Journal of Computer Science and Applications.
ISSN 2231-1270 Volume 9, Number 1 (2017), pp. 33-37

http://www.irphouse.com

Speech/Music Change Point Detection using Perceptual Linear Prediction and GMM

R. Thiruvengatanadhan

Department of Computer Science and Engineering,
Annamalai University, Annamalainagar, Tamil Nadu, India.

Abstract

Category change point detection of acoustic signals into significant regions is an important part of many applications. Changes in audio signal characteristics help in detecting the category change point between different categories. In this paper, Perceptual Linear Prediction (PLP) features are extracted which are used to characterize the audio data. Gaussian mixture model (GMM) is used to detect change point of audio. The results achieved in our experiments illustrate the potential of this method in detecting the change point between speech and music changes in audio signals.

Keywords: Speech, Music, Feature Extraction, PLP, GMM.

I. INTRODUCTION

Category change point detection of acoustic signals into significant regions is an important part of many applications. Systems which are developed for speech/music classification, indexing and retrieval usually take segmented audios rather than raw audio data as input. A first content characterization could be the categorization of an audio signal as one of speech, music, or silence [1]. Changes in audio signal characteristics help in detecting the category change point between different categories. In speech/music change point detection the audio signal can be segmented into speech and music regions. A human listener can easily distinguish audio signals into these different audio types by just listening to a short segment of an audio signal. However, solving this problem using computers has proven to be very difficult [2].

Category change points in an audio signal such as speech to music, music to advertisement and advertisement to news are some examples of segmentation boundaries. Systems which are designed for classification of audio signals into their corresponding categories usually take segmented audios as input. However, this task

in practice is a little more complicated as these transitions are not so obvious all the times [3]. For example, the environmental sounds may vary while a news report is broadcast. Thus, many times it is not obvious even to a human listener, whether a category change point should occur or not.

II. ACOUSTIC FEATURE EXTRACTION

Acoustic feature extraction plays an important role in constructing an audio change point detection system. The aim is to select features which have large between-class and small within-class discriminative power. Discriminative power of features or feature sets tells how well they can discriminate different classes. Feature selection is usually done by examining the discriminating capability of the features.

A. Perceptual Linear Prediction (PLP)

Hermansky developed a model known as PLP. It is based on the concept of psychophysics theory and discards unwanted information from the human pitch [4]. It resembles the procedure to extract LPC parameters except that the spectral characteristics of the speech signal are transformed to match the human auditory system.

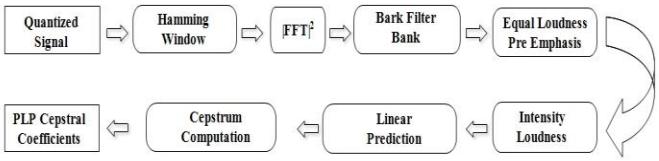

Fig. 1 PLP Parameter Computations.

PLP is the approximation of three aspects related to perceptron namely resolution curves of the critical band, curve for equal loudness and the power law relation of intensity loudness. The process of PLP computation is shown in Fig. 1. The audio signal is hamming windowed to reduce discontinuities. The Fast Fourier Transform (FFT) transforms the windowed speech segment into the frequency domain [5].

The auditory warped spectrum is convolved with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing. Critical band is the frequency bandwidth created by the cochlea, which acts as an auditory filter. The cochlea is the hearing sense organ in the inner ear. Bark scale corresponds to 1 to 24 critical bands. The power spectrum of the critical band masking curve and auditory warped spectrum are convoluted to simulate the human hearing resolution. The equal loudness pre-emphasis needs to compensate the unequal perception of loudness at varying frequencies. An all pole model normally applied in Linear Prediction (LP) analysis is used to approximate the spectral samples. Either the coefficients can be used as such for representing the signal or they can further be

transformed to Cepstral coefficients. In this work, a 9^{th} order LP analysis is used to approximate the spectral samples and hence obtained a 9-dimensional feature vector for a speech signal of frame size of 20 milliseconds is obtained.

III. TECHNIQUES

A. Gaussian mixture models (GMM)

The probability distribution of feature vectors is modeled by parametric or non parametric methods. Models which assume the shape of probability density function are termed parametric. In non parametric modeling, minimal or no assumptions are made regarding the probability distribution of feature vectors [6]. In this section, we briefly review Gaussian mixture model (GMM), for audio classification. The basis for using GMM is that the distribution of feature vectors extracted from a class can be modeled by a mixture of Gaussian densities.

The iterative Expectation Maximization (EM) algorithm is used to estimate the parameters of GMM. EM algorithm is one of the most popular clustering algorithms used to estimate the probabilistic models for each Gaussian component. The Expectation step (E-step) and Maximization step (M-step) are iterated till the convergence of the parameter [7]. EM algorithm finds out maximum likelihood estimation of parameters. The E-step computes Expectation of likelihood assuming parameters and M-step computes maximum likelihood estimates of parameters by maximizing the expected likelihood found in E-step.

IV. EXPERIMENTAL RESULTS

A. The database

Performance of the proposed audio change point detection system is evaluated using the Television broadcast audio data collected from Tamil channels, comprising different durations of audio namely speech and music from 5 seconds to 1 hour. The audio consists of varying durations of the categories, i.e. music followed by speech and speech in between music etc., Audio is sampled at 8 kHz and encoded by 16-bit.

B. Acoustic feature extraction

9 PLP features are extracted a frame size of 20 ms and a frame shift of 10ms of 100 frames as window are used. Hence, an audio signal of 1 second duration results in 100 × 9 feature vector. GMM models are used to capture the distribution of the acoustic feature vectors.

C. Category change point detection

The sliding window of 1 second is initially placed at the left end of the signal. The confidence score for the middle frame of the window is computed by averaging the

scores of the frames in the left half of the window. The window is shifted by 10 ms and the same procedure is repeated for the entire signal. The performance of the proposed speech/music change point detection system is shown in Fig. 2 for GMM.

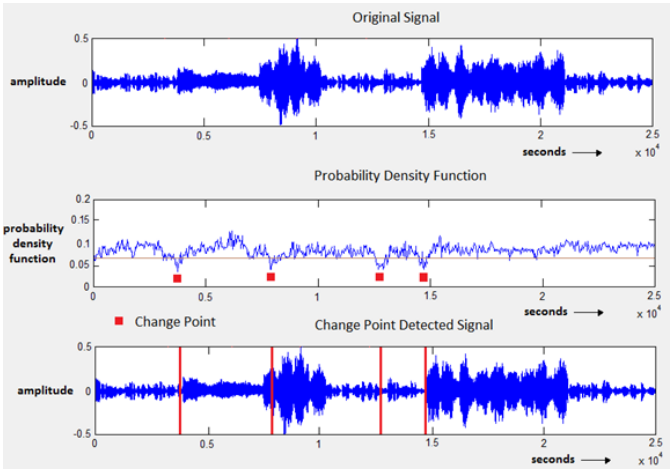

Fig. 2 Snapshot of Speech/Music Change Point Detection Systems Using GMM.

The performance of the speech/music change point detection system using GMM to detect the change point in terms of the various measures is shown in Fig. 3.

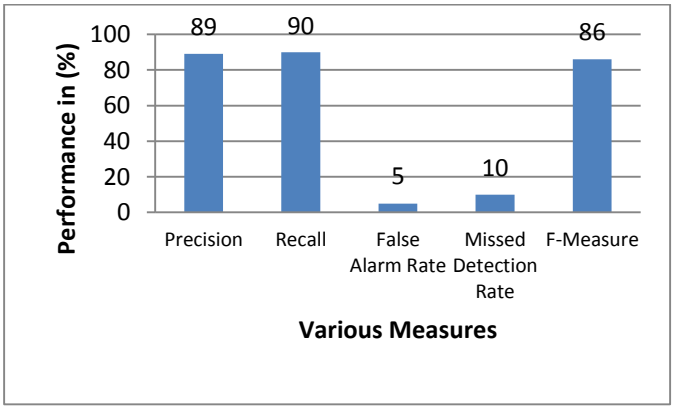

Fig. 3: Performance of detect the change point in terms of the various measures using GMM.

V. CONCLUSIONS

In this paper we have proposed a method for detecting the category change point between speech/music using Gaussian Mixture Model (GMM). The performance is studied using 9 dimensional PLP features. GMM based change point detection gives a better performance of 86% F-measure is achieved.

REFERENCES

[1] D. Li, I. K. Sethi, N. Dimitrova, and T. Mc Gee, "Classification of General Audio Data for Content Based Retrieval," Pattern Recognition Letters, vol. 22, no. 1, pp. 533-544, 2001.

[2] G. M. Bhandari, R. S. Kawitkar, M. C. Borawake, "Audio Segmentation for Speech Recognition using Segment Features," International Journal of Computer Technology and Applications, vol. 4, no. 2, pp. 182-186, 2013.

[3] Francis F. Li, "Nonexclusive Audio Segmentation and Indexing as a Pre-processor for Audio Information Mining," 26th International Congress on Image and Signal Processing, IEEE, pp: 1593-1597, 2013.

[4] Peter M. Grosche, Signal Processing Methods for Beat Tracking, Music Segmentation and Audio Retrieval, Thesis, Universit¨at des Saarlandes, 2012.

[5] PetrMotlcek, Modeling of Spectra and Temporal Trajectories in Speech Processing, PhD thesis, Brno University of Technology, 2003.

[6] Tang, H., S.M. Chu, M. Hasegawa-Johnson, T.S. Huang, Partially Supervised Speaker Clustering. IEEE Transactionson Pattern Analysis and Machine Intelligence, 34(5): 959–971. 2012.

[7] Chien-Lin Huang, Chiori Hori and Hideki Kashioka, "Semantic Inference Based on Neural Probabilistic Language Modeling for Speech Indexing," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8480-8484, 2013.