EXPERIMENTS ON SPEECH TRACKING IN AUDIO DOCUMENTS USING GAUSSIAN MIXTURE MODELING

Mouhamadou SECK, Ivan MAGRIN-CHAGNOLLEAU and Frédéric BIMBOT

IRISA (CNRS & INRIA) - Campus universitaire de Beaulieu - 35042 Rennes cedex - FRANCE
E-mails: mseck@irisa.fr - ivan@ieee.org - bimbot@irisa.fr

ABSTRACT

This paper deals with the tracking of speech segments in audio documents. We use a cepstral-based acoustic analysis and gaussian mixture models for the representation of the training data. Three ways of scoring an audio document based on a frame-level likelihood calculation are proposed and compared. Our experiments are done on a database composed of television programs including news reports, advertisements, and documentaries. The best equal error rate obtained is approximately 12%.

1. INTRODUCTION

In this paper, we deal with the difficult problem of tracking speech segments in audio documents containing speech, music, speech + music, and noise segments. This problem has already been addressed in the literature for instance in [1, 2, 3, 4]. It is a very important problem for audio indexing, as speech and music tracking is usually one of the first steps to index an audio document.

We propose here a tracking method using a cepstral-based acoustic analysis and a gaussian mixture modeling (GMM), and we test three ways of scoring an audio document which are all based on a frame-level log-likelihood calculation. We also compare GMMs with diagonal or full covariance matrices. The goal here is to test several system configurations to find what are the sensitive tuning parameters for such a system.

Our algorithms are tested on a database composed of various television programs in French including news reports, advertisements, and documentaries.

2. METHODS

2.1. Gaussian Mixture Models (GMM)

A gaussian mixture model is represented by:

$$p(\mathbf{x}) = \sum_{i=1}^{n} \pi_i \cdot \mathcal{N}_i(\mathbf{x}; \mu_i, \Sigma_i),$$

where $\mathbf{x}$ is a feature vector, n the number of gaussian probability density functions (pdf's) in the mixture, π_i the weight associated with the pdf i, and $\mathcal{N}_i$ a gaussian pdf with mean μ_i and covariance matrix Σ_i. The weights π_i satisfy the two following properties:

$$\pi_i \in]0, 1[\text{ for } 1 \leq i \leq n \text{ and } \sum_{i=1}^{n} \pi_i = 1.$$

Given a sequence of feature vectors, a GMM is trained using the expectation-maximization (EM) algorithm [5, 6].

2.2. Tracking Algorithms

To track speech segments in an audio document, two training models are required: one model representing speech data, noted S, which corresponds in our experiments to *sponly* and *spmu* data (see section 3.1); and one model representing non-speech data, noted $\bar{S}$, which corresponds in our experiments to *muonly* data (see section 3.1).

The first phase of the tracking algorithm consists in calculating for each feature vector the likelihood obtained with model S and the likelihood obtained with model $\bar{S}$:

$$p(\mathbf{x}_t|S) = \sum_{i=1}^{n} \pi_i^{(S)} \cdot \mathcal{N}_i(\mathbf{x}_t; \mu_i^{(S)}, \Sigma_i^{(S)}),$$

and

$$p(\mathbf{x}_t|\bar{S}) = \sum_{i=1}^{n} \pi_i^{(\bar{S})} \cdot \mathcal{N}_i(\mathbf{x}_t; \mu_i^{(\bar{S})}, \Sigma_i^{(\bar{S})}).$$

We now propose three ways of scoring based on these likelihood calculations.

2.2.1. Smoothed Log-Likelihood Ratio (SLLR)

The first scoring is a smoothed log-likelihood ratio (SLLR). A log-likelihood ratio is calculated for each feature vector of the audio document:

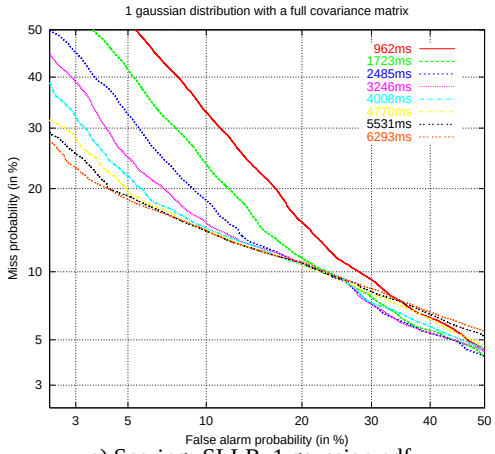
$$\mathcal{R}(\mathbf{x}_t|S; \bar{S}) = \log p(\mathbf{x}_t|S) - \log p(\mathbf{x}_t|\bar{S})$$

The log-likelihood ratio is then smoothed over a sequence of several consecutive feature vectors to attenuate its sharp variations. The smoothing is done by calculating the arithmetic mean of several consecutive vectors around the vector under consideration, after applying to this sequence a Hamming window $w = \{w_{-t_0}, ..., w_{t_0}\}$ of length $2t_0 + 1$:

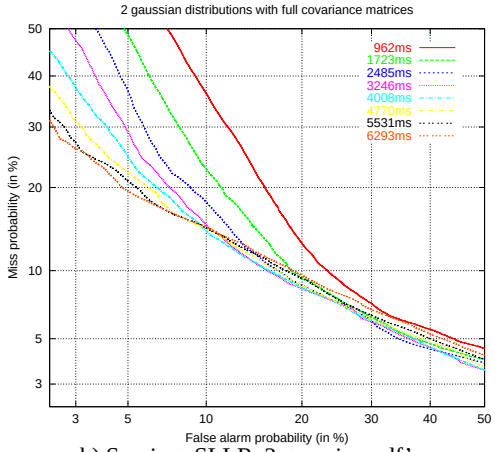
$$s_1(t) = \frac{\sum_{\tau=-t_0}^{t_0} w_\tau \cdot \mathcal{R}(\mathbf{x}_{t+\tau}|S; \bar{S})}{\sum_{\tau=-t_0}^{t_0} w_\tau}$$

Ivan Magrin-Chagnolleau is now with the computer laboratory of Avignon (LIA), 339 chemin des Mainajariès, BP 1228, 84911 Avignon cedex 9, France.

A part of this work has been done in the framework of the French national research project AGIR.



a) Scoring: SLLR. 1 gaussian pdf.



b) Scoring: SLLR. 2 gaussian pdf's.

Fig. 1. Results for gaussian pdf's with full covariance matrices and for various durations of the smoothing window (962ms, 1723ms, 2485ms, 3246ms, 4008ms, 4770ms, 5531ms, and 6293ms). Scoring: SLLR.

Finally, $s_1(t)$ is compared to a threshold θ_1 and the corresponding feature vector is classified as speech if $s_1(t)$ is higher than the threshold, as non-speech otherwise.

The smoothing is entirely defined by the number $l = 2t_0 + 1$ of vectors used for the arithmetic mean calculation. Several values of l were tested for each experiment reported.

2.2.2. Mixture Model (MM)

Let $p(\mathbf{x}_t|\alpha; S; \bar{S})$ define the mixture between the likelihood of $\mathbf{x}_t$ given the model S and the likelihood of $\mathbf{x}_t$ given the model $\bar{S}$. We have:

$$p(\mathbf{x}_t|\alpha; S; \bar{S}) = \alpha \cdot p(\mathbf{x}_t|S) + (1-\alpha) \cdot p(\mathbf{x}_t|\bar{S})$$

with

$$\alpha \in [0, 1].$$

We can then define two other ways of scoring a feature vector $\mathbf{x}_t$. The first one, that we will call MM, consists in estimating $\hat{\alpha}$ which maximizes the quantity $p(\mathbf{x}_{t-t_0}, ..., \mathbf{x}_{t+t_0}|\alpha; S; \bar{S})$ in the sense of the maximum likelihood:

$$s_2(t) = \hat{\alpha} = \arg\max_{\alpha \in [0,1]} p(\mathbf{x}_{t-t_0}, ..., \mathbf{x}_{t+t_0}|\alpha; S; \bar{S}),$$

with

$$p(\mathbf{x}_{t-t_0}, ..., \mathbf{x}_{t+t_0}|\alpha; S; \bar{S}) = \prod_{\tau=-t_0}^{t_0} p(\mathbf{x}_{t+\tau}|\alpha; S; \bar{S}).$$

$\hat{\alpha}$ can be estimated using the EM algorithm.

Finally, $s_2(t)$ is compared to a threshold θ_2 and the corresponding feature vector is classified as speech if $s_2(t)$ is higher than the threshold, as non-speech otherwise.

2.2.3. Generalized Likelihood Ratio (GLR)

We can also define the generalized likelihood ratio as:

$$s_3(t) = \frac{\sup_{\alpha \in [0,1]} p(\mathbf{x}_{t-t_0}, ..., \mathbf{x}_{t+t_0}|\alpha; S; \bar{S})}{\prod_{\tau=-t_0}^{t_0} p(\mathbf{x}_{t+\tau}|\bar{S})}$$

Finally, $s_3(t)$ is compared to a threshold θ_3 and the corresponding feature vector is classified as speech if $s_3(t)$ is higher than the threshold, as non-speech otherwise.

3. EXPERIMENTS AND RESULTS

3.1. Database

For our experiments, we used a subset of a database provided by the French national institute for audiovisual (INA) in the framework of the European research project DIVAN. This subset was composed of 6 CD-ROMS, noted CD1 to CD6, containing various extracts of television programs including news reports, advertisements, and documentaries. We labeled the whole subset in function of four sound categories: *sponly* corresponding to speech with no background, *spmu* corresponding to speech with a musical background, *muonly* corresponding to music only, and *other* corresponding to various kind of noises. Table 1 gives the total durations of each category for each CD-ROM.

CD	*sponly*	*spmu*	*muonly*	*other*	Total
CD1	19'25"	7'42"	6'10"	2'15"	35'34"
CD2	0'48"	7'17"	5'38"	0'58"	14'43"
CD3	9'09"	0'52"	1'20"	0'59"	12'23"
CD4	8'32"	1'40"	1'44"	1'20"	13'17"
CD5	12'19"	1'34"	1'03"	0'28"	15'25"
CD6	10'43"	2'25"	1'10"	0'30"	14'48"

Table 1. Total durations of each category for each CD-ROM of the database.

The CD-ROMS CD2 to CD6 were used for the training phase. We trained three models for the three categories *sponly*, *spmu*, and *muonly*. The model for speech data was obtained by combining the model for *sponly* and the model for *spmu* with equal weights. The model *muonly* was used as the model for non-speech data. Finally, the tracking was done on CD1 without the segments of the category *other*.

3.2. Acoustic Analysis

Each audio segment, sampled at 22.05 kHz, was decomposed in frames of 10 ms extracted every 10 ms. A Hamming window was applied to each frame. The signal was pre-emphasized with a coefficient 0.95. For each frame, a fast Fourier transform was computed and provided 256 square modulus values representing the short term power spectrum in the 0-11025 Hz band. This Fourier power spectrum was then used to compute 24 filterbank coefficients, using triangular filters placed on a linear frequency scale. We took the base 10 logarithm of each filter output and multiplied the result by 10, to form a 24-dimensional vector of filterbank coefficients in dB. Then, we computed cepstral coefficients c_1 to c_{16} [7], augmented by their Δ coefficients (calculated over 5 vectors) [8], and by the Δ energy (calculated also over 5 vectors). We finally obtained 33-dimensional feature vectors.

3.3. Evaluation

Results of the various systems were measured by a DET curve [9]. This representation is a way of showing all the possible operating points of a system (false alarm rate vs. miss rate) corresponding to various thresholds in a scale which makes the result curves rather linear, when the distributions of scores for speech and non-speech segments are both gaussian. For this task, the false alarm rate and the miss rate were defined as follows:

$$\mathcal{R}_{FA} = \frac{\text{Number of non-speech vectors labeled as speech}}{\text{Number of non-speech vectors}}$$

$$\mathcal{R}_{MI} = \frac{\text{Number of speech vectors labeled as non-speech}}{\text{Number of speech vectors}}$$

3.4. Type of Covariance Matrices

In this section, we report results obtained with the smoothed log-likelihood ratio (SLLR) score for various number of gaussian pdf's with full covariance matrices (Fig. 1) or diagonal covariance matrices (Fig. 2).

The best results obtained with diagonal covariance matrices correspond to 64 gaussian pdf's. We tested 128 gaussian pdf's but the results degrade, which is probably due to the lack of data to estimate the parameters of the GMMs.

The best results obtained with full covariance matrices correspond to 1 gaussian pdf. We tested 4 gaussian pdf's, but the results degrade, probably also because of the lack of data to estimate the parameters of the GMMs.

GMMs with full covariance matrices perform slightly better than GMMs with diagonal covariance matrices (12 % of equal error rate vs. 14 %). This may be due to the fact that the correlations between various cepstral coefficients characterize very well one of the two classes of sounds that we want to discriminate (probably the music class).

A long smoothing window (between 4 and 6 seconds) performs better than a shorter one (around 1 or 2 seconds), but the right value to use depends on the configuration of the system. We tried longer windows but the results degrade again. It is also very interesting to notice that, in the full case, most of the DET curves converge to the same values when the false alarm rate increases.

3.5. Type of Scoring

Fig. 3 gives the results for the three types of scoring using 1 gaussian pdf with a full covariance matrix. When the duration of the smoothing window is small, the SLLR score does not give as good performance as the two other scores. But when the duration of the smoothing window is between 1 and 5 seconds, the three scores perform very similarly, and give an equal error rate of approximately 12 %. If the duration of the smoothing window is increased, the performances degrade again.

4. CONCLUSIONS AND PERSPECTIVES

In this paper, we presented experiments on speech tracking in audio documents using gaussian mixture models. Three scoring methods based on a frame-level likelihood calculation were proposed and compared. And various configurations of the GMMs were tested. The best result was obtained with a GMM composed of one gaussian pdf with a full covariance matrix. In that configuration, the optimal duration for the smoothing window was between 1 and 5 seconds, and the three scores performed similarly, giving an equal error rate of approximately 12 %.

This work can be extended in several directions. First, we intend to build a new model for non-speech data including noise data in order to be able to apply speech tracking on the complete data (including noise segments). We also want to apply our tracking algorithm for music tracking complementary to speech tracking. Finally, it would be interesting to study the distribution of the segment durations to correlate this distribution with the optimal duration for the smoothing window.

5. REFERENCES

[1] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature music/speech discriminator," in *Proceedings of ICASSP 97*, Apr. 1997, vol. II, pp. 1331–1334.

[2] Mouhamadou Seck, Frédéric Bimbot, Didier Zugaj, and Bernard Delyon, "Two-class signal segmentation for speech / music detection in audio tracks," in *Proceedings of EUROSPEECH 99*, Sept. 1999, pp. 2801–2804.

[3] Jonathan Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proceedings of ICME 2000*, Aug. 2000.

[4] Mouhamadou Seck, *Détection de ruptures et suivi de classes de son pour l'indexation sonore*, Ph.D. thesis, IRISA, Rennes, France, Dec. 2000.

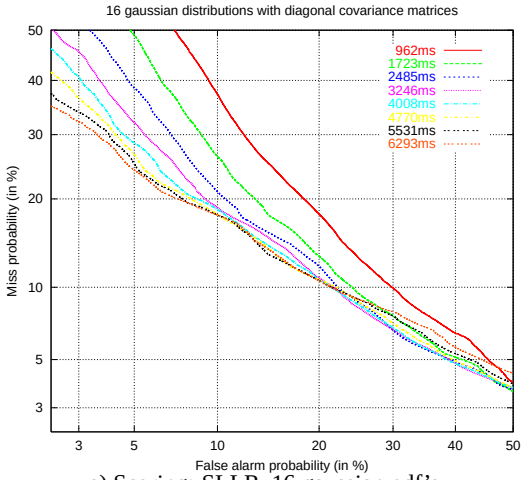
[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 6, no. 39, pp. 1–38, 1977.

[6] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, John Wiley and sons, 1985.

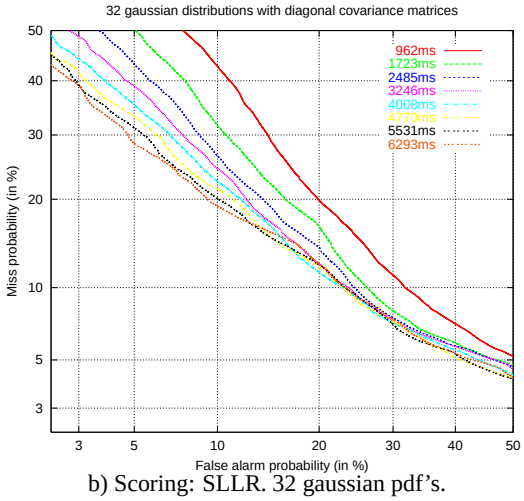
[7] Alan V. Oppenheim and Ronald W. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, June 1968.

[8] Sadaoki Furui, "Comparison of speaker recognition methods using static features and dynamic features," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 342–350, June 1981.

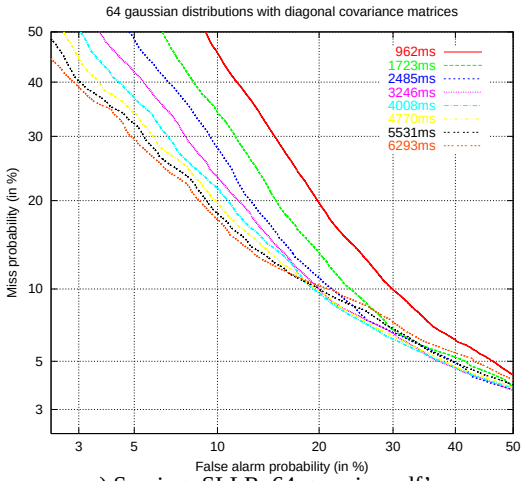
[9] A. Martin et al., "The DET curve in assessment of detection task performance," in *Proceedings of EUROSPEECH 97*, Sept. 1997, vol. 4, pp. 1895–1898, Rhodes, Greece.



a) Scoring: SLLR. 16 gaussian pdf's.

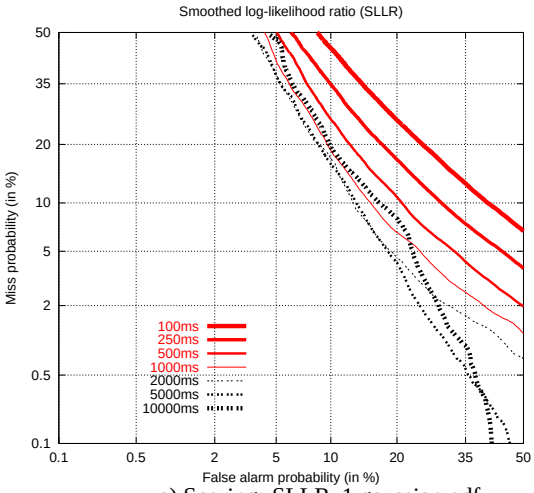


b) Scoring: SLLR. 32 gaussian pdf's.

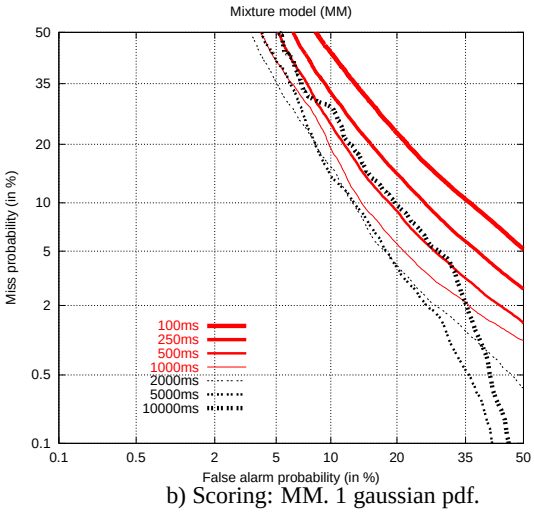


c) Scoring: SLLR. 64 gaussian pdf's.

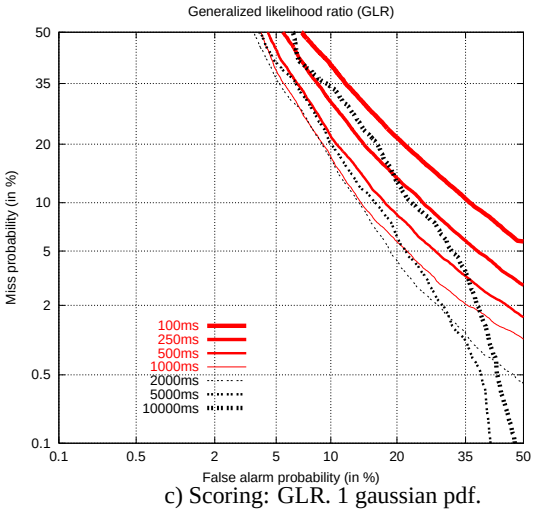
Fig. 2. Results for gaussian pdf's with diagonal covariance matrices and for various durations of the smoothing window (962ms, 1723ms, 2485ms, 3246ms, 4008ms, 4770ms, 5531ms, and 6293ms). Scoring: SLLR.



a) Scoring: SLLR. 1 gaussian pdf.



b) Scoring: MM. 1 gaussian pdf.



c) Scoring: GLR. 1 gaussian pdf.

Fig. 3. Results for 1 gaussian pdf with a full covariance matrix for the various scores and for various durations of the smoothing window (100ms, 250ms, 500ms, 1000ms, 2000ms, 5000ms, and 10000ms).