16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

ROBUST AUDIO SPEAKER SEGMENTATION USING ONE CLASS SVMS

Hachem Kadri[1], *Manuel Davy*[2], *Asma Rabaoui*[1], *Zied Lachiri*[1] *and Noureddine Ellouze*[1]

[1]Unité de recherche Signal, Image et Reconnaissance des formes
ENIT, BP 37, Campus Universitaire, 1002 le Belvédère, Tunis Tunisia.
[2] LAGIS, UMR CNRS 8146, and INRIA SequeL Team, BP 48, Cité Scientifique
59651 Villeneuve d'Ascq Cedex, Lille France.
emails: hachem.kadri@gmail.com, Manuel.Davy@inria.fr

ABSTRACT

This paper presents a new technique for segmenting an audio stream into pieces, each one contains speeches of only one speaker. Speaker segmentation has been used extensively in various tasks such as automatic transcription of radio broadcast news and audio indexing. The segmentation method used in this paper is based on a discriminative distance measure between two adjacent sliding windows operating on preprocessed speech. The proposed unsupervised detection method which does not require any pre-trained models is based on the use of the exponential family model and 1-SVMs to approximate the generalized likelihood ratio. Our 1-SVM-based segmentation algorithm provides improvements over baseline approaches which use the Bayesian Information Criterion (BIC). The segmentation results achieved in our experiments illustrate the potential of this method in detecting speaker changes in audio streams containing overlapped and short speeches.

1. INTRODUCTION

Speaker segmentation has been generally referred to as speaker change detection and was closely related to acoustic change detection. For a given speech/audio stream, speaker segmentation/ change detection systems find the times when there is a change of speaker in the audio [1, 2, 3]. Detection of speaker changes is a crucial element of speech recognition and speaker recognition engines. Besides improving automatic speech recognition systems, audio segmentation is also useful in many other interesting practical applications such as content based audio classification and retrieval, audio archive management, surveillance, etc [4, 5, 6].

Recently, three main domains of application for speaker segmentation have received special attention [5] :

- 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.

The BIC is probably the most extensively used model-selection segmentation method due to its simplicity and effectiveness. Several speaker segmentation approaches using BIC have been proposed. Initially, in [1] a multiple changing point detection algorithm was proposed. Then, [3, 7] present one or two-pass algorithms using a growing window with inner variable length analysis segments to iteratively find the changing points. In [8], some ways to make the algorithm faster were developed. Even with the efforts to speed up the processing of BIC, it is computationally more intensive than other statistics-based metrics when used to analyze the signal with high resolution, but its good performance has kept it as the algorithm of choice in many applications. This is why some people have proposed a hybrid approach in which BIC is used as the second pass (refinement) of two-pass speaker segmentation system [4, 9, 10]. As described earlier, an important step in this direction is taken with the use of Hotelling's T^2 distance as a first step for detecting short speaker changes [10, 6].

The main focus of this paper 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 objective is to determine whether N consecutive audio frames constitute a single homogeneous window W or two different windows: W_1 and W_2. In order to detect if an abrupt change occurred at the i^{th} frame 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 speech and very short changes. To solve this problem, our proposed technique uses 1-SVMs and exponential family model to maximize the generalized likelihood ratio with any probability distribution of windows.

The remainder of this paper is organized as follows. Section 2 details audio segmentation techniques based on BIC. The proposed speaker change detection method is presented in section 3. Experimental results are provided in Section 4. Section 5 concludes the paper with a summary and discussion.

2. BIC BASED SEGMENTATION TECHNIQUES

BIC [1] 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 applied to it indicates how well the model fits the data, and is determined by:

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

$\log L(X_i, M_i)$ is the log-likelihood of the data given the considered model, λ is a free design parameter dependent on the data being modeled; N_i is the number of frames in the considered segment and $\#(M_i)$ the number of free parameters to estimate in model M_i.

The BIC-based segmentation procedure is based on the measure of the ΔBIC value between two consecutive audio segments containing parameterized acoustic vectors X_1 and X_2 of length N_1 and N_2 respectively . ΔBIC represents the difference in BIC scores between two models: one

suppose that the two segments are generated by the same Gaussian distribution $M(\mu,\Sigma)$ and the other suppose that each segment is generated by a different Gaussian distribution ($M_1(\mu_1,\Sigma_1) \neq M_2(\mu_2,\Sigma_2)$).

$$\Delta BIC = \log \frac{L(X_1+X_2,M)}{L(X_1,M_1)+L(X_2,M_2)} - \Delta\#(M_i)\log(N_1+N_2)$$

where $\Delta\#(i,j)$ is the difference between the number of free parameters in the combined model versus the two individual models. A positif ΔBIC value indicate the presence of a speaker change between the two audio segments

BIC-based segmentation is more suitable to validate speaker changes when we have enough data for a good estimation. That is why we associate with BIC a metric-based segmentation allowing obtaining a hybrid approach. In our previous work [10], we developed a hybrid segmentation technique called $DIS\_T^2\_BIC$ that uses the Hotelling's T^2 statistic as a first step followed by the BIC as a second step. Speaker change detection by Hotelling's T^2 distance is based on the following concept: for the two audio segments X_1 and X_2, if they can be modelled by multivariate Gaussian distributions: $M_1(\mu_1,\Sigma_1)$ and $M_2(\mu_2,\Sigma_2)$, we assume their covariances are equal but unknown, then the only difference between them is the mean values reflected in the T^2 distance as:

$$T^2 = \frac{N_1 \cdot N_2}{N_1+N_2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \quad (2)$$

Under the equal covariance assumption, we can use more data to estimate the covariance and reduce the impact of insufficient data in the estimation. Hence, The T^2 distance can detect more short speaker changes than BIC.

3. SVM BASED SPEAKER SEGMENTATION

3.1 1-class SVM

The One-class approach [11] has been successfully applied to various problems such as outlier detection and novelty detection [12, 13]. Its first application is outlier detection, to detect uncharacteristic objects from a dataset, examples which do not resemble the rest of the dataset in some way. 1-SVM [11] distinguishes one class of data from the rest of the feature space given only a positive data set and never sees the outlier data. Instead it must estimate the boundary that separates those two classes based only on data which lies on one side of it. The problem therefore is to define this boundary in order to minimize misclassifications.

The aim of 1-SVMs is to use the training dataset $\mathcal{X}$ in $\mathbb{R}^d$ so as to learn a function $\mathsf{f}_{\mathcal{X}} : \mathbb{R}^d \mapsto \mathbb{R}$ such that most of the data in $\mathcal{X}$ belong to the set $\mathcal{R}_{\mathcal{X}} = \{x \in \mathbb{R}^d \text{ with } \mathsf{f}_{\mathcal{X}}(x) \geq 0\}$ while the volume of $\mathcal{R}_{\mathcal{X}}$ is minimal. This problem is termed *minimum volume set (MVS) estimation* and we see that membership of x to $\mathcal{R}_{\mathcal{X}}$ indicates whether this datum is overall similar to $\mathcal{X}$, or not. 1-SVMs solve MVS estimation in the following way. First, a so-called *kernel function* $k(\cdot,\cdot)$; $\mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is selected, here, we assume a Gaussian RBF kernel such that $k(x,x') = \exp\ -\|x-x'\|^2/2\sigma^2$. This kernel induces a so-called *feature space* denoted $\mathcal{H}$ and makes the evaluation of $k(x,x')$ a linear operation in $\mathcal{H}$, whereas it is nonlinear in $\mathbb{R}^d$. In practice, let the separating hyperplane $\mathcal{W} = \{h(\cdot) \in \mathcal{H} \text{ with } \langle h(\cdot), w(\cdot)\rangle_{\mathcal{H}} - \rho = 0\}$, then its parameters $w(\cdot)$ and ρ results from the optimization problem

$$\min_{\mathrm{w},\xi,\rho} \frac{1}{2}\|w(\cdot)\|^2_{\mathcal{H}} + \frac{1}{\nu m}\sum_{j=1}^{m}\xi_j - \rho \quad (3)$$

subject to (for $i = 1,\ldots,m$)

$$\langle w(\cdot), k(x_j,\cdot)\rangle_{\mathcal{H}} \geq \rho - \xi_j, \quad \text{and } \xi_j \geq 0 \quad (4)$$

where ν tunes the fraction of data that are allowed to be on the wrong side of $\mathcal{W}$ (these are the outliers and they do not belong to $\mathcal{R}_{\mathcal{X}}$) and ξ_j's are so-called slack variables. It can be shown [11] that a solution of (3)-(4) is such that $w(\cdot) = \sum_{j=1}^{m}\alpha_j k(x_j,\cdot)$ where the α_j's verify the dual optimization problem

$$\min_{\alpha} \frac{1}{2}\sum_{j,j'=1}^{m}\alpha_j\alpha_{j'}k(x_j,x_{j'}) \quad (5)$$

subject to

$$0 \leq \alpha_j \leq \frac{1}{\nu m}, \quad \sum_j \alpha_j = 1 \quad (6)$$

Finally, the decision function is $\mathsf{f}_{\mathcal{X}}(x) = \sum_{j=1}^{m}\alpha_j k(x_j,x) - \rho$ and ρ is computed by using that $\mathsf{f}_{\mathcal{X}}(x_j) = 0$ for those x_j's in $\mathcal{X}$ that are located onto the boundary, i.e., those that verify both $\alpha_j \neq 0$ and $\alpha_j \neq 1/\nu m$.

3.2 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:

$$p(x|\eta) = h(x)\ exp\{\eta^T T(x) - A(\eta)\} \quad (7)$$

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 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 [14] :

$$p(x|\eta) = h(x)\ \exp\{\langle\eta(.),k(x,.)\rangle_{\mathcal{H}} - A(\eta)\} \quad (8)$$

with

$$A(\eta) = \log\int \exp\{\langle\eta(.),k(x,.)\rangle_{\mathcal{H}}h(x)dx \quad (9)$$

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

Novetly change detection using SVM and exponential family is proposed by Canu and Smola [15] [14]. 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 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 :

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

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

$$L(z_1,\ldots,z_{2N}) = \prod_{i=N+1}^{2N}\frac{\widehat{\mathbb{P}}_y(z_i)}{\widehat{\mathbb{P}}_x(z_i)} \quad (10)$$

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 (see eq 16):

$$\mathbb{P}_x(z) = h(z) \ \exp\{\langle \eta_x(.), k(z,.)\rangle_{\mathcal{H}} - A(\eta_x)\} \qquad (11)$$

and

$$\mathbb{P}_y(z) = h(z) \ \exp\{\langle \eta_y(.), k(z,.)\rangle_{\mathcal{H}} - A(\eta_y)\} \qquad (12)$$

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:

$$\widehat{\mathbb{P}}_x(z) = h(z) \exp(\sum_{i=1}^{N} \alpha_i^{(x)} k(z, z_i) - A(\eta_x) \qquad (13)$$

$$\widehat{\mathbb{P}}_y(z) = h(z) \exp(\sum_{i=N+1}^{2N} \alpha_i^{(y)} k(z, z_i) - A(\eta_y) \qquad (14)$$

where $\alpha_i^{(x)}$ i is determined 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_{t+N} to z_{2N}). Using these three hypotheses, the generalized likelihood ratio test is approximated as follows:

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

A speaker change in the frame z_n exist if :

$$L(z1, \ldots, z_{2t}) > s_x \quad \Leftrightarrow$$

$$\sum_{j=N+1}^{2N} (\sum_{i=N+1}^{2N} \alpha_i^{(y)} k(z_j, z_i) - \sum_{i=1}^{N} \alpha_i^{(x)} k(z_j, z_i)\ ) > s'_x \qquad (15)$$

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 :

$$\sum_{j=N+1}^{2N} (-\sum_{i=1}^{N} \alpha_i^{(x)} k(z_j, z_i)) > s'_x \qquad (16)$$

3.4 The proposed speaker segmentation technique

In subsection 3.3, we show that a speaker changes exist if the condition defined by the equation (16) 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 homogenous 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 A can be written like this:

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

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

$$\sum_{j=1}^{N} (-\sum_{i=N+1}^{2N} \alpha_i^{(y)} k(z_j, z_i)) > s'_y \qquad (17)$$

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 (16) and (17). Using the criteria (16), we can detect only changes number 1 and 3 and using the criteria (17), we can detect only changes number 2 and 4. For these reason it is more appropriate to use the criterion described as follow:

$$\sum_{j=N+1}^{2N} (-\sum_{i=1}^{N} \alpha_i^{(x)} k(z_j, z_i)) + \sum_{j=1}^{N} (-\sum_{i=N+1}^{2N} \alpha_i^{(y)} k(z_j, z_i)) > S \qquad (18)$$

In this case and as illustrated in Figure 1, we can detect easily all speaker changes. Our technique detects speaker turns by computing the distance detailed in equation (18) 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.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 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 [16]:
   - 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.

4.2 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 [17]. Type-I

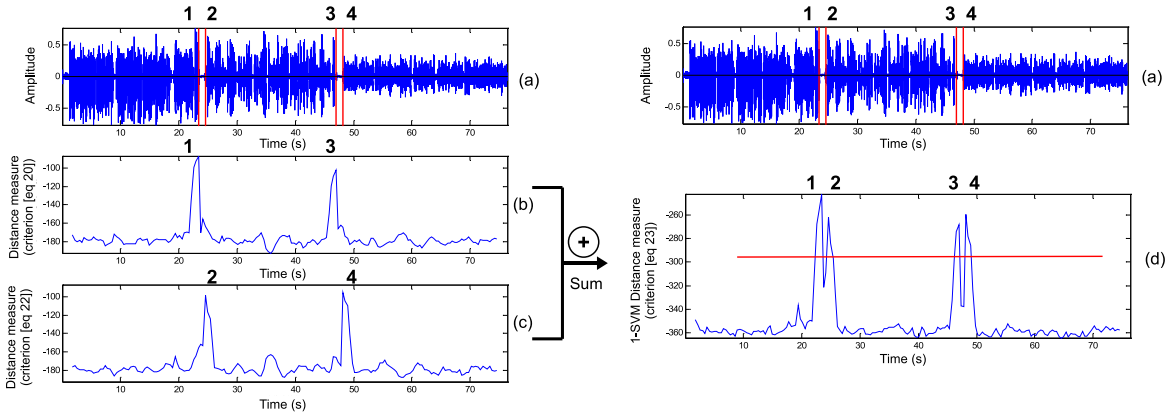

Figure 1: Segmentation of an audio segment containing four speaker changes. (a): The audio segment. (b), (c) and (d): 1-SVM based distance measures using criteria defined respectively by (16), (17) and (18). Distance (16) detects speaker changes **1** and **3** and distance (17) detects changes **2** and **4**. The proposed distance (18) detects all the changes.

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:

$$\text{PRC} = \frac{\text{number of correctly found changes}}{\text{Total number of changes found}} \quad (19)$$

$$\text{RCL} = \frac{\text{number of correctly found changes}}{\text{Total number of correct changes}} \quad (20)$$

$$(21)$$

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

$$F = \frac{2.0 \times \text{PRC} \times \text{RCL}}{\text{PRC} + \text{RCL}} \quad (22)$$

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

4.3 Evaluation

4.3.1 Audio parametrization

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 lowpass and highpass filtering of the discrete time-domain signal. This is called the Mallat algorithm or Mallat-tree decomposition [18].

4.3.2 Segmentation 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(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.

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(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 modelized 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 modelized 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 presents it's 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.

5. CONCLUSION

In this paper, 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.

Adding to cepstral coefficients the discrete wavelet coefficients 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 allow to deal practically with this high dimensional acoustic features vector. Experi-

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

	1-SVM method				$DIS\_T^2\_BIC$ method			
	Features	RCL	PRC	F	Features	RCL	PRC	F
IDIAP meetings (1)	39 MFCC+DWC_5	0.8	0.79	0.79	13 MFCC	0.69	0.68	0.68
IDIAP meetings (2)	39 MFCC+DWC_5	0.68	0.67	0.67	13 MFCC	0.58	0.56	0.57
Broadcast news	39 MFCC+DWC_6	0.75	0.75	0.75	39 MFCC	0.63	0.66	0.64
Telephone conversation	39 MFCC+DWC_3	0.72	0.71	0.71	13 MFCC	0.56	0.58	0.57

mental 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

[1] 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.

[2] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *ICASSP*, 2000, pp. 1423–1426.

[3] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Eurospeech*, 1999, pp. 679–682.

[4] P. Delacourt and C. Wellekens, "DISTBIC: a speaker based segmentation for audio data indexing," *Speech Communication*, vol. 32, pp. 111–126, 2000.

[5] 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.

[6] B. Zhou and J. Hansen, "Unsupervised audio stream segmentation and clustering via the bayesian information criterion," in *ICSLP*, Beijing, China, 2000, pp. 714–717.

[7] L. Lu and H. Zhang, "Real-time unsupervised speaker change detection," in *ICPR'02*, Quebec City, Canada.

[8] M. Cettolo and M. Vescovi, "Efficient audio segmentation algorithms based on the bic," in *ICASSP*, 2003.

[9] S. Sian-Cheng and H. Min-Wang, "Metric-seqdac: a hybrid approach for audio segmentation," in *Interspeech'04*.

[10] H. Kadri, Z. Lachiri, and N. Ellouze, "Hybrid approach for unsupervised audio speaker segmentation," in *Europpean Signal Processing Conference (EUSIPCO)*, Florence, Italy, 2006.

[11] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, USA: MIT Press, 2002.

[12] B. Fergani, M. Davy, and A. Houacine, "Speaker diarization using one class support vector machines," in *Speech Communication, to appear*, 2008.

[13] M. Davy and S. Godsill, "Detection of Abrupt Spectral Changes using Support Vector Machines. An Application to Audio Signal Segmentation," in *IEEE ICASSP*, vol. 2, Orlando, USA, May 2002, pp. 1313–1316.

[14] A. Smola, "Exponential families and kernels," Berder summer school, http://users.rsise.anu.edu.au/smola/teaching/summer2004/, Tech. Rep., 2004.

[15] S. Canu and A. Smola, "Kernel methods and the exponential family," in *ESANN'05*, Brugge, Belgium, 2005.

[16] D. Moore, "The idiap smart meeting room," in *IDIAP-COM 07*, 2002.

[17] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE SignalProcessing Letters*, pp. 649–651, 2004.

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