International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 01 | Jan 2019 www.irjet.net p-ISSN: 2395-0072

Speech/Music Change Point Detection using SVM and GMM

R. Thiruvengatanadhan

Assistant Professor/Lecturer (on Deputation), 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. Systems which are developed for speech/music classification, indexing and retrieval usually take segmented audios rather than raw audio data as input. This paper presents a new technique to identify the change point detection of audio data between speech and music category. The change point detection method used in this paper is based on Mel-Frequency cepstral coefficients (MFCC) features which are used to characterize the audio data. Support Vector Machine (SVM) and Gaussian Mixture Model (GMM) are 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.*

Key Words: *Mel-Frequency cepstral coefficients (MFCC), Support Vector Machine (SVM) and Gaussian Mixture Model (GMM).*

1. INTRODUCTION

Speech\Music change point detection or segmentation aims at finding change points between two successive categories (speech and music) in an audio stream. The topic has drawn a great deal of interest in recent years, since detecting audio category changes is an important preprocessing step for various subsequent tasks such as speech\music classification, speech & music indexing and retrieval [1].

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 [2]. 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. Fig.1 shows the speech/music change point detection of audio signals.

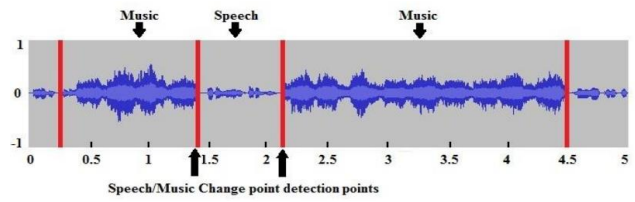

Fig -1: Speech/Music Change point detection

2. 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 [3].

2.1 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) are short-term spectral based and dominant features and are widely used in the area of audio and speech processing. The mel frequency cepstrum has proven to be highly effective in recognizing the structure of music signals and in modeling the subjective pitch and frequency content of audio signals [4]. The MFCCs have been applied in a range of audio mining tasks, and have shown good performance compared to other features. MFCCs are computed by various authors in different methods. It computes the cepstral coefficients along with delta cepstral energy and power spectrum deviation which results in 26 dimensional features. The low order MFCCs contains information of the slowly changing spectral envelope while the higher order MFCCs explains the fast variations of the envelope [5].

MFCCs are based on the known variation of the human ears critical bandwidths with frequency. The filters are spaced linearly at low frequencies and logarithmically at high frequencies to capture the phonetically important characteristics of speech and audio [6]. To obtain MFCCs, the audio signals are segmented and windowed into short frames of 20 ms. Fig. 2 describes the procedure for extracting the MFCC features.

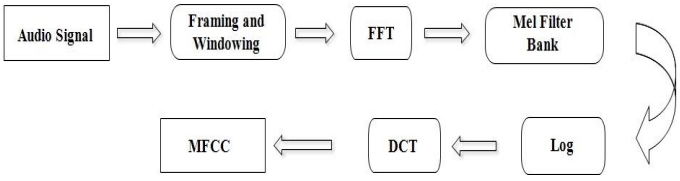

Fig -2: Extraction of MFCC from Audio Signal.

Magnitude spectrum is computed for each of these frames using fast Fourier transform (FFT) and converted into a set of Mel scale filter bank outputs. A popular solution is therefore filter bank analysis since this provides a much more straightforward route to obtain the desired non-linear frequency resolution. The magnitude coefficients are then binned by correlating them with each triangular filter. Here binning means that each FFT magnitude coefficient is

multiplied by the corresponding filter gain and the results are accumulated. Thus, each bin holds a weighted sum representing the spectral magnitude in that filter bank channel.

Logarithm is then applied to the filter bank outputs followed by discrete cosine transformation to obtain the MFCCs. Because the Mel spectrum coefficients are real numbers, they may be converted to the time domain using the Discrete Cosine Transform (DCT). In practice the last step of taking inverse DFT is replaced by taking discrete cosine transform (DCT) for computational efficiency. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Typically, the first 13 MFCCs are used as features.

3. SPEECH/MUSIC CHANGE POINT DETECTION TECHNIQUES

3.1 Support Vector Machine

A machine learning technique which is based on the principle of structure risk minimization is support vector machines. It has numerous applications in the area of pattern recognition [7]. SVM constructs linear model based upon support vectors in order to estimate decision function. If the training data are linearly separable, then SVM finds the optimal hyper plane that separates the data without error [8]. Fig. 3 shows an example of a non-linear mapping of SVM to construct an optimal hyper plane of separation. SVM maps the input patterns through a non-linear mapping into higher dimension feature space. For linearly separable data, a linear SVM is used to classify the data sets [9]. The patterns lying on the margins which are maximized are the support vectors.

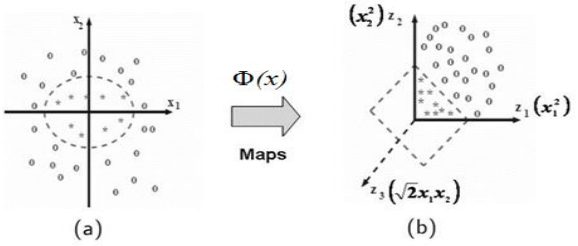

Fig -3: Example for SVM Kernel Function Φ(x) Maps 2-Dimensional Input Space to Higher 3-Dimensional Feature Space. (a) Nonlinear Problem. (b) Linear Problem.

The support vectors are the (transformed) training patterns and are equally close to hyperplane of separation. The support vectors are the training samples that define the optimal hyperplane and are the most difficult patterns to classify.

3.2 Gaussian Mixture Model

The probability distribution of feature vectors is modeled by parametric or nonparametric methods. Models which assume the shape of probability density function are termed parametric. In nonparametric modeling, minimal or no assumptions are made regarding the probability distribution of feature vectors. In this section, we briefly review Gaussian mixture model (GMM). 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 as shown in Fig. 4.

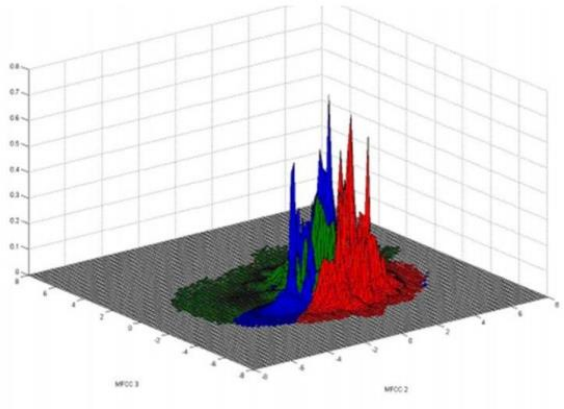

Fig -4: Gaussian Mixture Model.

GMM's represent the feature vectors using Gaussian components and are characterized by the mean vector and the co-variance matrix [10]. Even in the absence of other information, GMM models have the capability to form an arbitrarily shaped observation density [11].

4. THE PROPOSED AUDIO CHANGE POINT DETECTION ALGORITHM

4.1 Support Vector Machine

1. The sliding window is initially placed at the left end of the signal.
2. The classifier SVM model is trained to map the distribution of the feature vectors in the left and right half of the window over the hyper plane.
3. The misclassification rate of the left and right half feature vectors of the window are used for testing.
4. The SVM misclassification rates are obtained using the relations are misclassification rates of MFCC for (+1) class and (-1) class, respectively and α is a scaling factor which varies from 0 to 1.

The above process is repeated by moving the window with a shift of 10 milliseconds until it reaches the right end of the signal.

The category change points are detected from the misclassifications by applying a threshold.

A low misclassification indicates that the characteristics of the signal in the right half of the window are different from the signal in the left half of the window, and hence, the middle of the window is a category change point.

4.2 Gaussian Mixture Model

1. For the feature vectors in the left half of the window, a Gaussian distribution is fit using EM algorithm.
2. The probability density function of the feature vectors in the right of the window belonging to the GMM is computed.
3. An average probability density function is computed.
4. The sliding window proceeds with a shift of 10 milliseconds until it reaches the right end of the signal.
5. A higher probability density function indicates the characterization of the feature vectors on the right half of the window differ from those in the left of the window.
6. Hence, a change point is detected based on a threshold.

4.3 Performance Measures

The performance of speech/music change point detection is assessed in terms of two types of error namely false alarms and missed detections. A false alarm (α) of category change point detection occurs when a detected category change point is not a true one. A missed detection (β) occurs when a true category change point cannot be detected. The false alarm rate (αr) and missed detection rate (βr) are defined as

$$\alpha_r = \frac{Number\ of\ false\ alarms}{Number\ of\ actal\ category\ change\ points + number\ of\ false\ alarms} \quad (1)$$

$$\beta_r = \frac{Number\ of\ missed\ detections}{Number\ of\ actal\ category\ change\ points} \quad (2)$$

5. EXPERIMENTAL RESULTS

5.1 Database

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

5.2 Database

13-dimensional MFCC features are extracted. A frame size of 20 ms and a frame shift of 10 ms of 100 frames as window are used. Hence, an audio signal of 1 second duration results in 100 13 feature vector. SVM and GMM models are used to capture the distribution of the acoustic feature vectors.

5.3 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 proposed speech/music change point detection algorithm of confidence scores, miss classification error and probability density function of SVM and GMM, respectively are shown in Fig. 5.

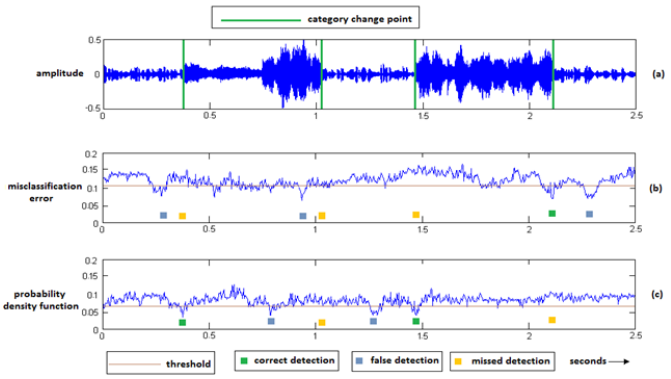

Fig -5: Speech/music Change Point Detection, (a) Sample Waveform. (b) SVM Misclassification Error of Sample Waveform. (c) GMM Probability Density Functions of Sample Waveform.

The performance of the proposed speech/music change point detection system using SVM and GMM in terms of the precision and recall measures is shown in chart-1.

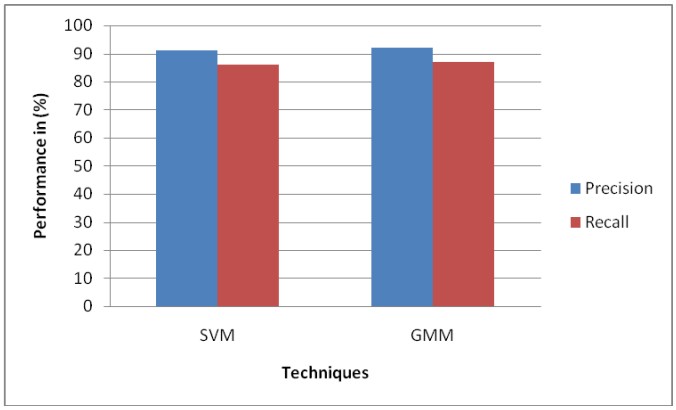

Chart -1: Performance of Speech/Music Change Point Detection using SVM and GMM.

The performance comparison of the speech/music change point detection using SVM and GMM in terms of F-measures is shown in Table 1.

Table -1: A Comparison of the performance of speech/music Change point detection using SVM and GMM in terms of the F-measures

Techniques	F-Measure
SVM	82%
GMM	85%

6. CONCLUSION

In this paper we have proposed a method for detecting the Category change point between speech/music using Support Vector Machine (SVM) and Gaussian Mixture Model (GMM). The performance is studied using 13 dimensional MFCC features. GMM based change point detection gives a better performance of 85% F-measure when compared to SVM based change point detection.

REFERENCES

[1] Li, F.F., "Nonexclusive audio segmentation and indexing as a pre-processor for audio information mining," Image and Signal Processing (CISP), 2013 6th International Congress on , vol.03, no., pp.1593,1597, 16-18 Dec. 2013

[2] 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.

[3] J. Aucouturier and F. Pachet, "Representing Musical Genre: A State of Art," Journal of New Music Research, 2002.

[4] O.M. Mubarak, E. Ambikai rajah and J. Epps, "Novel Features for Effective Speech and Music Discrimination," IEEE Engineering on Intelligent Systems, pp. 342-346, 2006.

[5] A. Meng and J. Shawe-Taylor, "An Investigation of Feature Models for Music Genre Classification using the Support Vector Classifier," *International Conference on Music Information Retrieval*, Queen Mary, University of London, UK, pp. 604-609, 2005.

[6] Francois Pachet and Pierre Roy, "Analytical Features: A Knowledge-Based Approach to Audio Feature Generation," Journal on Applied Signal Processing, 2009.

[7] Chungsoo Lim Mokpo, Yeon-Woo Lee, and Joon-Hyuk Chang, "New Techniques for Improving the practicality of a SVM-Based Speech/Music Classifier," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1657-1660, 2012.

[8] Andriansyah Zakaria, Rizal R Isnanto and Oky Dwi Nurhayati. Particle Swarm Optimization and Support Vector Machine for Vehicle Type Classification in Video Stream. International Journal of Computer Applications 182(18):9-13, September 2018.

[9] Lim and Chang, "Enhancing Support Vector Machine-Based Speech/Music Classification using Conditional Maximum a Posteriori Criterion," Signal Processing, IET, vol. 6, no. 4, pp. 335-340, 2012.

[10] Rafael Iriya and Miguel Arjona Ramírez, "Gaussian Mixture Models with Class-Dependent Features for Speech Emotion Recognition," IEEE Workshop on Statistical Signal Processing, pp. 480-483, 2014.

[11] Tang, H., Chu, S. M., Hasegawa-Johnson, M. and Huang, T. S., "Partially Supervised Speaker Clustering," IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 5, pp. 959-971, 2012.