A Video Scene Detection Method Using Audio Information

Yu Li Shein
University of Computer Studies, Yangon
ulishein@gmail.com

Abstract

The role of audio in the context of multimedia applications involving video is becoming increasingly important. In this paper, an approach to automatic segmentation and classification of audiovisual data based on audio content analysis is proposed. Specifically, an audio classification scheme is developed to partition the sound-track of a video into homogeneous audio segments such as speech, music and speech with music background signal. Audio features which are extracted from both time and frequency domain are employed to ensure the feasibility in scene change detection. A hierarchical classification approach is applied for fast segmentation and detection. A Support Vector Machine (SVM) is firstly used to detect scene with music. Then Gaussian Mixture Model (GMM) is adopted to classify the rest scenes into either scene containing speech only or scene consisting of speech with music background. The experiments on real documentary videos show that proposed approach provides satisfactory detection rates.

1. Introduction

The task of automatic segmentation, indexing, and retrieval of audiovisual data has important applications in professional media production, audiovisual archive management, education, entertainment, surveillance, and so on. For example, a vast amount of audiovisual material has been archived in television and film database. If these data can be properly segmented and indexed, it will facilitate the retrieval of desired video segments for the editing of a documentary or an advertisement video clip. To give another example, in audiovisual libraries or family entertainment applications, it will be convenient to users if they are able to retrieve and watch video segments of their interests. As the volume of the available material becomes huge, manual segmentation and indexing is impossible. Automatic segmentation and indexing through computer processing based on video content analysis is clearly the trend. Acceptable results have been obtained with classification schemes, hidden Markov models (HMMs) in [1], zero crossing rate (ZCR) in [2], fixed thresholds [3, 4] and maximum-likelihood and entopic prior HMM in [5].

Recent works on audio classification have shown that hierarchical classification has better performance than a single binary classifier for different classes of audio. Therefore, in this paper, SVM and GMM classifiers are used serially to classify scenes into speech, music and speech with music background scenes. First, the sound track of documentary videos is supplied into system as input. The features are calculated from the input audio data. The feature sets are fed into the SVM classifier which classifies speech and pure music. The output speech segments of SVM classifier are again fed into the GMM classifier which classifies speech and speech with music.

2. Background

2.1. Feature extraction

As the first step, audio features are extracted from audio files. To be an efficient system, features are extracted from time domain as well as from frequency domain. In feature extraction process of this system, the features such as short-time energy (STE), zero-crossing rate (ZCR), high zero-crossing rate ratio (HZCRR), low short-time energy ratio (LSTER), spectral flux(SFLUX) and a time-complexity measure (sample entropy) are employed.

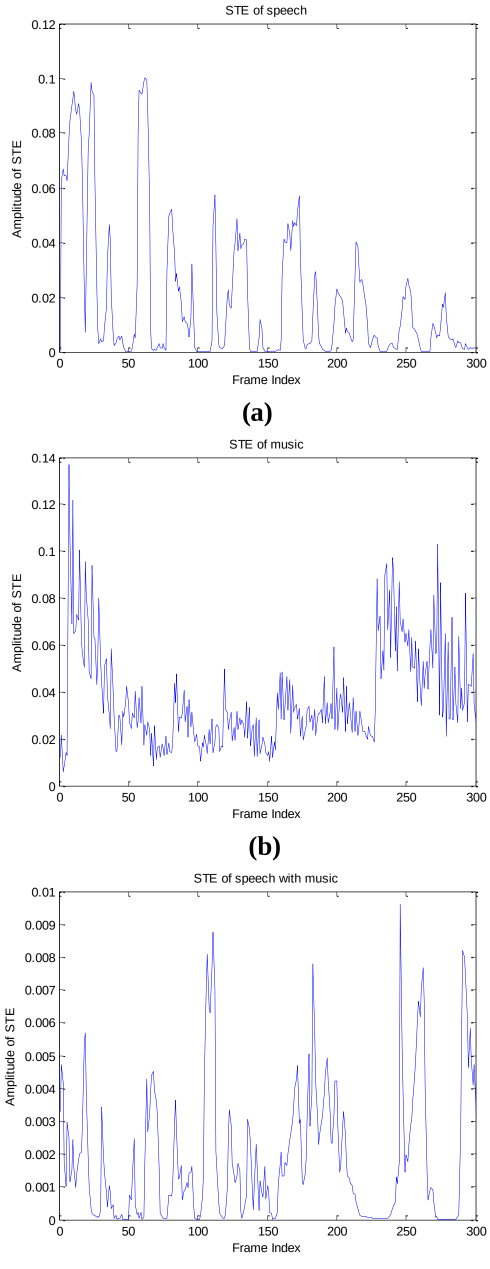
Given the audio track (with sampling frequency of 22kHz) of documentary videos, it is uniformly segmented into non-overlapping 1s clips, and then 9 features are extracted to represent each clip, which are chosen based on their effectiveness in capturing both temporal and spectral structures of different audio classes. For every 20ms audio frame which advances for every 10 ms, each feature is calculated. In the following sub-sections, the details of the features extraction process are described.

2.1.1. Short-time energy

STE provides a convenient representation of the signal's amplitude evolutions over time. The short time energy function of an audio signal is defined as

$$E_n = \frac{1}{N}\sum\left[x(m)w(n-m)\right]^2 \qquad (1)$$

where *x(m)* is discrete time audio signal; *n* is time index of the short-time energy; *w(m)* is window of length N. Mean, Variance of short-time energy then can be calculated. Figure 1(a), 1(b) and 1(c) show the plots of short time energy calculated over a speech signal, a music excerpt and a speech with music background of 3s long.



(a)

(b)

(c)

Fig 1. Short time energy of 3s (a) speech (b) music and (c) speech with music background

2.1.2 Low short time energy ratio

LSTER can be defined as the ratio of the number of frames whose STE values may be less than 0.5 times of the average STE, to the total number of frames.

2.1.3 Short-time zero-crossing rate

The short-time zero crossing rate is defined as:

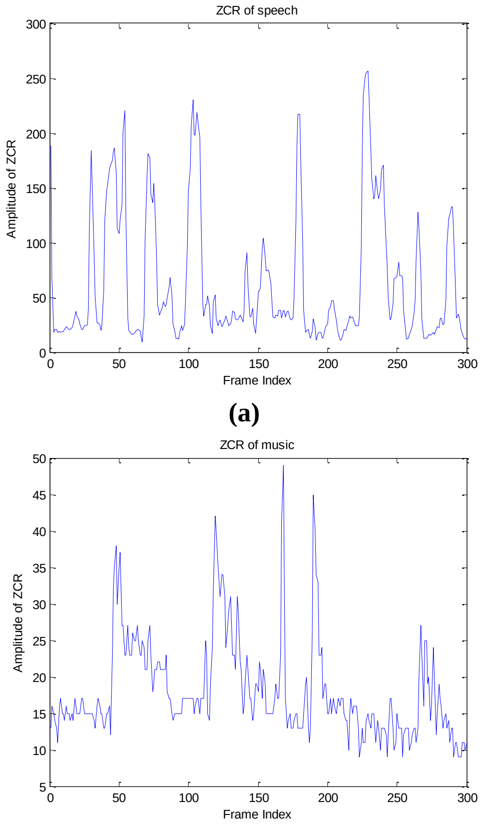
$$Z_n = \frac{1}{2}\sum_m \left|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\right| w(n-m) \qquad (2)$$

where $\mathrm{sgn}[x(n)] = \begin{cases} 1, x(n) \geq 0, \\ -1, x(n) \langle 0 \end{cases}$

Mean and variance of short-time zero-crossing rate then can be calculated. Figure 2 shows the examples of zero crossing rates calculated over speech, music and speech with music background signals which have been used for calculation of short-time energy.

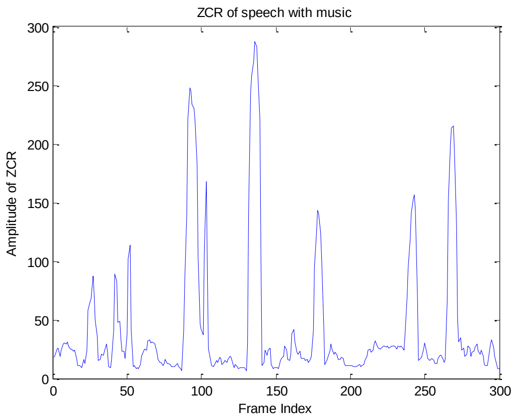
2.1.4. High zero crossing ratio

HZCRR can be defined as the ratio of the number of frames whose ZCR may be around 1.5 fold average ZCR rate, to the total number of frames



(a)

(b)



(c)

Fig 2. Zero crossing rate of 3s (a) speech (b) music and (c) speech with music background

2.1.5. Spectrum flux

Spectrum flux is defined as the average variance value of spectrum between the adjacent two frames in a window.

$$SF = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} \left[\log\left(A(n,k)+\delta\right) - \log\left(A(n-1,k)+\delta\right)\right]^2 \quad (3)$$

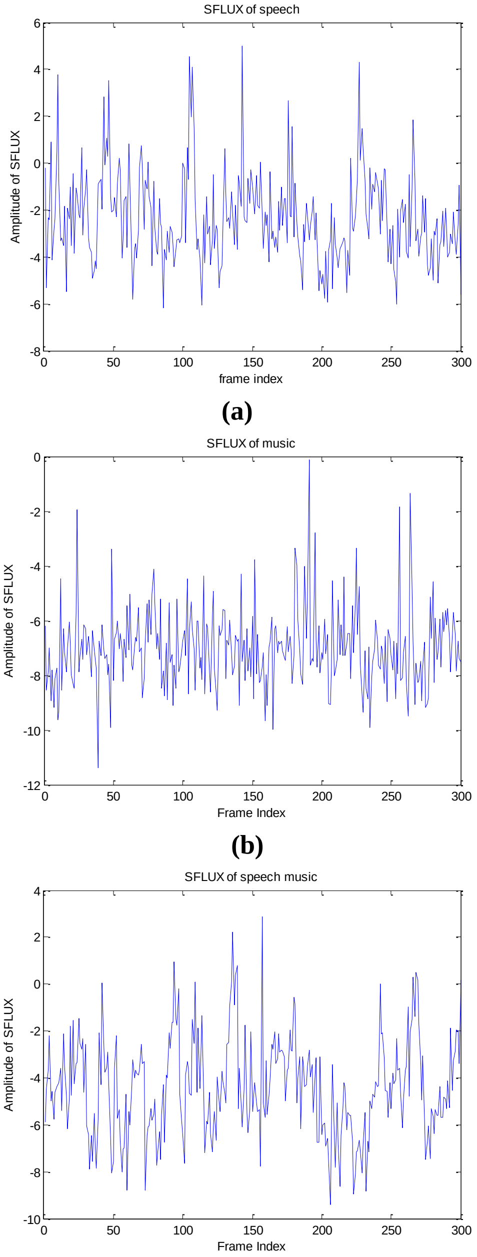
where *A(n,k)* is the discrete Fourier transform of the n^{th} frame of input signal

$$A(n,k) = \left| \sum_{m=-\infty}^{\infty} x(m) w(nL-m) e^{-j(2\pi/L)km} \right| \quad (4)$$

x(m) is the original audio data, *L* is the window length, *K* is the order of DFT, *N* is the total number of frames, δ is the very small value to avoid calculation overflow . Figure 3 demonstrates the examples of spectrum flux computed for previously used three audio signals.

2.1.6. Sample entropy

Sample entropy is a measure of regularity that can be applied to the typically short and noisy time series of clinical data. This method examines times series for similar epochs: more frequent and more similar epochs lead to lower value of sample entropy. SampEn *(m,r,N)* is precisely the negative natural logarithm of the *CP* (conditional probability) that a dataset of length *N*, having repeated itself within a tolerance *r* for *m* points, will also repeat itself for *m* 1 points, without allowing self-matches. The parameter *r* is that it is commonly expressed as a fraction of the *SD* of the data. Figure 4(a) shows an example of sample entropy sequence for a pure speech signal. In Figure 4(b), the plot of sample entropy for music signal and the sample entropy sequence calculated over speech with music signal is shown in Figure 4(c). As expected, the regularity is highest in music signal which has more repeated patterns while speech with music signal has less regularity than pure music and speech has more ups and downs as it contain less repeated patterns.



(a)

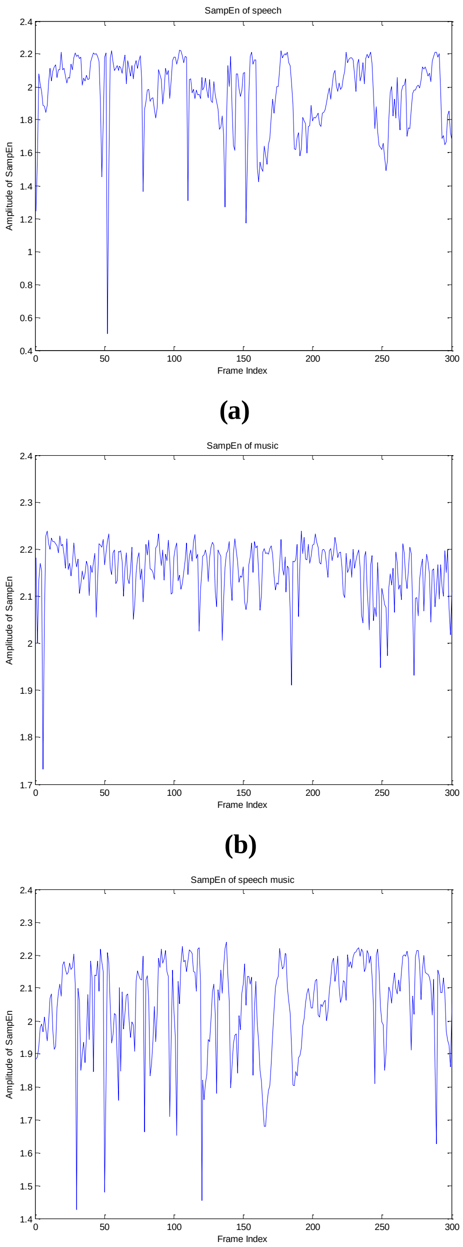
(b)

(c)

Fig 3. Spectrum flux of 3s (a) speech (b) music (c) speech with music background

2.2. Classifiers

This video scene detection system uses two classifiers: support vector machine and Gaussian mixture model in hierarchical approach. The following sub-sections describe brief introduction on these classifier.



(a)

(b)

(c)

Fig 4. Sample Entropy of 3s (a) speech (b) music and (c) speech with music background

2.2.1. Support vector machine

SVM (Support Vector Machine) is a useful technique for data classification. A classification task usually involves with training and testing of some data instances. Each instance in the training set contains one “target value” (class labels) and several “attributes” (features). The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes. Given a training set of instance-label pairs (x_i, y_i), i=1,………,l where $x_i \in R^n$ and $y \in \{1,-1\}^l$, the support vector machines require the solution of the following optimization problem:

$$\min(w,b,\mathrm{E}) \quad \frac{1}{2} w^T w + c\mathrm{E},$$

$$\textbf{Subject to} \quad y_i\left(w^T\varphi(x_i)+b\right) \geq 1-\mathrm{E}_i$$

$$\mathrm{E}_i \geq 0$$

Training vectors x_i are mapped into a higher dimensional space by the function φ. Then SVM finds a linear separating hyper plane with the maximal margin in this higher dimensional space, c>0 is the penalty parameter of the error term. Furthermore, $K(x_i, y_i)$ = $\varphi(x_i)^T\varphi(x_i)$ is called the kernel function. There are some basic kernel functions such as linear, polynomial, radial and sigmoid. In this paper, Radial Basic Kernel function is used for discriminating music from non-music signal. In literature, SVM has been used as an efficient classifier in applications of text classifications [6] and large scale biological applications [7, 8].

2.2.2. Gaussian mixture model

Gaussian mixture model has been successfully applied in numerous applications. Some example applications include language identification [9] and blind separation [10]. In GMM, each Gaussian component is defined as

$$G_k = \frac{1}{(2\pi)^{\frac{1}{2}} \, |V_k|^{\frac{1}{2}}} . e^{[{}^{1}\!/_{2}(X-M_k}{)}^T V_k^{-1}(X-M_k)] \qquad (5)$$

Free parameters of the Gaussian mixture model consist of the means and covariance matrices of the Gaussian components and the weights indicating the contribution of each Gaussian to the approximation of P(X | Cj). These parameters are tuned using a complex iterative procedure called the estimate-maximize (EM) algorithm, that aims at maximizing the likelihood of the training set generated by the estimated PDF. The likelihood function L for each class j can be defined as:

$$L_f = \prod_{i=0}^{N_{train}} P(X_i \,|\, C_j) \qquad (6)$$

3. Proposed method

The main components of video scene analysis using audio information are illustrated in the Figure 5. Firstly, the videos are converted into audio wave files to put into the system as input. Each audio file is segmented into 1s audio clip without overlapping.

Then 9 features are extracted to represent the clip. After extracting the features from audio file, they are put into the SVM and GMM classifiers. Nine features: mean and variance of ZCR, HZCRR, mean and variance of STE, LSTER, SFLUX and mean and variance of SampEn are used to train the classifiers. The different feature vectors are used to train SVM and GMM. To discriminate music from speech signal, 9-mensional feature vector is needed. In classifying speech signal from speech with music background, only two features are utilized from time series regularity: mean and variance of sample entropy. The SVM classifier separates the input data into speech and music by using trained SVMStruct. Speech is further separated into speech and speech with music background by using the GMM classifier. At the final stage of the method, the outputs from two classifiers are integrated into three label results.

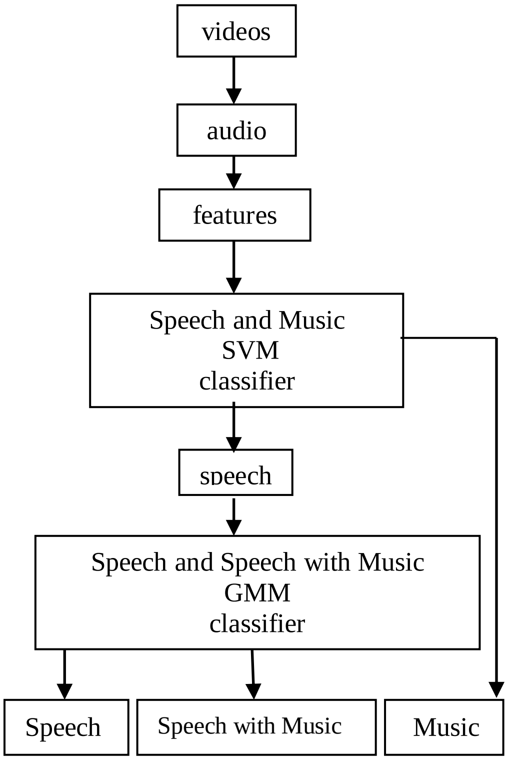

videos
audio
features
Speech and Music SVM classifier
speech
Speech and Speech with Music GMM classifier
Speech
Speech with Music
Music

Fig 5. Proposed System

4. Experimental Study

4.1. Data

In this paper, the videos downloaded from http://www.cctv.com are used as experimental data. All video clips are documentary types. The training data for SVM classifier includes in length of 247s of speech, 888s of music, 49s of mixed audio type (speech with music background) extracted from 3 video episodes and music CDs. The training data for GMM classifier contains 648s of pure speech and 498s of speech with music background. This training data is collected from video clips trained for SVM classifier and speech from various sources. There is no overlapping between testing data and training of this experimental study.

In the testing phase, for every 1s clip, it is fed into SVM classifier to determine it is a music clip or speech clip. Clips identified as music types are then eliminated from the test sequence. The clips which are labeled as speech are again fed into GMM classifier. The task of GMM classifier is to discriminate speech and speech with music background clips. The proposed algorithm is also tested with videos from CCTV channel. Three video episodes from series "Tibet Serf" are used as test data. Here, these episodes are named as Tibet Serf 1, Tibet Serf 2 and Tibet Serf 3. Labels for each 1s clip of the tested videos are given manually. Tibet Serf 1 video includes 72s of speech, 54s of music and 124 of speech with music background. In Tibet Serf 2, 6s of music and 73s of speech with music are contained. Tibet Serf 3 video is composed of 40s of music and 86s of speech with music. Environmental sounds which are not included in training stage are labeled as music clips.

4.2. Performance

The performance of the proposed approach is measured with classification accuracy measured for each sound class. The ratio of correctly and incorrectly classified 1s clips over total 1s clips contained in each class is defined as true class and false class as given Table 1.

Table 1: Performance criteria information

Sound Type	**Speech**	**Music**	**Speech with Music**
Speech	Trspeech	Spasmusic	Spasspmusic
Music	Muasspeech	Trmusic	Muasspmusic
Speech With Music	Spmuaspeech	Spmuasmusic	Trspmusic

Table 2 to Table 4 show the detected scene results for the tested 3 videos described above. .

Table 2: Detection Result for Tibet Serf 1

Scene	Speech	Music	Speech w/ Music
Speech	37.5%	13.89%	48.61%
Music	10.71%	75%	14.29%
Speech w/ Music	28.23%	17.74%	54.03%

Table 3: Detection Result for Tibet Serf 2

Scene	Speech	Music	Speech w/ Music
Speech	-	-	-
Music	16.67%	50%	33.33%
Speech w/ Music	24.66%	13.70%	61.64%

Table 4: Detection Result for Tibet Serf 3

Scene	Speech	Music	Speech w/ Music
Speech	-	-	-
Music	20%	65%	15%
Speech w/ Music	36%	8.14%	55.81%

From these tables, it is observed that the method obtains the highest accuracy for music class. The relatively low accuracy is found for speech and speech with music background. This is due to insufficiency of training data with extracted feature dimension and labeling of environmental sound into music class. The proposed method detects speech class as speech with music class and speech with music scene as speech scene since these two classes are very close to each other. However, the false detection rate for pure speech as pure music and pure music as pure speech is found to be as low as 10.71%. Better classification performance can be achieved by testing with videos containing pure audio clips and audio with less noisy audio background.

5. Conclusion

Video scene detection such as speech, music and speech with music are performed on the audio signal nature in this paper. Two types of classifiers, SVM and GMM are hierarchically used. Different feature sets are fed into SVM and GMM. Through initial investigation of this study, it is found that the SVM is not suitable for classification of complex audio types. GMM is observed to be more robust for complex audio types such as speech with music signal. The proposed approach is tested with documentary video sequences containing three main audio classes (speech, music, speech with music background) with different durations. The results show that this method offers acceptable detection accuracy. Specifically, the method is favorable to video consisting of pure music and speech with music background scenes.

.

6. References

[1] D.Kimber and L.Wilcox, "Acoustic Segmentation for Audio Browsers," *Sydney*, Australia, Jul. 1996, pp. 83-87

[2] J.Saunders, "Real-time Discrimination of Broadcast Speech/Music," in *Proc. Int. Conf. Acoust., Speech, Signal Process (ICASSP)*, Atlanta, GA, May1996, vol. II, pp.993-996

[3] S.Srinivasan, D. Petkovic, and D.Ponceleon, "Toward Robust Features for Classifying Audio in the Cue Video system," in *Proc. ACM Multimedia*, 1999, pp.393-400

[4] T. Zhang and C.-C, Kuo, "Audio Content Analysis for Online Audio-Visual Data Segmentation," *in IEEE Trans. Speech Audio Process*, vol.9,no.4, ,Jul.2001, pp.441-457.

[5] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang, "Comparing MFCC and MPEG-7 Audio Features for Feature Extraction, Maximum Likelihood HM and Entropic Prior HMM for Sports Audio Classification," in *Proc. ICME*, 2003, pp.397-400

[6] S. Tong, D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification" in *Journal of Machine Learning Research* ,2001, pp.45-66

[7] C. Xue, F. Li, T. He, G. P. Liu, Y. d. Li, X.g. .Zhang,"Classification of Real and Pseudo MicroRNA Precursors Using Features and Support Vector Machine", in *BMC Bioinformatics*, 2005, pp.6:310.

[8] H. Zhu, F. S. Domingues, I. Sommer, T. Lengauer, "NOXclass: Prediction of Protein-Proteinlocal Structure-Sequence Interaction Types", in *BMC Bioinformatics*, 2006, pp.7:27

[9] P. A. Torres-Carrasquillo, Elliot Singer, Mary A Kohler, Richard J. Greene, Douglas A. Reynolds, and J.R.Deller, Jr., "Approaches to Language to Identification Using Gaussian Mixture Models and Shifted Data Cepstral features", in *Proc. Int'l. Cof. Spoken Language Processing, Denver*, Sep.2002(in press), pp.89-92

[10] K. Todros and J. Jaboikian " Applications of Gaussian Mixture Models for Blind Separation of Independent Sources", in *Lecture Notes in Computer Science*, 2004, pp.382-389