ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Speech/music change point detection using PNCC and GMM

R. Thiruvengatanadhan

Assistant Professor, Dept. of Computer Science and Engineering, Annamalai University, Annamalainagar,

Tamilnadu, India

ABSTRACT: Speech/music segmentation is the task of partitioning an audio stream into speech and music segments. This is the first fundamental step toward content based audio information retrieval because most of the signals processing algorithms are separately designed for these classes. In this paper, Power Normalized Cepstral Coefficients (PNCC) 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, PNCC, GMM.

I.INTRODUCTION

Automatic audio segmentation divides a digital audio signal into segments, each of which contains audio information from a specific acoustic type such as speech, music, non-verbal human activity sounds, animal vocalizations, environmental sounds, noises, etc. The ultimate goal of segmentation is to produce a sequence of discrete utterances with particular characteristics remaining constant within each one. The characteristics of choice depend on the overall structure of the recognition system [1]. The segmentation algorithm was directed at locating points in the audio stream where there was a change in the acoustic class. The degree of detail in audio class analysis depends on the application. For example, in radio broadcast signals segmentation, the interest falls in the detection of the audio parts that contain speech, music, silence and noises.

The segmentation that is done automatically implements it through the division of a digitalized audio signal into smaller segments, which contain information in the form of audio using a particular type of acoustic feature [2]. Segmentation is used to produce sequential characteristic possessing utterances in the discrete form with constant characteristics [3].

Model-based change point detection involves defining the set of models from different speaker classes and training them before the segmentation process begins. Segmentation can detect the sites where acoustic feature changes in these boundaries where changes take place are regarded as the segment boundaries. It looks for the speaker that it sequenced in such a way that the time alignment with respect to time is the best and it accomplishes the segmentation process, where there are chances of finding the acoustic changes [4].

In metric-based change point detection the audio signal is fragmented into smaller chunks that are known to contain only one type of segment [5]. Audio change point detection is used to measure dissimilar values between acoustic feature vectors in two consecutive windows. Consecutive distance values are often filtered using the low pass filters. Hybrid change point detection combines both approaches, namely metric based and model based. This algorithm is used for segmentation that lays its base on distance that is meaningful in the production of speech and music model initial sets.

In decoder-guided change point detection, the decoding process is done on the audio data that are fed to the system, after doing that, the useful segments are generated by segmenting the input at the silence locations that the decoder generates. There has been the traditional implementation of the way involving the segments used for recognizing the speech [6]. There are two systems that fall in this category. One of them is based on energy and the other one is based on decoder.

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.

Power Normalised Cepstral Coefficients (PNCC) is well known for the high accuracy of automatic speech recognition systems even in high-noise environments [7]. PNCC is an acoustic feature which performs the computation using online algorithms in real-time and provides high accuracy even in noisy conditions [8]. (PNCC) is well known for the accuracy of automatic speech recognition systems, even in high-noise environments. In Fig. 1 Shows the block diagram for the extraction of PNCC features.

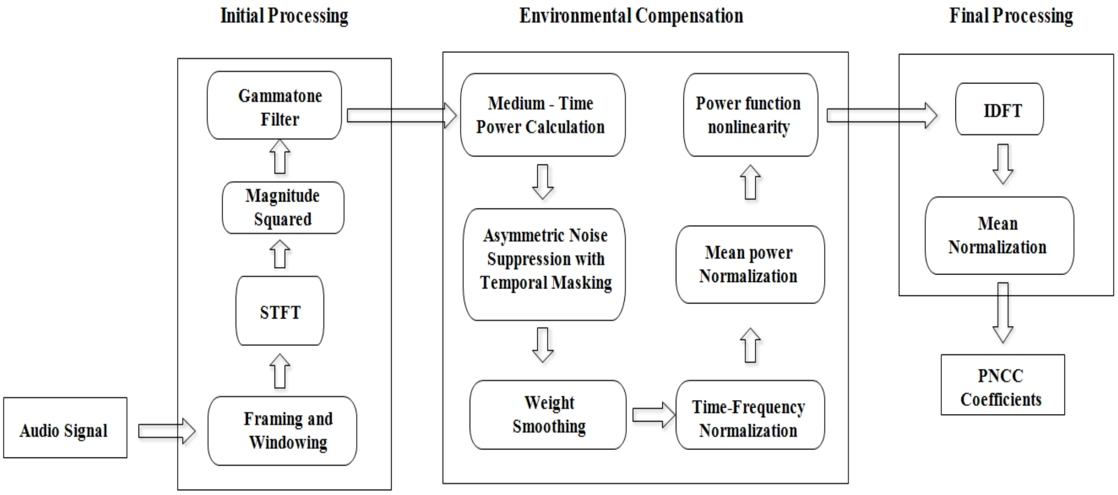

Fig. 1 PNCC Feature Extractions.

III.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 [9]. 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 [10]. 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

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.

Acoustic feature extraction

13 set of PNCC feature is extracted from each frame of the audio by using the feature extraction techniques. PNCC feature will be calculated for the given wav file. The above process is continued for 600 wav files. The feature values for all the wav files will be stored separately for speech and music.

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.

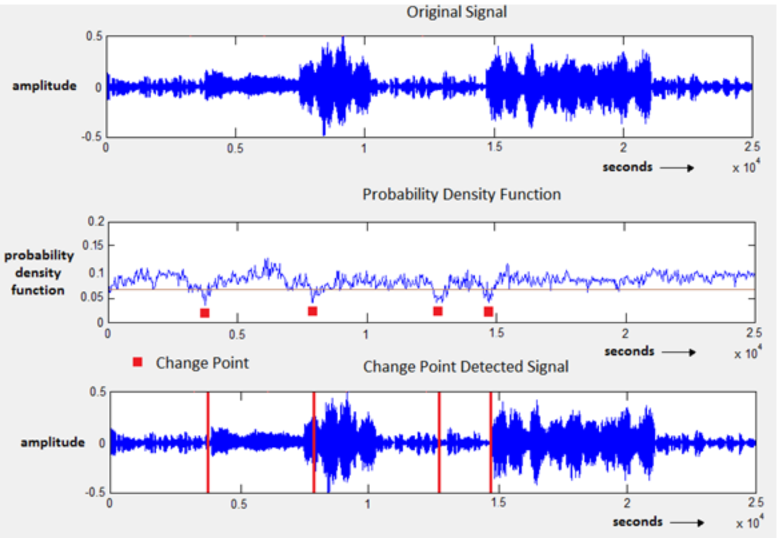

Fig. 2 Snapshot of Speech/Music Change Point Detection Systems Using GMM.

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.

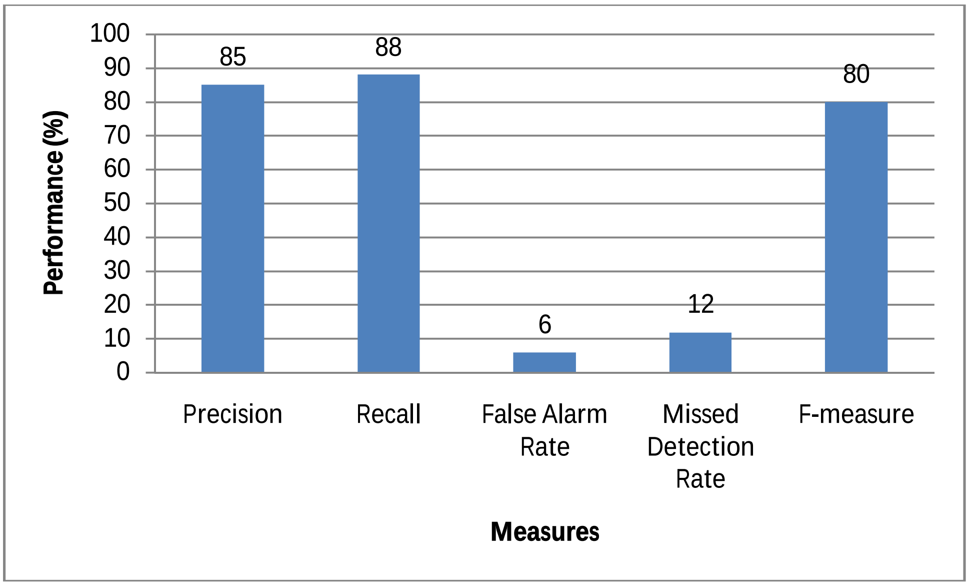

Fig. 3: Performance of detect the change point in terms of the various measures using GMM.

V.CONCLUSION

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 13 dimensional PNCC features. GMM based change point detection gives a better performance of 80% F-measure is achieved.

REFERENCES

[1] Kim H.-G. and Sikora T., “Automatic Segmentation of Speakers in Broadcast Audio Material,” *IS&T/SPIE’s Electronic Imaging 2004*, San Jose, CA, USA, January 2004.

[2] Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., and Wactlar, H., “Informedia, Digital Video Library,” *Communications of the ACM*, vol. 4, no. 38, pp. 57-58, 1995.

[3] Yali Amit, Alexey Koloydenko, and ParthaNiyogi, “Robust Acoustic Object Detection,” *Journal of the American Acoustic Association*, vol. 118, pp. 2634-2648, 2005.

[4] Vincenzo Dimattia, *An Automatic Audio Segmentation System for Radio Newscast*, Thesis, Department de Teoria, UPC, March 2008.

[5] P. Woodland, M. Gales, D. Pye and S. Young, “The Development of the 1996 HTK Broadcast News Transcription System,” *Proceedings of the Speech Recognition Workshop*, pp. 73-78, 1997.

[6] F. Kubala et al., “The 1996 BBN Byblos Hub-4 transcription system,” *Proceedings of the Speech Recognition Workshop*, pp. 90-93, 1997.

[7] Xin Yan and Ying Li, “Anti-noise Power Normalized Cepstral Coefficients for Robust Environmental Sounds Recognition in Real Noisy Conditions,” Fourth International Conference on Computational Intelligence and Communication Networks, pp. 263-267, 2012.

[8] Chanwoo kim, Stern, R.M. “Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp:4101 –4104, 25-30 March 2012.

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

[10] 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.