the rupture detection, consists in an hypothesis test: the sequence is represented by only one class (hypothesis H_0) versus the sequence is represented by two classes (hypothesis H_1). The model for H_0 is supposed to be a single gaussian: $(x_1, ..., x_{N_X}) \sim N(\mu_X, \sigma_X)$ and for H_1 each sub sequence is also represented by a single gaussian: $(x_1, ..., x_i) \sim N(\mu_{X_1}, \sigma_{X_1})$ and $(x_{i+1}, ..., x_{N_X}) \sim N(\mu_{X_2}, \sigma_{X_2})$.This hypothesis test leads to the equation:

$$\Delta BIC(i) = -\frac{N_X}{2} \log |\Sigma_X| + \frac{N_{X_1}}{2} \log |\Sigma_{X_1}| + \frac{N_{X_2}}{2} \log |\Sigma_{X_2}| - \lambda \frac{1}{2} (p + \frac{1}{2} p(p+1)) \log N_X \quad .$$

A negative value for this equation means the two Gaussian model is better than the single model. Consequently, a rupture is detected on the segment X if

$$\{\min_i \Delta BIC(i)\} < 0 \quad .$$

The maximum likelihood estimate for the rupture on the segment X is given by:

$$\hat{t} = \arg\min_i \Delta BIC(i) \quad .$$

The BIC segmentation system used in this study is a three step algorithm as described in [4]: first there is a rough rupture detection with a large window, then we adjust each rupture found more precisely and finally we suppress bad ruptures. The weight λ is optimized on the training corpus.

After segmentation, a binary test is performed on each segment, the binary test principle consists in comparing the likelihood ratio between a class model C_i and the anti-class model $\bar{C}_i$ to a decision threshold β. The class is considered present in the segment if

$$\frac{1}{T} \ln \left(\prod_t \frac{p(x_t|C_i)}{p(x_t|\bar{C}_i)} \right) > \beta \quad .$$

In the experiments presented below, the anti-class model parameters were estimated for each class on all the data not labeled as belonging to that class. The threshold was set experimentally on the training corpus.

3. EXPERIMENTS

3.1. Corpus and performance measures

Results are given on two different corpora. The first one is a soundtrack of a tennis video. The first *set* is used as the training corpus while tests are carried out on the second one. The two *sets* were manually segmented according to the following sound classes: *speech*, *music*, *applause*, *tennis* and *noise*. The *tennis* class corresponds to tennis noises such as ball hits, player screams, etc. The *noise* class represents any other noise such as crowd noise. The training corpus duration is about 50 minutes and contains 500 segments. The test corpus has 700 segments over 40 minutes of signal, 40% of those segments being labeled with multiple classes.

The second corpus is a set of documentaries on scientific and music topics. All of them are about 13 minutes long. Two of them, a documentary on the cavitation and one on malgach music are used as a training corpus. The tests were made on the two others, one on wildlife and one on traditionnal music. All these documentaries have been segmented manually using three classes: *music*, *speech* and *background*. The *speech* class, here, corresponds to the commentator speech, and to the interviews. The *music* class corresponds to the jingles and background music, and the *background* class is everything else happening in the documentary. The training corpus has 96 segments of which 48% are multi segments while the test corpus has 81 segments with 49% of multi segments. Each corpora has very different caracteristics: the tennis corpus is a live match, so the sound quality is worse and it is less structured than documentaries. Overall scores should be better on the second corpus.

For each classes of each corpora, a 64 component GMM with diagonal covariance matrices was estimated using only the segments corresponding to the single class. Cepstral coefficients plus first and second order derivatives are used except for the model combination convolution approach where filter-bank energies are considered.

Though this work aims at detecting simultaneous classes, the performances are evaluated on the base of the single classes since this is what we are interested in at the far end of the process. Therefore, when comparing two segments that possibly have multiple labels, a correct match is counted for each label present in both segments. A substitution error is counted for every pair of non-matching labels after removing the correct matches. Finally, an insertion (resp. deletion) occurs for each additional label in the hypothesized (resp. reference) classes after removing the pairs counted as substitutions. Note that since the class *noise* carries no information for the the video indexing system, it is discarded from multiple labels when computing the performances.

In addition to the above performance measures, we also compute the rate of segments with multiple labels correctly recognized, the rate of segments with multiple labels detected as having a single label and the rate of segments with a single label detected as having multiple labels.

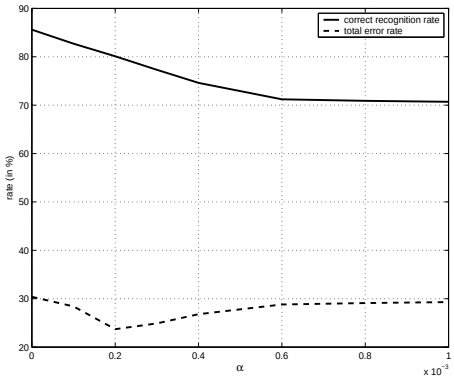

Fig. 2. Error and correct segmentation rates as a function of the model complexity penalty (tennis corpus).

For the binary test, in order to evaluate the quality of the segmentation, we used a recall/ precision measure, where

$$\text{recall} = \frac{\text{\# correct ruptures found}}{\text{\# reference ruptures}} \quad .$$

$$\text{precision} = \frac{\text{\# correct ruptures found}}{\text{\# ruptures found}} \quad .$$

A rupture found at time t is correct if there is a reference inside an interval $[t-\Delta, t+\Delta]$. We used a Δ of 0.5 seconds. Because of the unprecision of the manual segmentation a smaller value of Δ is unuseful. For our problem, the recall value is much more important than the precision: it is better to have the highest number of good ruptures possible because the binary classification cannot correct missing ruptures.

The baseline system performances are given in the first row of table 1.

3.2. Model concatenation

The effect of replacing the log-likelihood by an information criterion as given in (1) is illustrated here on the concatenation approach. Figure 2 shows the error and correct segmentation rates as a function of the weight α given to the penalty.

As can be seen, the introduction of a complexity penalty, leads to slightly better results for small values of α, the error rates decreasing rapidly with a small penalty. This is principally due to the fact that there is a lot of insertion errors without the penalty since the system tends to use the multi-class models instead of the simple ones. Indeed, the insertion rate is 22.7% without the penalty and drops down to

7.2% with $\alpha = 2e^{-4}$. However, in the same time, the rate of correctly recognized multi-class segments goes down from 58% without the penalty to 33%. Finally, the rate of single class segments recognized as multi-class segments goes from 38.4% for $\alpha = 0$ to 12% with $\alpha = 2e^{-4}$.

As the weight of the model complexity penalty increases, results rapidly degrades to the performance of the baseline system since the penalty takes precedence over the likelihoods in the search algorithm and no more multiclass segments are hypothesized.

Similar results were obtained on the documentary corpus with an optimal penalty weight $\alpha = 6^{-5}$. The difference of weight between the two corpora is due to the link between the class insertion penalty factor and the complexity penalty. Indeed as the class penalty is not the same for the two corpora (-50 for the documentaries and -150 for the tennis), their sensibility to the insertion of a second penality can differ greatly.

Three class model combination has also been experimented but, as expected, no improvement with respect to the two class model combination was observed because of the few three-multiple segments on the corpus.

These results clearly show that the concatenation approach is a good candidate to multi-class segments classification, outperforming significantly the baseline system. They also illustrate the benefit of using an information criterion instead of the likelihood.

3.3. Model convolution

As mentioned, the convolution approach was tested with filter-bank energy based features for which the additivity property holds. The feature therefore consists of the energies at the output of a 24 channel filter-bank. The baseline system with these features gives a correct classification rate of 67.1% with an error rate of 32.9% which is comparable to the performances of the baseline system using cepstral coefficients as given in table 1. Good results were obtained with the convolutive combination of models with a recognition rate of 74.4% and 18% of multi-class segments correctly detected.

However, one problem with the convolution approach is that the number of components of the combined models can be quite large (4096 in our case). It might therefore be interesting to reduce this number of components by selecting the components with the highest weights. Results are given in figure 3 for 64, 128, 256 and all the 4,096 Gaussians. They show that selecting some of the components severely degrades the results. In fact, no multi-class segment is correctly recognized with the truncated models (18% when all the components are kept). The reason may be that the weights of a combined model are the product of the weights in the single class models. In consequence, the weights have low and quite close values thus making the pruning quite arbitrary and inefficient.

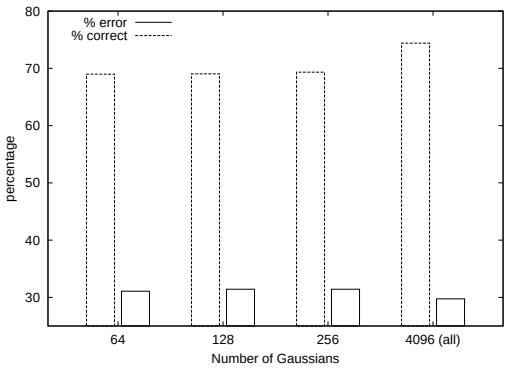

Fig. 3. Error and correct recognition rates as a function of the number of Gaussians kept in the convolutive model combination approach (tennis corpus).

Though the assumption that the features are additive does not hold for cepstral coefficients or for the modules of a filter-bank (square root compression of the energy), we tried the convolution approach on such features. Unsurprisingly, poor results were obtained with the filter-bank modules with only 3% of correctly classified multi-class segments. The results with the cepstral coefficients were similar to those of the baseline system since, the base assumption being false, the combined models do not correspond to the features and are therefore never chosen by the alignment algorithm.

The results of these experiments show that, as the concatenation approach, the convolution approach is also a good candidate for detecting multiple class segments as long as additive features are used. It also turned out that, for the present task, spectral coefficients give results comparable to those obtained with cepstral features.

3.4. Binary tests

The BIC segmenter can lead to very different quality of segmentation, depending on the value of the weight λ. For high values, it will keep only few segments, with a poor recall value. On the tennis, for $\lambda = 0.3$ we have only 134 segments left with a precision of 80% but with a recall of only 20%. We therefore prefer a smaller $\lambda = 0.08$ with a recall of 60%. The Viterbi segmentation gives similar results than the $\lambda = 0.08$ BIC segmenter. For this reason the results shown in table 1 and 2 are quite the same for the two methods.

The binary tests gave good results for the documentaries corpus, almost matching the concatenation method. Figure 4 shows the variations of the recognition and error rates as a function of the likelihhod ratio threshold. When the

threshold is low the system detects a class more easily and the error rate increases very rapidly. When the threshold is high the system converge to the baseline system performances. This is due to an implementation choice: when the score of all the classes are below the threshold, we consider that only the class with highest likelihood is present. The $\beta = -1000$ likelihood threshold yields to the best results for both Viterbi and BIC segmentation. The difference of recognized multiple segments performance is due to the oversegmentation of the Viterbi segmenter as opposed to the BIC one since the oversegmentation increases the probability of detecting a wrong class.

On the other hand, poor results were obtained on the tennis corpus with this method. On this corpus, the best threshold performance corresponds to the baseline system which means that no multiple classes has been detected. With low threshold values, the error rate rapidly increases: there is no real optimal threshold for this corpus. The mean rank for each class score gives a possible explanation to for this problem. Indeed, if we suppose that the class repartition is approximately uniform, the mean rank (using five classes) should be around 2.5. But it appears that one class has a mean rank above 4 which means that this class has a different range of value for the likelihood ratio from the others. Since a unique threshold is used, if the score of a class is not in a range comparable to the scores of the other classes, the system is not able to recognize that class anymore. The problem might be fixed by setting an independent threshold value for each class.

3.5. Comparison of the methods

Table 1 summarizes the results obtained with the various methods. The first column corresponds to the baseline Viterbi system, the second one to the concatenation up to two classes, the next one to the concatenation up to three classes, the fourth column stands for the convolution, and the last two columns to binary tests with Viterbi segmentation and BIC segmentation respectively. For each method we give the correct recognition rate, the total error rate and the multi class segment recognized.

As discussed in the two previous sections, the two model combination methods, concatenation and convolution, gave comparable results. The concatenation approach gave the best results and is much faster than the convolution approach since the combined mixture models have far less Gaussians. However, the convolution is also a promising method, the main difficulty being to have additive features that provide a good representation of the data. A linear discriminant analysis has been tested to improve the filter-bank energies features but without any better result.

The approach based on binary hypothesis testing gave mitigated results. Relatively good results for the documentary corpus were obtained, but with no progression for the multiple class recognition, compared to the concatenation method: 66% are recognized with the concatenation while 30% for the BIC+binary tests method and only 17% for Viterbi+binary tests. It turned out that the decision threshold determined on the training corpus was not optimal on the test set and that the system lacks score normalization. By decreasing the threshold, we were able to achieve 38% of correctly recognized multi-class segments but with only an increase of the total error rate by 1%.

		tennis		**documentaries**	
		prior	post	prior	post
viterbi	%c	62.4	65.2	85.5	87
	%e	37.6	34.8	39.9	40
	%m	0	0	17	17.8
bic	%c	63.22	66.2	78.8	84.7
	%e	37.2	33.8	42.9	43.9
	%m	2.8	0	30.4	38

Tab. 2. For each corpora, a priori (left column) and a posteriori (right column) results on the test corpus. %c stands for correct recognitionrate, %e for total error rate, %m for correct multiple segment detected and β for the corresponding likelihood threshold

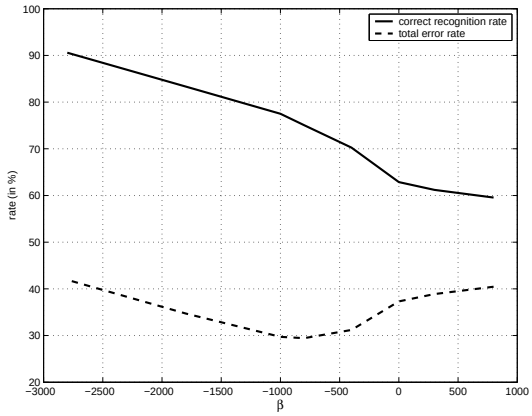

Fig. 4. error and recognition rate as a function of the threshold β for the binary tests (documentaries corpus).

4. CONCLUSION

In this paper, we have presented a system to extract audio information from video sound tracks. As the Viterbi baseline system is not able to detect simultaneous events, we proposed several approaches to solve this problem. As estimating multi-class model parameters is often intractable due to the lack of data, we investigated approaches to derive these models from the single class ones. Two methods

	viterbi		concat. (2 cl.)		concat. (3 cl.)		convol.	viterbi + bin.		bic + bin.	
	ten.	doc.	ten.	doc.	ten.	doc.	ten.	ten.	doc.	ten.	doc.
% correct	68.3	60.2	78.8	85.1	79.3	87	74.4	62.4	85.5	63.2	78.8
% error	31.7	39.8	28.4	25.5	28.8	28.3	29.7	37.6	39.9	37.2	42.9
% mcorrect	0	0	33.4	66.5	32.2	59.3	18.0	0	17	2.8	30.4

Tab. 1. Recognition and error rates for the various methods. For each method the left column gives results for the tennis corpus and the right one for the documentaries.

were tested (concatenation and convolution) and gave good results, especially for concatenation models. The convolution approach gave slightly worse results with much lower performance due to the model complexity involved. Finally, a method based on binary hypothesis tests was also studied but yield to mitigated results. However, we belive that these poor results are mainly due to the lack of score normalization. We will investigate this issue in the near futur in order to improve this method.

5. ACKNOWLEDGMENTS

The work described in this paper was carried out in the framework of the Domus Videum project which focuses on video abstracting. The project is partially funded by the French Reseau National pour la Recherche en Telecommunication (National Network for Research in Telecommunications).

6. REFERENCES

[1] *Proc. of the NIST Speaker Recognition Workshop*, 2002.

[2] Mauro Cettolo and Marcello Federico. Model selection criteria for acoustic segmentation. In *ISCA Workshop on Automatic Speech Recognition*, 2000.

[3] Scott Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Broadcast News Transcription and Understanding Workshop*, pages 127–132, 1998.

[4] Perrine Delacourt. *La segmentation et le regroupement par locuteurs pour l'indexation audio*. PhD thesis, Eurecom, Sophia-Antipolis, 2000.

[5] C. Saraceno and R. Leonardi. Audio as a support to scene change detection and characterization of video sequences. In *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, pages 2597–2600, 1997.

[6] Mouhamadou Seck. *Detecting changes and tracking sound classes for audio indexing*. PhD thesis, Université de Rennes 1, January 2001. (in French).

[7] Matthew Siegler, Uday Jain, Bhihsha Raj, and Richard Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *DARPA Speech Recognition Workshop*, 1997.

[8] T. Zhang and C.-C. Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Trans. on Speech and Audio Processing*, 9(4):441–457, May 2001.