Robust Unsupervised Speaker Segmentation for Audio Diarization

Kadri Hachem, Manuel Davy, Noureddine Ellouze

Robust Unsupervised Speaker Segmentation for Audio Diarization

Hachem Kadri[1], Manuel Davy[1] and Noureddine Ellouze[2]
[1] *LAGIS, UMR CNRS 8164 and INRIA SequeL Team*
France
[2] *Unité de Recherche Signal, Image et Reconnaissance de Formes*
Tunisia

1. Introduction

Audio diarization (Reynolds and Carrasquillo, 2005) is the process of partitioning an input audio stream into homogeneous regions according to their specific audio sources. These sources can include audio type (speech, music, background noise, ect.), speaker identity and channel characteristics. With the continually increasing number of larges volumes of spoken documents including broadcasts, voice mails, meetings and telephone conversations, diarization has received a great deal of interest in recent years which significantly impacts performances of automatic speech recognition and audio indexing systems. A subtype of audio diarization, where the speech segments of the signal are broken into different speakers, is speaker diarization (Anguera et al., 2006, 2007; Tranter and Reynolds, 2006). It generally answers to the question "Who spoke when?" and it is divided in two modules: speaker segmentation and speaker clustering. The goal of speaker segmentation is finding the times when there is a change of speaker in the audio stream. Speaker clustering consists in merging speech segments, detected by the speaker segmentation step, related to a same speaker.

Recently, three main domains of application for speaker segmentation have received special attention (Reynolds and Carrasquillo, 2004):

- Broadcast news : Radio and TV programs with various kinds of programming, usually containing commercial breaks and music, over a single channel.
- Recorded meetings: meetings or lectures where multiple people interact in the same room or over the phone. Normally recordings are made with several microphones.
- Phone conversations: single channel recordings of phone conversations between two or more people.

Segmenting this types of audio stream in terms of speakers is useful in many application. In Automatic Speech Recognition (ASR) (Moraru et al., 2003), for example, an initial segmentation is required in terms of homogeneous speech and non-speech regions. Having segmented speech regions, it is also often necessary to segment these further in terms of homogeneous speaker turns. In addition to improving ASR systems, speaker turn information can be helpful for speaker adaptation in rich transcription of videos and meetings (Bonastre et al., 2000) and for content based audio classification and retrieval (Hansen et al., 2005) which have a wide range of applications in the entertainment industry, audio archive

management, surveillance, etc. Audio segmentation would also be an important tool in summarizing meetings, which has recently gained a lot of interest in the research community. For example, segmentation of the speech data in terms of speakers could help in efficient navigation through audio documents like meeting recordings (Dielmann and Renals, 2007; Jin and Schultz, 2004). Using these segmentation queues, an interested user can directly access a particular segment of the speech made by a particular speaker.

1.1 Previous works

Recent research on audio segmentation mostly focused on four categories: energy based, model-based (Kemp et al., 2000), metric-based (Delacourt and Wellekens, 2000), and information criterion-based approaches (Cettolo and Federico, 2000; Cettolo and Vescovi, 2003; Chen and Gopalakrishnan, 1998). Energy audio segmentation only detects change-points at silence segments, which generally are not directly connected with the acoustic changes of the audio signals. Model-based segmentation approach requires predefined audio classes and complete training data. The metric-based approach are not stable and need thresholds generally selected from experiments results. The information criterion-based scheme are proposed for evaluating models constructed by various estimation procedures when the specified family of probability distributions does not contain the distribution generating the data. The so-called Delta Bayesian information criterion (BIC) segmentation algorithm is widely employed in many studies (Chen and Gopalakrishnan, 1998). The BIC is intended to provide a measure of the weight of evidence favoring one model over another. According to previous research, the Delta-BIC is threshold-free and suitable for unknown acoustic conditions. However, this method, extremely computationally expensive, can introduce an estimation error due to insufficient data when the speaker turns are close to each other (Chen and Gopalakrishnan, 1998; Huang and Hansen, 2004). In order to minimize these effects, Delacourt and Wellekens (2000) tested different metric criteria to associate them to the BIC criterion such as the Kullbach-Leibler distance, the similarity measure and the Generalized Likelihood Ratio measure (GLR). Still, this method encountered problems in case of short segments and requires also a high computation cost. On another issue, Zhou and Hansen (2000) recommend the use of the T^2-Statistic for metric-based segmentation in the aim to reduce this computation cost. However its technique, T^2-BIC, depends on many empiric parameters which affect the quality of the detection of speaker turns. In our previous work (Kadri et al., 2006), we developed a hybrid segmentation algorithm called $DIS\_T^2\_BIC$ to improve the detection of speaker turns close to each others using a fixed threshold independent of the type of the audio stream with a low computation cost. Nevertheless all of these techniques suppose that the audio signal don't contains different acoustic changes and simultaneous speeches of two or more speakers and then find difficulties in segmenting streams containing background noise and overlapped speeches.

1.2 Contributions and Chapter organization

The main focus of this chapter is to introduce a new unsupervised speaker segmentation technique robust to different acoustic conditions. In most commonly used model selection segmentation techniques like BIC segmentation, the basic problem may be viewed as a two-class classification where the object is to determine whether N consecutive audio frames constitute a single homogeneous of frames W or two such windows: W_1 and W_2 with the boundary frame or change occurring at the i^{th} frame. In order to detect if a speaker change occurred within a window of N frames, two models are built. One which represents the entire window by a Gaussian characterized by μ (mean) , Σ (variance) ; a second which

represents the window up to the i^{th} frame, W_1 with μ_1,Σ_1 and the remaining part, W_2, with a second Gaussian μ_2,Σ_2. This representation using a Gaussian process is not totally exact when the audio stream contains overlapped speeches and very short segments. To solve this problem, our proposed segmentation technique use the one class SVM and exponential family model to maximize the generalized likelihood ratio with any probability distribution of windows (Kadri et al., 2008). Moreover, we use the discrete wavelet coefficients (DWC) to improve the detection of speaker changes in the presence of background noise. The use of this coefficient is suitable since our technique is insensitive to the dimension of acoustic features.

The remainder of this chapter is organized as follows. Section 2 details previous audio segmentation techniques based on BIC. Section 3 reviews the support vector machines approach and the exponential family model. The proposed speaker change detection method is illustrated in section 4. Experimental results are provided in Section 5. Section 6 concludes the paper with a summary and discussion.

2. Previous techniques: BIC based segmentation techniques

Model selection based speaker segmentation is proposed by Chen and Gopalakrishnan (Chen and Gopalakrishnan, 1998). Their method employs the Bayesian information criterion as model selection criterion, illustrating several desirable properties such as robustness, threshold independence and optimality.

2.1 BIC Segmentation

BIC (Chen and Gopalakrishnan, 1998) is a model selection criterion penalized by the model complexity (amount of free parameters in the model). For a given acoustic segment X_i, the BIC value of a model M_i indicates how well the model fits the data, and is determined by:

$$BIC(X, M) = \log L(X_i, M_i) - \frac{\lambda}{2}\#(M_i) \cdot \log(N_i) \tag{1}$$

$\log L(X_i, M_i)$ is the log-likelihood of the data given the considered model, N_i is the number of frames in the considered segment, $\#(M_i)$ the number of free parameters to estimate in model M_i and λ is a free design parameter dependent on the data being modeled. λ determines the 'weight' applied to model parameters, theoretically 1, but tunable in practice. Given several different candidate models to explain a single dataset, the model with the largest BIC gives the best fit according to this criterion.

The BIC-based segmentation procedure is as follows: A sequence of d-dimensional audio feature vectors $X = x_i \in \mathbb{R}^d : i = 1, \ldots, N$ are modeled as independent draws from either one or two multivariate Gaussian distributions. The null hypothesis is that the entire sequence is drawn from a single distribution:

$$H_0 = \{x_1, \ldots, x_N\} \sim \mathcal{N}(\mu_0, \Sigma_0)$$

where $N(\mu, \Sigma)$ denotes a multivariate Gaussian distribution with mean vector μ and full covariance matrix Σ. The null hypothesis is compared to the hypothesis of having a segment boundary after sample t i.e. that the first t points are drawn from one distribution and that

the remaining points come from a different distribution:

$$\begin{aligned} H_1 : \{x_1, \ldots, x_t\} &\sim \mathcal{N}(\mu_1, \Sigma_1) \\ \{x_{t+1}, \ldots, x_N\} &\sim \mathcal{N}(\mu_2, \Sigma_2) \end{aligned}$$

The difference in BIC scores between these two models is a function of the candidate boundary position t:

$$\Delta BIC(t) = \log(\frac{\mathcal{L}(X \backslash H_0)}{\mathcal{L}(X \backslash H_1)}) - \frac{\lambda}{2} \frac{d^2 + 3d}{2} \log(N) \tag{2}$$

where $\mathcal{L}(X \backslash H_0)$ is the likelihood of X under hypothesis H_0 etc., and $(d^2 + 3d)/2$ is the number of extra parameters in the two-model hypothesis H_1. When $\Delta BIC(t) > 0$, we place a segment boundary at time t, and then begin searching again to the right of this boundary and the search window size N is reset. If no candidate boundary t meets this criteria, the search window size is increased, and the search across all possible boundaries t is repeated. This continues until the end of the signal is reached.

2.2 T^2-BIC

T^2-BIC (Zhou and Hansen, 2000) is a variant of BIC segmentation technique which validates each speaker change point detected by Hotelling's T^2-statistic using the BIC criterion. Hotelling's T^2-statistic is a multivariate analogue of the square of the t-distribution (Anderson, 1985). The T^2-statistic is used when we wish to test if the mean of one normal population is equal to the mean of the other where the covariance matrices are assumed equal but unknown. In terms of segmentation (Wegmann et al., 1999), the problem can be viewed as testing the hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_0 : \mu_1 \neq \mu_2$ where μ_1, μ_2 are, respectively, the means of two samples of the audio stream, one containing the frame $[1, b]$ and the second contains $[b, N]$. The likelihood ratio test is given by the following T^2-statistic:

$$T^2 = \frac{b(N-b)}{N} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \tag{3}$$

where Σ represent the common covariance matrix. The T^2 value defined in (3) can be considered as a distance measure of two samples. Obviously, the smaller the value of T^2, the more similar the two samples distributions. The T^2-BIC algorithm operates by fixing an analysis frame with L second length from the beginning of the parameterized audio stream and calculating the T^2 value in different points situated on this frame; the point that represents the highest value of T^2 is more probable to be a real speaker turns; then it can be validated by the BIC criterion. The T^2-BIC segmentation presents certainly some advantages. The selection, from the statistical criteria T^2, of a candidate speaker change permits to reduce computational costs. Thus, T^2-BIC offers a reduced calculation time compared to the BIC segmentation. Besides, this technique works with an automatic threshold and presents a low false alarm. However, T^2-BIC is not reliable for the segmentation of audio documents that contain speaker changes close to each other. In fact, it requires the use of a time delay τ (Zhou and Hansen, 2000) between two consecutive speaker turns which can lead missing some break points.

2.3 DIS_T^2_BIC

Like T^2-BIC, $DIS\_T^2\_BIC$ (Kadri et al., 2006) is a speaker segmentation algorithm which process with a fixed threshold and low computation cost. It is proposed to improve speaker

turns detection even they are close to each other. $DIS\_T^2\_BIC$ is based in a hybrid concept which is organized in two steps: the detection of most probable speaker turns and the validation of changes already detected. Speaker turns are detected by computing the value of T^2 between a pair of adjacent windows of the same size shifted by a fixed step along the whole parameterized speech signal. In the end of this procedure we obtain the curve of the variation of T^2 in time. A speaker change point is characterized by the presence of a high value peak. To differentiate high peaks from low peaks, a fixed threshold is defined as below:

$$T^2 > \frac{(N-2)p}{N-p-1} F_{p,N-p-1}(\alpha) = T_0^2 \tag{4}$$

where $F_{p,N-p-1}$ is the F-point for p and $N-p-1$ degrees of freedom with significance level α. A T^2 value lower than T_0^2 shows that the two samples are homogeneous and consequently don't present a speaker change. So, break points can be detected by searching the local maxima of the T^2 curve that verify the criterion (4). The validation of already detected break points is made using the BIC criterion. Denote $\{T_1, ..., T_N\}$ as the set of speaker turns found in the first step, a ΔBIC value is computed for each pair of windows $[T_{i-1}, T_i]$ $[T_i, T_{i+1}]$. When this value is positive, a speaker turn is identified at time i. Otherwise, the point i is discarded from the candidate set, then the ΔBIC value is applied again for a larger pair of windows $[T_{i-1}, T_{i+1}][T_{i+1}, T_{i+2}]$. At this stage, when segments are large enough, BIC criterion gives better validation results since model estimation becomes more accurate. Detecting speaker changes from the curve of T^2 gives to $DIST\_T^2\_BIC$ the advantage to detect speaker turns close to each others and the use of the T^2-statistic criteria permits to reduce the computation cost and to have an automatic threshold decision independent of the type of the audio stream. However, like others BIC based segmentation technique, suppose that the audio signal don't contains different acoustic changes and simultaneous speeches of two or more speakers and then find difficulties to segment audio streams containing background noise and overlapped speeches.

3. Background information

This section provides a brief review of reproducing kernel Hilbert spaces (Aronszajn, 1950; Schölkopf and Smola, 2002), One-class Support Vector Machines (Desobry et al., 2005; Schölkopf et al., 2000) and exponential families (Canu and Smola, 2005).

3.1 Reproducing kernel Hilbert spaces (Aronszajn, 1950)

Let $\mathcal{X}$ be a set, and $\mathcal{H}$ be a Hilbert space included in the set of all functions on $\mathcal{X}$. The Hilbert space $\mathcal{H}$ is called reproducing kernel Hilbert space (RKHS) if the evaluation functional $e_x : \mathcal{H} \ni f \longmapsto f(x) \in \mathbb{R}$ is continuous on $\mathcal{H}$ for any $x \in \mathcal{X}$.

A function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a *positive kernel* if it is symmetric and for any points $x_1, ...x_n$ in $\mathcal{X}$ the matrix $(k(x_i, x_j))_{i,j}$ is positive semidefinite, i.e., for any sequence of scalar $\alpha_1, ...\alpha_n$ the inequality $\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$ is verified.

Using Riesz's theorem, If $\mathcal{H}$ is a RKHS on $\mathcal{X}$ then there exists a function $k(., x) \in \mathcal{H}$, called *reproducing kernel*, such that $e_x(f) = f(x) = \langle f(.), k(., x)\rangle_{\mathcal{H}}$, where $\langle\ ,\ \rangle_{\mathcal{H}}$ is the inner product of $\mathcal{H}$. The function $k(x, y)$ is a positive definite kernel, because it is symmetric from $k(y, x) = \langle k(., x), k(., y)\rangle_{\mathcal{H}} = \langle k(., y), k(., x)\rangle_{\mathcal{H}} = k(x, y)$, and positive definite from $\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \| \sum_i \alpha_i k(., x_i)\|^2_{\mathcal{H}} \geq 0$.

In the other hand, it is known that for a positive definite kernel k on $\mathcal{X}$ there uniquely exists a Hilbert space $\mathcal{H}_k$ such that $\langle f(.), k(., x)\rangle_{\mathcal{H}_k} = f(x)$ holds for any $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$.

This propriety means that $\mathcal{H}_k$ is a RKHS with a reproduicing kernel k. given a RKHS $\mathcal{H}$ and its reproducing kernel $k(.,x)$, because of the uniqueness of the reproducing kernel, we can conclude that the Hilbert space $\mathcal{H}_k$ constructed by k is identical to $\mathcal{H}$. So there is a bijection between the set of all possible RKHS and the set of all positive kernels.

3.2 One-Class SVM

The One-class approach was proposed by Schölkopf et al. (2000) and has been successfully used for abrupt change detection (Davy and Godsill, 2002; Davy et al., 2006; Desobry et al., 2005). 1-SVM distinguishes one class of data from the rest of the feature space given only a positive data set. Based on a strong mathematical foundation, 1-SVM draws a nonlinear boundary of the positive data set in the feature space using a parameter to control the noise in the training data and another one to control the smoothness of the boundary.

The 1-class SVM is a method that aims at learning a single class, by determining its contours. To explain 1-class SVM, we can begin by giving a kernel. A kernel $k(x,y)$ is a positive and symmetric function of two variables lying in a Reproducing Kernel Hilbert Space with the scalar product:

$$\langle f,g\rangle_{\mathcal{H}} = \sum_{i=1}^{k}\sum_{j=1}^{l} f_i g_i k(x_i,y_j)$$

In this framework, the 1-class SVM problem with the sample $(x_i)_{i=1,\dots,m}$ is the solution of the following optimization problem under constraints for $f \in H$:

$$\begin{cases} \min_{f,\rho,\xi} \frac{1}{2}\|f\|^2_{\mathcal{H}} + C\sum_{i=1}^{m}\xi_i - \rho \\ \text{s.t.} \quad f(x_i) > \rho - \xi_i \quad i = 1,\dots,m \\ \text{and} \quad \xi_i \geq 0, \quad i = 1,\dots,m \end{cases}$$

where C is a scalar that adjusts the smoothness of the decision function, ρ is a scalar called bias and ξ are slack variables. The dual formulation is:

$$\begin{cases} \max_{\alpha\in\mathbb{R}^m} \frac{-1}{2}\ \alpha^T K\alpha \\ \text{s.t.} \quad \alpha^T e = 1 \\ \text{and} \quad 0 < \alpha_i < C, \quad i = 1,\dots,m \end{cases}$$

where K is the kernel matrix $K_{ij} = k(x_i,x_j)$ and $e = [1,\dots,1]^T$. The 1-class SVM solution is then given by solving a quadratic optimization problem of dimension m under box constraints. The decision function is $D(x) = sign(f(x) - \rho)$. The input points are considered as part of the current class as long as the decision function is positive.

3.3 Exponential family

The exponential family covers a large number (and well-known classes) of distributions such as Gaussian, Multinomial and Poisson. A general representation of a exponential family is given by the following probability density function:

$$p(x|\eta) = h(x)\ exp\{\eta^T T(x) - A(\eta)\} \tag{5}$$

where $h(x)$ is called the base density which is always ≥ 0,
η is the natural parameter,
$T(x)$ is the sufficient statistic vector

$A(\eta)$ is the cumulant generating function or the log normalizer.

The choice of $T(x)$ and $h(x)$ determines the member of the exponential family. Also we know that since this is a density function,

$$\int h(x)\ exp\{\eta^T T(x) - A(\eta)\}dx = 1$$

then,

$$A(\eta) = \log \int \exp[\eta^T T(x)]h(x)dx$$

For a Gaussian distribution, $p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}}\exp(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - log\sigma)$. In this case, $h(x) = \frac{1}{\sqrt{2\pi}}$, $\eta = [\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}]$ and $T(x) = [x, x^2]$. Thus, Gaussian distribution is included in the exponential family.

The density function of a exponential family can be written in the case of presence of an reproducing kernel Hilbert space $\mathcal{H}$ with a reproducing kernel k as :

$$p(x|\eta) = h(x)\ \exp\{\langle \eta(.), k(x,.)\rangle_{\mathcal{H}} - A(\eta)\}$$

with

$$A(\eta) = \log \int \exp\{\langle \eta(.), k(x,.)\rangle_{\mathcal{H}} h(x)dx$$

4. SVM based speaker segmentation

4.1 Speaker change detection using 1-class SVM and exponential family

Novelty change detection using SVM and exponential family is proposed by Canu and Smola (2005). Let $X = \{x_1, x_2, \ldots, x_N\}$ and $Y = \{y_1, y_2, \ldots, y_N\}$ two adjacent windows of acoustic feature vectors extracted from the audio signal ,where N is the number of data points in one window. Let Z denote the union of the contents of the two windows having $2N$ data points. The sequences of random variables X and Y are distributed according respectively to $\mathbb{P}x$ and $\mathbb{P}y$ distribution. We want to test if there exist a speaker turn after the sample x_N between the two windows. The problem can be viewed as testing the hypothesis $H_0 : \mathbb{P}_x = \mathbb{P}_y$ against the alternative $H1 : \mathbb{P}x \neq \mathbb{P}y$. H_0 is the null hypothesis and represents that the entire sequence is drawn from a single distribution, thus there not exist a speaker turn. While H_1 represents the hypothesis that there is a segment boundary after sample X_n. The likelihood ratio test of this hypotheses test is the following:

$$L(z_1, \ldots, z_{2N}) = \frac{\prod_{i=1}^{N} \mathbb{P}_x(z_i) \prod_{i=t+1}^{2N} \mathbb{P}_y(z_i)}{\prod_{i=1}^{2N} \mathbb{P}_x(z_i)} = \prod_{i=N+1}^{2N} \frac{\mathbb{P}_y(z_i)}{\mathbb{P}_x(z_i)} \tag{6}$$

since both densities are unknown the generalized likelihood ratio (GLR) has to be used :

$$L(z_1, \ldots, z_{2N}) = \prod_{i=N+1}^{2N} \frac{\widehat{\mathbb{P}}_y(z_i)}{\widehat{\mathbb{P}}_x(z_i)} \tag{7}$$

where $\hat{\mathbb{P}}_x$ and $\hat{\mathbb{P}}_y$ are the maximum likelihood estimates of the densities.

Assuming that both densities $\mathbb{P}_x$ and $\mathbb{P}_y$ are included in the generalized exponential

family, thus it exists a reproducing kernel Hilbert space $\mathcal{H}$ embedded with the dot product $< \cdot, \cdot >_{\mathcal{H}}$ with a reproducing kernel k such that:

$$\mathbb{P}_x(z) = h(z)\ \exp\{\langle \eta_x(.), k(z,.)\rangle_{\mathcal{H}} - A(\eta_x)\}$$

and

$$\mathbb{P}_y(z) = h(z)\ \exp\{\langle \eta_y(.), k(z,.)\rangle_{\mathcal{H}} - A(\eta_y)\}$$

Using One class SVM and the exponential family, a robust approximation of the maximum likelihood estimates of the densities $\mathbb{P}_x$ and $\mathbb{P}_y$ can be written as:

$$\widehat{\mathbb{P}}_x(z) = h(z) \exp\left(\sum_{i=1}^{N} \alpha_i^{(x)} k(z, z_i) - A(\eta_x)\right) \tag{8}$$

$$\widehat{\mathbb{P}}_y(z) = h(z) \exp\left(\sum_{i=N+1}^{2N} \alpha_i^{(y)} k(z, z_i) - A(\eta_y)\right) \tag{9}$$

where $\alpha_i^{(x)}$ is computed by solving the one class SVM problem on the first half of the data (z_1 to z_N), while $\alpha_i^{(y)}$ is given by solving the one class SVM problem on the second half of the data (z_{N+1} to z_{2N}). Using these three hypotheses, the generalized likelihood ratio test is approximated as follows:

$$L(z_1, \ldots, z_{2N}) = \prod_{j=N+1}^{2N} \frac{\exp\left(\sum_{i=N+1}^{2N} \alpha_i^{(y)} k(z_j, z_i) - A(\eta_y)\right)}{\exp\left(\sum_{i=1}^{N} \alpha_i^{(x)} k(z_j, z_i) - A(\eta_x)\right)} \tag{10}$$

A speaker change in the frame z_n exist if :

$$L(z_1, \ldots, z_{2N}) > s_x \quad \Leftrightarrow \quad \sum_{j=N+1}^{2N} \Big(\sum_{i=N+1}^{2N} \alpha_i^{(y)} k(z_j, z_i) - \sum_{i=1}^{N} \alpha_i^{(x)} k(z_j, z_i) \Big) > s'_x$$

where s_x is a fixed threshold. Moreover, $\sum_{i=N+1}^{2N} \alpha_i^{(y)} k(z_j, z_i)$ is very small and can be neglect in comparison with $\sum_{i=1}^{N} \alpha_i^{(x)} k(z_j, z_i)$. Then a speaker turn is detected when :

$$\sum_{j=N+1}^{2N} \Big(-\sum_{i=1}^{N} \alpha_i^{(x)} k(z_j, z_i)\Big) > s'_x \tag{11}$$

4.2 Proposed speaker segmentation technique

In section 4.1, we show that a speaker changes exist if the condition defined by the equation (11) is verified. This speaker change detection approach can be interpreted like this: to decide if a speaker change exit between the two windows X and Y, we built an SVM using the data X as learning data, then Y data is used for testing if the two windows are homogeneous or not.

On the other hand, since H_0 represent the hypothesis of $\mathbb{P}_x = \mathbb{P}_y$ the likelihood ratio test of the hypotheses test described in section 4.1 can be written like this:

$$L(z_1, \ldots, z_{2N}) = \frac{\prod_{i=1}^{N} \mathbb{P}_x(z_i) \prod_{i=t+1}^{2N} \mathbb{P}_y(z_i)}{\prod_{i=1}^{2N} \mathbb{P}_y(z_i)} = \prod_{i=1}^{N} \frac{\mathbb{P}_x(z_i)}{\mathbb{P}_y(z_i)} \tag{12}$$

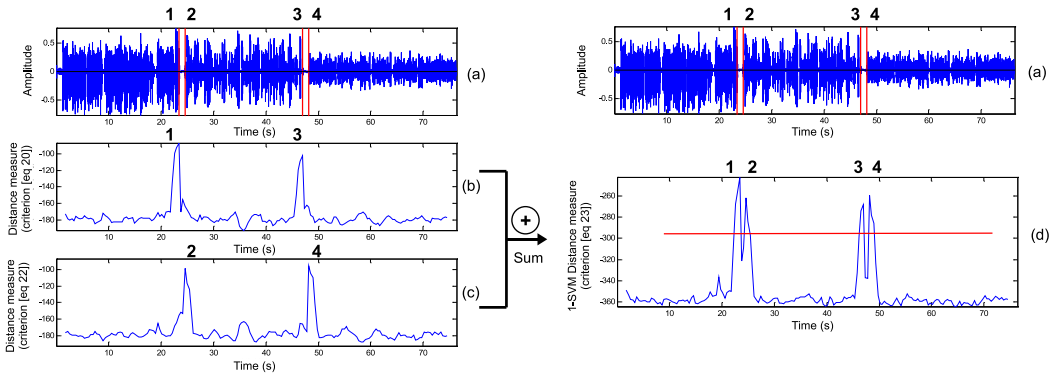

Figure 1: Segmentation results of an audio stream extracted from NIST RT-02 broadcast news data using criteria defined by eq. (11) (subplot b), eq. (13) (subplot c) and eq. (14) (subplot d).

Using the same gait, a speaker change has occurred if:

$$\sum_{j=1}^{N}(-\sum_{i=N+1}^{2N}\alpha_i^{(y)}k(z_j,z_i)) > s'_y \tag{13}$$

Experimental tests show that in some case is more appropriate when we use Y data for learning and X data for testing. Figure 1 presents the segmentation of an audio stream which presents four speaker changes. This audio stream is a sample of broadcast news extracted from NIST RT-02 data. Figures (b) and (c) represent the result of segmentation using respectively (11) and (13). Using the criteria (11), we can detect only changes number 1 and 3 and using the criteria (13), we can detect only changes number 2 and 4. For these reason it is more appropriate to use the criterion described as follow:

$$\sum_{j=N+1}^{2N}(-\sum_{i=1}^{N}\alpha_i^{(x)}k(z_j,z_i)) + \sum_{j=1}^{N}(-\sum_{i=N+1}^{2N}\alpha_i^{(y)}k(z_j,z_i)) > S \tag{14}$$

In this case and as illustrated in figure 1, we can detect easily all speaker changes.

4.2 Our segmentation method

Our technique detects speaker turns by computing the distance detailed in equation (14) between a pair of adjacent windows of the same size shifted by a fixed step along the whole parameterized speech signal. In the end of this procedure we obtain the curve of the variation of the distance in time. The analysis of this curve shows that a speaker change point is characterized by the presence of a "significant" peak. A peak is regarded as "significant" when it presents a high value. So, break points can be detected easily by searching the local maxima of the distance curve that presents a value higher than a fixed threshold.

Algorithm 1: Speaker change detection algorithm

Step 0: Initialization

- initialize the interval $[a, b]$, $a = 0, b =$ SIZE_WINDOW

Step 1: Computing detection criterion

- Compute the distance measure $d1$ according to equation (11) with $[a, b/2]$ testing data and $[b/2 + 1, b]$ training data.
- Compute the distance measure $d2$ according to equation (13) with $[b/2 + 1, b]$ testing data and $[a, b/2]$ training data
- Compute the decision criterion $d = d1 + d2$
- a=$a + pas$ and $b = b + pas$; go to step 1

Step 2: speaker turns detection

- detecting peaks of d-curve, $p = p_i$
- decision:
  - if $d(p_i) > s$ a speaker change is detected,
  - if $d(p_i) < s$ no speaker change is detected,

---

5. Experiments

5.1 Data set

In order to evaluate 1-SVM-based segmentation method, experiments are based essentially on the segmentation of IDIAP meetings Corpus. This database contains two separate test sets sampled at 16 kHz. The first test set contains only single speaker segments without overlapping. However the second one contains a short overlap segment included at each speaker change. Further, to generalize our experiments, we used also other types of audio streams like broadcast news and telephone conversations. These audio streams are extracted from the Rich Transcription-04 MDE Training Data Speech corpus created by Linguistic Data Consortium (LDC). Description of the used datasets is presented below:

1. IDIAP meetings (Moore, 2002):
   - Test set 1: contains only single speaker segments without overlap segments. This test set groups nine files, each of them contains 10 speaker turns constructed in a random manner with segments duration varying from 5 to 20 seconds. The total test set duration was 20 minutes.
   - Test set 2: contains a short overlap segment included at each speaker change. The test set is formed by six files, each containing 10 single speaker segments (of between 5-17 seconds duration), interleaved with 9 segments of dual-speaker overlap (of between 1.5-5 seconds duration).
2. Broadcast news data: is composed of three approximately 10-minute excerpts from three different broadcasts. The broadcasts were selected from programs from NBC, CNN and ABC, all collected in 1998.
3. Telephone conversation: is composed of a 10-minute excerpt from a conversation between two switchboard operators.

5.2 Evaluation criteria

For evaluating the performance of the segmentation task, we use Type-I errors: precision (PRC) and Type-II errors: recall (RCL) was widely used in previous research (Ajmera et al., 2004). Type-I errors occur if a true change is not spotted (missed alarm) within a certain window. Type-II errors occur when a detected change does not correspond to a true change in the reference (false alarm). Precision (PRC) and recall (RCL) are defined as below:

$$\text{PRC} = \frac{\text{number of correctly found changes}}{\text{Total number of changes found}} \tag{15}$$

$$\text{RCL} = \frac{\text{number of correctly found changes}}{\text{Total number of correct changes}} \tag{16}$$

In order to compare the performance of different systems, the F-measure is often used and is defined as

$$F = \frac{2.0 \times \text{PRC} \times \text{RCL}}{\text{PRC} + \text{RCL}} \tag{17}$$

The F-measure varies from 0 to 1, with a higher F-measure indicating better performance.

5.3 Audio features components

In the experiments, two kinds of feature vectors are proposed: MFCCs and DWCs. Mel frequency cepstral coefficients (MFCCs) are a short-time spectral decomposition of audio that convey the general frequency characteristics important to human hearing. We calculate MFCCs by using overlapping frames of 30 ms. The Discrete Wavelet Coefficients (DWCs) are computed by applying the Discrete Wavelet Transform (DWT) which provides a time-frequency representation of the signal. It was developed to overcome the short coming of the Short Time Fourier Transform (STFT), which can also be used to analyze non-stationary signals. While STFT gives a constant resolution at all frequencies, the Wavelet Transform uses multi-resolution technique by which different frequencies are analyzed with different resolutions. The DWT is computed by successive low-pass and high-pass filtering of the discrete time-domain signal. This is called the Mallat algorithm or Mallat-tree decomposition (Mallat, 1998).

5.3.1 Mel frequency cepstral coefficient

MFCCs are a short-time spectral decomposition of audio that convey the general frequency characteristics important to human hearing. While originally developed to decouple vocal excitation from vocal tract shape for automatic speech recognition. In order to calculate MFCCs, the signal is first broken into overlapping frames, each approximately 25 ms long, a time scale at which the signal is assumed to be stationary. The log-magnitude of the discrete Fourier transform of each window is warped to the Mel frequency scale, imitating human frequency and amplitude sensitivity. The inverse discrete cosine transform decorrelates these "auditory spectra" and the so called "high time" portion of the signal, corresponding to fine spectral detail, is discarded, leaving only the general spectral shape

5.3.2 Discrete Wavelet transform

The Wavelet Transform provides a time-frequency representation of the signal. It was developed to overcome the short coming of the Short Time Fourier Transform (STFT), which can also be used to analyze non-stationary signals. While STFT gives a constant resolution

at all frequencies, the Wavelet Transform uses multi-resolution technique by which different frequencies are analyzed with different resolutions. The DWT is computed by successive low-pass and high-pass filtering of the discrete time-domain signal. This is called the Mallat algorithm or Mallat-tree decomposition (Mallat, 1998). Its significance is in the manner it connects the continuous-time mutiresolution to discrete-time filters.

5.4 Results

Table 1 illustrates speaker segmentation experiments conducted on the various audio documents previously described and their corresponding results using 1-SVMs and $DIS\_T^2\_BIC$ approaches. Segmentation using 1-SVMs outperforms $DIS\_T^2\_BIC$ based segmentation technique for all the tested audio documents. The segmentation of the IDIAP meetings test set 1 using the two methods presents the highest value of precision and recall. In fact, opposite to other types of audio streams, this corpus contains long speech segments allowing good estimation of data. As presented in the table 1, the PRC and RCL values obtained with IDIAP meetings(1) increases respectively from 0.69 to 0.8 and from 0.68 to 0.79.

Table 1: Segmentation results using the proposed 1-SVM and $DIS\_T^2\_BIC$ methods.

Audio	1-SVM method				$DIS\_T^2\_BIC$ method			
Streams	Features	RCL	PRC	F	Features	RCL	PRC	F
M. IDIAP1	39MFCC+DWC_5	0.8	0.79	0.79	13MFCC	0.69	0.68	0.68
M. IDIAP2	39MFCC+DWC_5	0.68	0.67	0.67	13MFCC	0.58	0.56	0.57
B. News	39MFCC+DWC_6	0.75	0.75	0.75	39MFCC	0.63	0.66	0.64
Tel. Conv	39MFCC+DWC_3	0.72	0.71	0.71	13MFCC	0.56	0.58	0.57

The proposed method based on 1-SVMs allows the improvement of speaker change detection in audio streams which contain overlapping speeches. The improvement in the PRC and RCL values using IDIAP meetings test set 2 is more than 10% with respect to $DIS\_T^2\_BIC$ method. Generally, BIC based segmentation techniques detect a speaker change between two adjacent analysis windows. Each window is modeled by a Gaussian distribution. This supposition is not true when the window contains overlapped speeches. In this case, it is more suitable to suppose that each window can be modeled by an exponential family.

Broadcast news segmentation results are enhanced by adding discrete wavelet coefficients to cepstral coefficients. The use of this kind of parametrization makes speaker changes detection possible in the presence of background noise. Further, deploying 1-SVMs permits to better put in evidence this characteristic since it is insensitive to the dimension of acoustic features. Also, the proposed method is more appropriate to detect speaker changes close each others. The F value obtained with the segmentation results of the telephone conversation is raised from 0.56 with $DIS\_T^2\_BIC$ method to 0.71 with 1-SVMS method.

6. Conclusion

In this chapter, we have proposed a new unsupervised detection algorithm based on 1-SVMs. This algorithm outperforms model-selection based detection methods. Using the exponential family model, we obtain a good estimation of the generalized Likelihood ratio applied on the known hypothesis test generally used in change detection tasks. By adding to cepstral coefficients the discrete wavelet coefficients. The use of this kind of parametrization permitted to detect speaker changes even in real-world conditions in which the environment

and context are so complex that the segmentation results are often affected. The use of support vector machines permit to deal practically with this high dimensional acoustic features vector. Experimental results present higher precision and recall values than those obtained with $DIS\_T^2\_BIC$ technique, the increase of PRC and RCL values obtained with various kinds of audio streams is roughly over 10%.

References

J. Ajmera, I. McCowan, and H. Bourlard. Robust speaker change detection. *IEEE Signal Processing Letters*, pages 649–651, 2004.

T. Anderson. *An introduction to multivariate statistical analysis.* John Wiley and Sons, New York, NY, 1985.

X. Anguera, C. Wooters, and J. LM. Pardo. Robust speaker diarization for meetings: ICSI RT06s evaluation system. In *ICSLP'06*, Pittsburgh, Pensilvania, USA, 2006.

X. Anguera, C. Wooters, and J. Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech and Language Processing*, 15:2011–2023, 2007.

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc*, 1950.

J. F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, and C. Wellekens. A speaker tracking system based on speaker turn detection for nist evaluation. In *ICASSP'00*, pages 1177–1180, Istanbul, Turkey, 2000.

S. Canu and A. Smola. Kernel methods and the exponential family. In *ESANN'05*, Brugge, Belgium, 2005.

M. Cettolo and M. Federico. Model selection criteria for acoustic segmentation. In *ISCA Tutorial and Research Workshop ASR*, 2000.

M. Cettolo and M. Vescovi. Efficient audio segmentation algorithms based on the BIC. In *ICASSP'03*, 2003.

S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

M. Davy and S. Godsill. Detection of abrupt spectral changes using Support Vector Machines. an application to audio signal segmentation. In *ICASSP'02*, volume 2, pages 1313–1316, Orlando, USA, May 2002.

M. Davy, F. Desobry, A. Gretton, and C. Doncarli. An online Support Vector Machine for abnormal events detection. *Signal Processing*, 86(8):2009–2025, August 2006.

P. Delacourt and C. J. Wellekens. DISTBIC: a speaker based segmentation for audio data indexing. *Speech Communication*, 32:111–126, 2000.

F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(5), May 2005.

A. Dielmann and S. Renals. Automatic meeting segmentation using dynamic bayesian networks. *IEEE Transactions on MultiMedia*, pages 25–36, 2007.

J. Hansen, R. Huang, B. Zhou, M. Deadle, J. Deller, A.R. Gurijala, M. Kurimo, and P. Angkititraku. Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word. *IEEE Trans. Speech Audio Process*, pages 712–730, 2005.

R. Huang and J. H. L. Hansen. Advances in unsupervised audio segmentation for the broadcast news and ngsw corpora. In *ICASSP'04*, pages 741–744, 2004.

Q. Jin and T. Schultz. Speaker segmentation and clustering in meetings. *INTERSPEECH'04*, pages 597–600, 2004.

H. Kadri, Z. Lachiri, and N. Ellouze. Hybrid approach for unsupervised audio speaker segmentation. In *Europpean Signal Processing Conference, EUSIPCO'06*, Florence, Italy, 2006.

H. Kadri, M. Davy, A. Rabaoui, Z. Lachiri, and N. Ellouze. Robust audio speaker segmentation using one class svms. In *Europpean Signal Processing Conference, EUSIPCO'08*, Lausanne, Switzeland, 2008.

T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *ICASSP'00*, pages 1423–1426, 2000.

S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998.

D. Moore. The idiap smart meeting room. In *IDIAPCOM 07*, 2002.

D. Moraru, S. Meignier, L. Besacier, J. F. Bonastre, and I. Magrin-Chagnolleau. The elisa consortium approaches in speaker segmentation during the nist 2002 speaker recognition evaluation. *ICASSP'04*, 2003.

D. Reynolds and T. Carrasquillo. The MIT Lincoln Laboratories RT-04F diarization systems: Applications to broadcast audio and telephone conversations. In *Fall 2004 Rich Transcription Workshop (RT04)*, Palisades, NY., USA, 2004.

D. Reynolds and T. Carrasquillo. Approaches and applications of audio diarization. *ICASSP'05*, 2005.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2002.

B. Schölkopf, R. Williamson, A. Smola, and J. Shawe-Taylor. Support Vector method for novelty detection. In *NIPS*, pages 582–588, 2000.

S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech and Language Processing*, pages 1557–1565, 2006.

S. Wegmann, P. Zhan, and L. Gillick. Progress in broadcast news transcription at dragon systems. In *ICASSP'99*, Phoenix, Arizona, USA, 1999.

B. W. Zhou and J. H. L. Hansen. Unsupervised audio stream segmentation and clustering via the bayesian information criterion. In *ICSLP'00*, pages 714–717, Beijing, China, 2000.