National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2012)
Proceedings published by International Journal of Computer Applications® (IJCA)

Audio Segmentation using Line Spectral Pairs

N. P. Jawarkar
B. N. College of Engineering, PUSAD, Dist-Yavatmal (MS) India

R. S. Holambe
SGGS Institute of Engg. and Technology, Vishnupuri, Nanded (MS), India

T. K. Basu
Institute of Technology and Marine Engineering, Jhinga (WB), India

ABSTRACT
This paper describes a technique for unsupervised audio segmentation. Main objective of the work presented in this paper is to study the performance of audio segmentation system using metric-based method. The system first classifies the audio signal into speech and nonspeech signal using variance of zero crossing rate. The feature Line spectral pair is used for automatically detecting the speaker change point. Hotelling T^2 distance metric is used in the first stage for coarse speaker change detection. The Bayesian information criterion (BIC) is used in the second stage to validate the potential speaker change point detected by the coarse segmentation procedure to reduce the false alarm rate. Database of four files containing the speech recorded from different combinations of male and female speakers mixed with nonspeech signal such as music and environmental sound are used for segmentation. The database-file with one male and one female gives the best performance with F_1 measure of 0.9474.

Keywords
Speaker segmentation, LSP, audio segmentation, VZCR

1. INTRODUCTION
There is a rapid increase in the volume of recorded speech. This includes television and audio broadcasts, voice mails, meetings, telephonic conversations and spoken documents resources. There is a growing need to apply automatic human language technologies to allow efficient and effective searching, indexing, and accessing of these information sources [1]. Speaker diarization is the process of automatically partitioning a multi-speaker conversation into the homogenous segments and grouping together all the segments that corresponds to the same speaker. The first part of the process is known as speaker segmentation or speaker change detection while the second one is called as the speaker clustering. Overview of automatic speaker segmentation and clustering is given in [2], [3].

Speaker segmentation is defined as the process by which a long speech signal is partitioned into homogenous segments by detecting changes of speaker identity. Speaker segmentation algorithms can be broadly classified into three categories: model based, metric based and hybrid. In the model-based segmentation, a set of models is derived and trained for different speaker classes from a training corpus. The incoming speech streams are classified using these models [4]. However, in many cases, the pre-knowledge of speakers and acoustic classes are often not available. Some of the model based classification methods are Gaussian Mixture Model (GMM), GMM with multilayer perceptron (MLP), K-nearest neighbor(KNN), Support Vector Machines(SVM), etc. [5]. Metric-based method accesses the similarity between neighbouring analysis windows over the audio stream by a distance function of their metric. A wide variety of distance metrics could be used. A commonly used metrics are Bayesian Information Criterion (BIC)[6],[7], the Kullback-Leibler divergence (Gaussian divergence) [8],[9] and second order hotelling T^2 statistics[7], [10], [11]. Hybrid method uses both model-based and metric-based approach.

Various features have been used for classifying audio signals into speech and nonspeech (music, environmental sounds, etc.) signals. Zhang and Kuo have used average zero crossing rate (ZCR), fundamental frequency and spectral peak tracks as their features [12]. Lu et al. [13] employed noise frame ratio, low short time energy ratio and four other features. Li [14] considered the total spectral power, sub-band power and also other features. R. Huang et al. [5] have used variance of the spectral flux (VSF) and variance of the zero crossing rate (VZCR) for speech and non-speech classification.

Speaker change can be detected using different features. Lu and Zhang [15] have used a multifeature set consisting of Mel frequency cepstral coefficients (MFCC), Line Spectrum and pitch features to detect change points. Perceptual linear prediction (PLP) cepstral coefficients are used in [16]. R. Huang et al. [5] have considered perceptual based minimum variance distortionless response (PMVDR), smooth zero crossing rate(SZCR), filterbank log energy coefficients (FBLC). Adami et al. have used Line spectral pairs (LSP) features [17]. T. K.Truong et al. [18] have used wavelet based features such as subband power, pitch frequency and other three features.

Main objective of the work presented in this paper is to study the performance of audio segmentation system using metric-based method. The VZCR feature is used for speech and nonspeech classification. The LSP feature is used for speaker change detection. Rest of the paper is organized as follows. System overview is discussed in section 2. Features such as VZCR and LSP, and distance measures used for potential speaker change detection used in the system are also discussed in this section. Section 3 gives the experimental work and performance analysis. Finally conclusion is given in section 4.

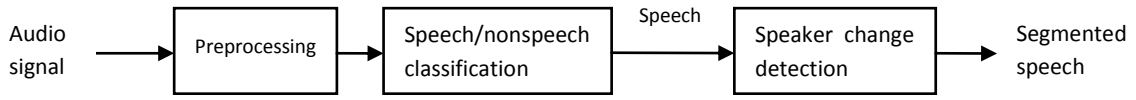


Fig. 1: Audio segmentation block diagram

2. SYSTEM OVERVIEW

Description of the system used for the audio segmentation along with the features used for voiced/unvoiced classification and speaker change detection is given in this section. The system block diagram is shown in Fig.1. It mainly consists of preprocessing, speech/nonspeech classification and speaker change detection.

2.1 Preprocessing

The audio signal is first passed through anti-aliasing filter. The signal is then sampled at the sampling frequency of 22050 Hz and converted into digital signal using digital to analog converter with 16-bit resolution. The silence removal stage removes the silence portion of the signal based on the energy threshold criterion. After silence removal, the voiced speech signal is pre-emphasized with the pre-emphasis factor of 0.97. This is followed by frame blocking with a frame length of 512 samples (23.22 ms) with 50% overlap with the neighbouring frames.

2.2 Speech/ Nonspeech classification

The audio signal, in general, may consist of speech signal and the nonspeech signal such as music, environmental sounds, etc. The variance of spectral flux of speech and variance of the zero crossing rate have proved to be better features for speech/nonspeech classification [5]. In our preliminary study, VZCR is found to better discriminator than VSF. Hence we have used VZCR for speech/nonspeech classification. VZCR is calculated on the basis of zero crossing rate (ZCR). ZCR is the number of zero-crossings within the frame and is calculated as under[19].

$$z(m) = \frac{1}{N}\sum_{n=m-N+1}^{m} \frac{|\mathrm{sgn}(s(n)) - \mathrm{sgn}(s(n-1))|}{2} w(m-n) \quad (1)$$

where N is the length of the frame, m is the endpoint of the frame, and $w(n)$ is the window function. With 23.22 ms-frame and 50% overlapping, there are 86 frames per second. The VZCR is calculated over 17 frames. Therefore, in one second of audio there will be about five subblocks, each resulting in a speech/nonspeech decision task as follows:

$$c_{ij} = \begin{cases} 0, & if \quad V_{ij} \geq \theta \\ 1, & otherwise \end{cases} \quad i=1, 2,.. , j=[1,5] \quad (2)$$

where 1 means speech, 0 means nonspeech, V_{ij} is the VZCR value of j^{th} subblock in the i^{th} audio block of 1-s duration and θ is the threshold. The final decision on the 1-s audio block is based on the vote

$$c_i = \begin{cases} 1, & if \sum_{j=1}^{5} c_{ij} \geq 3 \\ 0, & else \end{cases} \quad i=1, 2, . . \quad (3)$$

2.3 Speaker change detection

The aim of this step is to find the change over point in the speech signal between two speakers. Line spectrum pair (LSP) is used as feature for speaker change detection in our study. LSP was first introduced by Itakura [20] as an alternative to linear predictive coding (LPC) spectral representation. It has some important properties, such as all zeros of LSP polynomials lie on the unit circle and the corresponding zeros of the symmetric and anti-symmetric LSP polynomials are interlaced. Twelfth order LSP is computed for each frame having 512 samples. Speech signal is divided into overlapped segments, each of size 400 frames, where each segment is displaced by 20 frames with respect to its preceding segment as shown in Fig. 2. Hotelling T^2 distance metric is used in the first stage for coarse speaker change detection. T^2 distance is computed for two audio segments as under.

$$T^2 = \frac{a}{a+b}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) \quad (4)$$

where a and b are the number of frames of within each of the audio sub-segments, respectively and $\sum$ is the covariance of the segment of size (a+b) frames. Two audio segments are represented by multivariate Gaussian distributions: $N(\mu_1,\sum_1)$ and $N(\mu_2,\sum_2)$, where μ_i and $\sum_i$ represents the mean and covariance of the i^{th} segment, respectively.

If two distributions are corresponding to two different persons, there is a local peak in the distance measure at that point. To detect the potential change over point from the local peaks in the dissimilarity sequence, they must satisfy the following conditions.

(i) $T^2(i) > T^2(i-1)$
(ii) $T^2(i) > T^2(i+1)$
(iii) $T^2(i) >$ Threshold

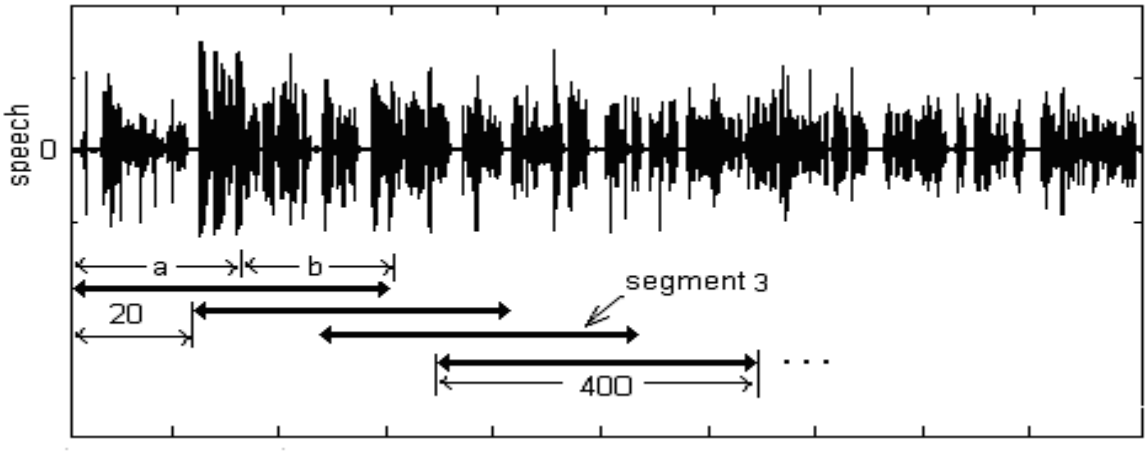


Fig 2: Illustration of overlapping of segments for distance computation

Preliminary study helped us to decide threshold equal to 1.5 times the mean of T^2 distance.

The Bayesian information criterion (BIC) is used in the second stage to validate the potential speaker change point detected by the coarse segmentation procedure to reduce the false alarm rate. The BIC difference between two models is given by:

$$\Delta BIC = \frac{1}{2}((N_i + N_j)\log|\Sigma| - N_i \log|\Sigma_i| - N_j \log|\Sigma_j| - \frac{1}{2}(\delta + \frac{1}{2}\delta(\delta+1))\log(N_i + N_j) \quad (5)$$

where γ is the penalty factor(set to 0.6), δ is the feature vector dimension, N_i is the number of frames in the i^{th} sub-segment, $\sum_i$ is the covariance matrix of the i^{th} sub-segment of length N_i, and $\sum$ is the covariance matrix of the segment of length (N_i+N_j).

3. EXPERIMENTAL WORK

Database having four audio files, namely 1M-1F, 2M-2F, 3M-3F and 4M-4F (where numeral indicates number of persons and; M and F indicate male & female, respectively) containing speech and nonspeech signal were generated. Each database file contains concatenated speech recording of male and female speakers in Marathi language, and nonspeech data such as music and environmental sound. The audio signal was first classified into the speech and nonspeech signal by using VZCR. To demonstrate the speech/nonspeech classification, a set of five speech clips from two speakers and four music/environmental sound clips each of 10-s duration were selected and concatenated to form a 90-s duration audio stream. The audio data were divided into overlapping frames each of 23.22 ms duration. ZCR was calculated for each frame and VZCR was calculated over 17 frames. The speech/nonspeech classification is carried out as mentioned in section 2. Fig. 3 shows the results of speech/nonspeech classification. It is found that, out of 90-s of audio data only1.11-s of nonspeech data are misclassified as speech data.

Following figures of merit were used for judging the performance of the system [21].

$$PRC = \frac{CFC}{DET} = \frac{CFC}{CFC + FA} \quad (6)$$

where PRC indicates the precision, CFC denotes the number of correctly found changes and DET=CFC+FA is the number of detected speaker changes.

$$RCL = \frac{CFC}{CFC + MD} \quad (7)$$

where RCL indicates the recall and MD denotes number miss detections.

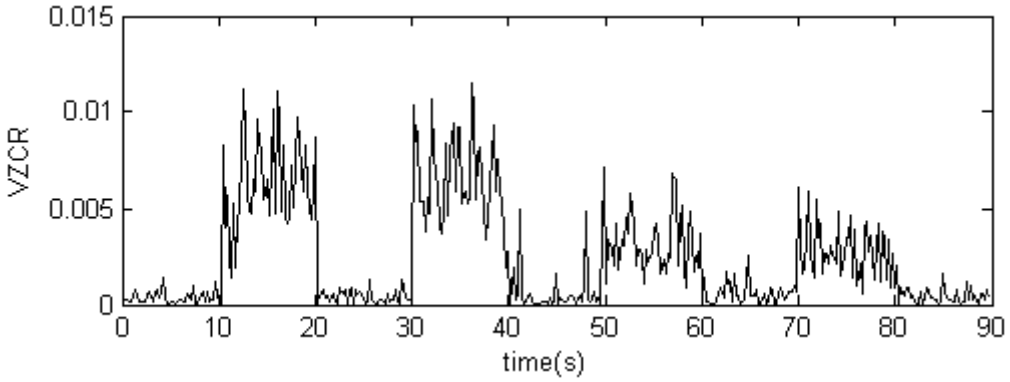
$$F_1 = 2\frac{PRC.RCL}{PRC + RCL} \quad (8)$$

F_1 measure takes the value between 0 and 1 and indicates the performance. For better performance the value of F1 should be nearly unity.

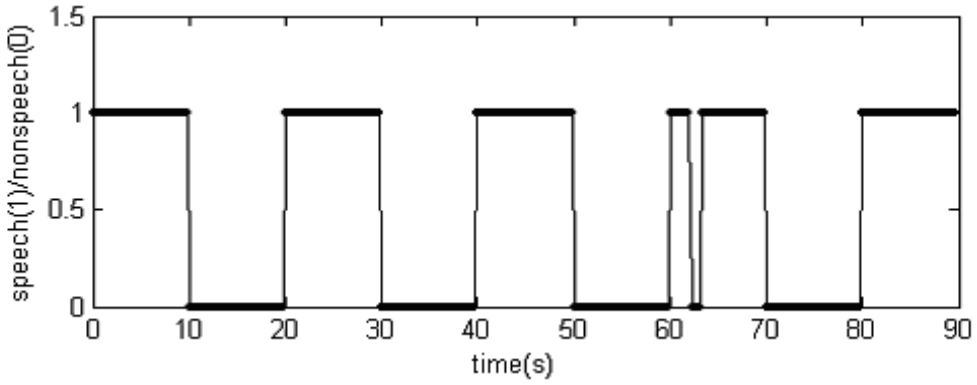
Table 1. Performance comparison for different database

Database file	CFC	FA	MD	PRC	RCL	F1
1M-1F	9	1	1	0.900	1.000	0.9474
2M-2F	21	1	1	0.954	0.913	0.9333
3M-3F	26	2	2	0.929	0.963	0.9455
4M-4F	22	3	3	0.880	0.956	0.9167

Four database files were tested for speaker change detection using LSP feature; and T^2 and ΔBIC distance metrics. The



(a) VZCR versus time



(b) Speech/nonspeech labeling versus time(s)

Fig 3: Speech/nonspeech classification using VZCR. (a) VZCR versus time (s) and (b)Speech/nonspeech labeling versus time(s)

results are shown in Table-I. It can be seen that the database file 1M-1F gives the best performance. Further as the number of speakers increases the performance degrades except for database file 3M-3F.

4. CONCLUSION

The audio segmentation system using the SZCR and LSP features has been developed and tested. It has been found that VZCR is better feature for speech/nonspeech classification. The performance for the database with one male and one female is found to the best amongst the four database files. The feature LSP is computationally simple and can be used in real time speaker segmentation.

5. REFERENCES

[1] A. Solomonoff, A. Mielke, M. Schmidt, H. Gish, "Clustering speakers by their voices," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Seattle, USA, pp. 757–760 , May 1998.

[2] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," IEEE Trans. Audio, Speech and language process., vol. 14, no. 5, pp. 1557-1565,Sept. 2006.

[3] M. Kotti,V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," Signal Processing, 88(2008), pp. 1091-1124.

[4] L. Lu, H. Jiang, and H. J. Zhang, "A robust audio classification and segmentation method," in Proc. 9th ACM Int. Conf. Multimedia, 2001, pp. 203–211.

[5] R Huang, J. H. L. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," IEEE Trans. Audio, speech, Language Process., vol. 1, no. 3, pp. 07919, May 2006.

[6] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in Proc. Broadcast News Transcr. Under. Workshop, Lansdowne, VA, 1998, pp. 127–132.

[7] B. Zhou and J. H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian information criterion," in Proc. ICSLP 2000, vol. 1, Beijing, China, Oct. 2000, pp. 714–717.

[8] M. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in Proc. DARPA Speech Recognition Workshop, Chantilly, VA, 1997, pp. 97–99.

[9] C. Barras, X. Zhu, S. Meigner, J. L. Gauvain, "Multistage Speaker diarization of broadcast news," IEEE Trans. Audio Speech Language Process. , vol. 14, no. 5, pp.1557-1565, Sept. 2006.

[10] M. Cettolo and M. Federico, "Model selection criteria for acoustic segmentation," in Proc. ISCA ITRWASR 2000 Workshop, Paris, France, Sep. 2000, pp. 221–227.

[11] S. Wegmann, P. Zhan, and L. Gillick, "Progress in broadcast news transcription at dragon systems," Proc. ICASSP, vol. 1, pp. 33–36, Mar. 1999.

[12] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," IEEE Trans. Speech Audio Process., vol. 9, no. 4, pp. 441–457, Jul. 2001.

[13] L. Lu, H. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," IEEE Trans. Speech Audio Process., vol. 10, no. 7, pp. 504–516, Oct. 2002.

[14] S. Z. Li, "Content-based audio classification and retrieval using the nearest feature line method," IEEE Trans. Speech Audio Process., vol. 8, no. 5, pp. 619-625,Sept. 2000.

[15]] L. Lu and H. Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis," in Proc. ACM Multimedia, Juan-les-Pins, France, Dec. 2002, pp. 602–610.

[16] S.E. Tranter, K. Yu, G. Evermann, P.C. Woodland, "Generating and valuating segmentations for automatic speech recognition of conversational telephone speech," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, pp. 433–477, May 2004.

[17] A. Adami, S. Kajarekar, and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," in Proc. ICASSP, vol. 4, Orlando, FL, 2002, pp. 13–17.

[18] T. K. Truong,C. Lin, S. Chen, " Segmentation of specific speech signals from multi-dialog environment using SVM and wavelet," Pattern Recognition Letters, 28, pp. 1307-1313, 2007.

[19] L. R. Rabiner and R. W. Schafer, Digital signal processing of speech signals, Englewood, NJ: Prentice-Hall,1978.

[20] Itakura F., "Line spectrum representation of linear predictive coefficients of speech signals," J. Acoust. Soc Am., 57, 537(A), 1975.

[21] J. Ajmera, I. McCowan, H. Bourlard, "Robust speaker change detection," IEEE Signal Process. Lett. 11 (8), pp. 649–651, August 2004.