TWO-CLASS SIGNAL SEGMENTATION FOR SPEECH / MUSIC DETECTION IN AUDIO TRACKS

Mouhamadou SECK, Frédéric BIMBOT, Didier ZUGAJ and Bernard DELYON

IRISA - SIGMA2 / INRIA & C.N.R.S.
Campus Universitaire de Beaulieu
35042 RENNES cedex, FRANCE

e-mail : {mseck, bimbot, dzugaj, delyon}@irisa.fr

Key-words : speech/audio signal processing, indexing, Gaussian Mixture Models, model order, likelihood ratio

Abstract

We present a technique for the segmentation of a sound track into two classes of segments. Each frame of signal is preprocessed by extracting cepstral coefficients and their first order derivatives. For each class, the distribution of the frame parameter vectors is modeled by a Gaussian Mixture Model (GMM). GMM order is selected using two criteria : the Minimum Description Length (MDL) criterion and the Akaike Information Criterion (AIC). Frame score is based on a weighted log-likelihood ratio in a window around the frame. Decision for each frame is taken by comparing its score to a threshold. Experiments are presented on speech / music segmentation in audio tracks. In these experiments, the MDL criterion leads to a reasonable GMM order. Using the MDL criterion for GMM order selection, frame classification error rate is around 20%. However, using GMMs with much lower orders, only decreases marginally performances.

1. Introduction

With the fast growth of the amount of audio visual data that have to be handled, the need for automatic indexing functionalities is increasing. One of the basic tasks in audio segmentation consists in organising audio tracks into homogeneous time intervals (or segments). For instance, segments may correspond to particular categories of sounds, such as speech, music, noise... In this work, we consider the case when segments belong to a set of two classes : target and non-target.

Section 2 presents an overview of the segmentation system and section 3 gives a detailed description. An experiment on speech / non-speech segmentation is presented in section 4 and results are discussed in section 5. The last section is dedicated to conclusions and perspectives.

2. System overview

The signal is first cut regularly into frames. Each frame is represented by a vector of acoustic parameters. For each class (target and non-target), the distribution of frame parameter vectors is modeled using training data. The models are then used for a frame-by-frame scoring and labeling (target / non-target). A segment is then defined as a sequence of frames which have the same label .

For parametrization, we use a vector of cepstral coefficients and their delta. The distribution of Target (resp. non-target) parameter vectors is modeled by a Gaussian Mixture Model (GMM) with diagonal covariance matrices. Frame score is based on a weighted log-likelihood ratio, in a window around the frame. Decision is taken frame-by-frame, by comparing frame score to a threshold.

3. System description

3.1. Problem

Let \( \mathcal{Y} = \{Y_1, \cdots, Y_n\} \) be a sequence of independent observations of a discrete time signal, and \( M_1^* \) and \( M_2^* \) be two distributions. We assume that there exist change times \( r_1 < r_2 < \cdots < r_k \), such that in each segment \( \{r_i + 1, r_{i+1}\} \) \( (r_0 = 0, r_{k+1} = n) \), the distribution of the signal is either \( M_1^* \) or \( M_2^* \).

Observations \( \mathcal{Y} \) correspond to a sequence of frame parameter vectors, extracted from a signal. \( M_1^* \) and \( M_2^* \) correspond respectively to target and non-target distributions.

3.2. Modeling

Each distribution \( M_i^* \) is modeled by a GMM \( \tilde{M}_i \) with diagonal covariance matrices. The probability density function (p.d.f.) of GMM \( M \) with order \( K \), for a vector \( y \), is :

\[
p(y|M) = \sum_{j=1}^{K} c_j \cdot g(y|\mu_j, \Sigma_j)
\]

where \( c_j \) is the weight of the \( j^{th} \) mixture component, and \( g(y|\mu_j, \Sigma_j) \) denotes the p.d.f. of a Gaussian distribution with mean \( \mu_j \) and covariance matrix \( \Sigma_j \).

When the GMM order \( K \) is fixed, one can estimate a GMM by Maximum Likelihood Estimator (MLE) [Sun73], using the standard Expectation Maximisation (EM) algorithm [DLR77]. The GMM order can be chosen using some criteria for selecting the number of pa-

1 Two contiguous segments have therefore different labels.
3.3. Gaussian Mixture Model order selection

In this section, \( M^* \) denotes either the target model \( M^*_1 \) or the non-target model \( M^*_2 \). When the mixture order is fixed, the MLE can be used to approximate the model \( M^* \). When no prior information is available about the GMM order, one has to choose one model among competing GMMs of different orders. This is a model order selection problem. To tackle this problem, conventional methods consist of introducing in the maximum likelihood criterion, a term which penalizes higher order models. In this work, we use two criteria the Akaike Information Criterion (AIC) [Aka74] and the Minimum Description Length (MDL) [Ris78] :

\[
\text{AIC}(j) = -2 \log[p(Y_n | \hat{M}(j))] + 2m(j)
\]

\[
\text{MDL}(j) = -2 \log[p(Y_n | \hat{M}(j))] + m(j) \log n
\]

where \( \hat{M}(j) \) is the MLE of GMM with order \( j \), and \( m(j) \) is the number of parameters in the model \( \hat{M}(j) \). In practice, these criteria are evaluated on the training data.

3.4. Score and Decision

The technique we propose for localising segments in the sequence of observations, is based on a weighted log-likelihood ratio, we define for a sequence of \( L = (2H + 1) \) observations \( \{Y_{i-H}, \cdots, Y_{i+H}\} \) around time \( t \) by :

\[
S(t) = \sum_{i=t-H}^{t+H} w(i-t+H+1) \log \frac{p(Y_i | \hat{M}_1)}{p(Y_i | \hat{M}_2)}
\]

where \( \{w(i); i = 1, \cdots, L\} \) are weights of a Hamming window with length \( L \). The classification decision for the observation at instant \( t \) is taken by comparison with a threshold \( \Theta \) :

\[
S(t) \begin{cases} \hat{M}_1^* & \text{if } \Theta \\ \hat{M}_2^* & \text{if } \Theta \end{cases}
\]

4. Experiments

In our experiments, we consider the problem of segmentation of a sound track into the following classes of segments :

- speech (target) class : sounds containing speech only
- non-speech (non-target) class : sounds including a non-speech component (possibly some speech).

4.1. Data base and experimental conditions

Our data base consists of 16 kHz sampled sound tracks of TV programs, collected by INA (Institut National de l’Audiovisuel) :

- training data : advertisements (14 min.) and news (13 min.);
- testing data : a documentary (35 min.).

The acoustic features extracted from the signal are 16 LPC based cepstral coefficients and their deltas, computed from frames of 26 ms duration.

4.2. Results of GMM order selection

Speech and non-speech models are estimated on training data, for different values of GMM orders : from 1 to 125. GMM estimation is done using the standard EM algorithm with 15 iterations. The system described in section 3 is applied to test data.

In figure 1, AIC and MDL criteria (3.3) are represented for speech models. Minimum value for the MDL criterion is obtained with a GMM order equal to 78. AIC do not show reasonable minimum. The same observation holds, for non-speech models; see figure 2. In this case, the minimum for the MDL criterion is the GMM with order 69.
Figure 3: Results of speech (darker) / non-speech (lighter) detection: reference segmentation and results obtained with various values of the analysis window length ($L$). Speech model is a $78 \text{ GMM}$ and non-speech model is a $69 \text{ GMM}$. The sound track is of 13 minutes duration.

Figure 4: Speech Detection Error Trade-off (DET) curves (miss versus false detection), with various values of the analysis window length ($L$). Speech model is a $78 \text{ GMM}$ and non-speech model is a $69 \text{ GMM}$. The sound track is of 13 minutes duration. For $L = 1001, 1201, 1401$, the EER is around $20\%$. 
4.3. Performance results

We present in this section, results obtained with GMMs selected by the MDL criterion, mentioned above, in section 4.2. In figure 3, the first band represents the correct segmentation, where black intervals correspond to speech segments. The other bands represent weighted log-likelihood ratio scores, for different values of the analysis window length ($L$). Darker intervals correspond to higher scores.

Performances are evaluated by the percentage of frames which are correctly recognized. Figure 4 shows speech Detection Error Trade-off (DET) curves, for different values of the analysis window length. A DET curve represents false detection versus false alarm, when the decision threshold $\theta$ varies, using a particular scale [MDOP97]. An Equal Error Rate (EER) around 20% is observed in this case.

5. Interpretation of results

We discuss here, influence on system performances, of two quantities : the analysis window length ($L$) and the GMM order for target and non target models.

We observe in figure 4 that performances are sensitive to the analysis window length. EER varies from 30% for $L = 101$ to 20% for $L = 1201$. Performances do not vary significantly when $L \geq 1201 \approx 31$ seconds.

![Figure 5: Speech / non-speech Equal Error Rate (EER), for different GMM orders for both speech and non-speech. Lower EERs correspond to darker boxes. EER is around 21% when GMM order is equal to 2 for speech model and 1 for non-speech model](image)

Figure 5 represents EER obtained with different values of GMM order (1 to 7) for both speech and non-speech models. Lower EERs correspond to darker boxes. Generally, best results are given by higher order GMMs for speech model and lower order GMMs for non-speech model. When using GMM with order 2 for speech and GMM with order 1 (a Gaussian model) for non-speech, an EER around 21% is obtained. Figure 6 shows DET curves obtained with these GMM models, for different values of the analysis window length. Remind that GMMs selected by the MDL criterion (orders : 6 and 9) give an EER around 20%. This means that GMM order selection methods lead here to performances that can be nearly obtained with much lower order GMMs.

![Figure 6: Speech Detection Error Trade-off (DET) curves (miss versus false detection), with various values of the analysis window length (L). GMM order is 2 for speech model and 1 for non-speech model. The sound track is of 13 minutes duration. For $L = 1001, 1201$ and 1401, the EER is around 21%](image)

6. Conclusion

Our experiments show that a reasonable level of performance can be obtained using GMM models for speech / music detection in audio tracks. However, 20% frames remain incorrectly classified and the model complexity seems to have little impact on the classification accuracy. Other directions for improving the performance should be explored, in particular the choice of more task-specific acoustic features (in particular time-domain features).

References