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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Speech/music Change Point Detection using Sonogram and SVM

R. Thiruvengatanadhan

Assistant Professor, Dept. of Computer Science and Engineering, Annamalai University, Annamalainagar, Tamilnadu, 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. In this paper, Sonogram features are extracted which are used to characterize the audio data. Support Vector Machine (SVM) 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, Sonogram, SVM.

I.INTRODUCTION

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 [1].

A digital audio recording is characterized by two factors namely sampling and quantization. Sampling is defined as the number of samples captured per second to represent the waveform. Sampling is measured in Hertz (Hz) and when the rate of sampling is increased the resolution is also increased and hence, the measurement of the waveform is more precise. Quantization is defined as the number of bits used to represent each sample. Increasing the number of bits for each sample increases the quality of audio recording but the space used for storing the audio files becomes large. Sounds with frequency between 20 Hz to 20,000 Hz are audible by the human ear [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.

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 [4]. Segmentation is used to produce sequential characteristic possessing utterances in the discrete form with constant characteristics [5].

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

DOI: 10.15680/IJIRCCE.2017. 504403

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 [6].

In metric-based change point detection the audio signal is fragmented into smaller chunks that are known to contain only one type of segment [7]. 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 [8]. 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. SONOGRAM

Pre-emphasis is performed for the speech signal followed by frame blocking and windowing. The speech segment is then transformed using FFT into spectrogram representation [9]. Bark scale is applied and frequency bands are grouped into 24 critical bands. Spectral masking effect is achieved using spreading function. The spectrum energy values are transformed into decibel scale [10]. Equal loudness contour is incorporated to calculate the loudness level. The loudness sensation per critical band is computed. STFT is computed for each segment of pre-processed speech. A frame size of 20 ms is deployed with 50% overlap between the frames. The sampling frequency of 1 second duration is 16 kHz. The block diagram of sonogram extraction is shown in Fig. 1.

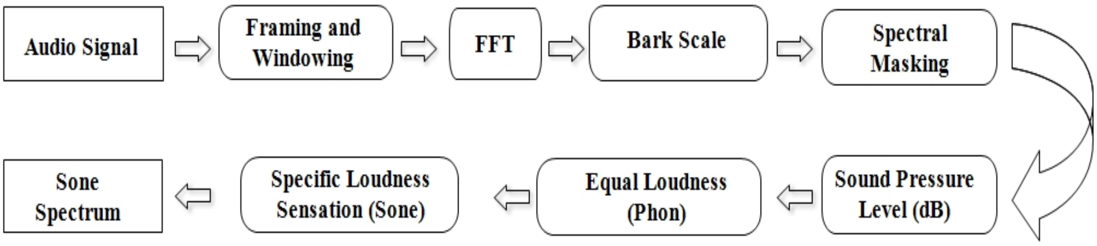

Fig. 1 Sonogram Feature Extractions.

A perceptual scale known as bark scale is applied to the spectrogram and it groups the frequencies based upon the perceptive pitch regions to critical bands. The occlusion of one sound to another is modelled by applying a spectral masking spread function to the signal [11]. The spectrum energy values are then transformed into decibel scale. Phone scale computation involves equal loudness curve which represents different perceptron of loudness at different frequencies respectively. The values are then transformed into a sone-scale to reflect the loudness sensation of the human auditory system [12].

III. SUPPORT VECTOR MACHINE (SVM)

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 [13]. 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.

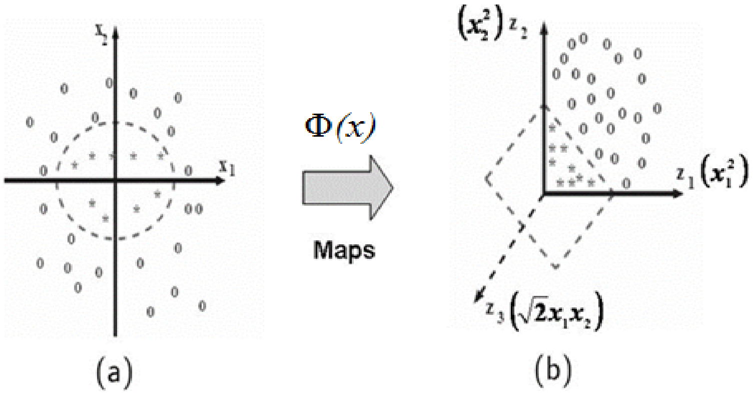

Fig. 2 Example for SVM Kernel Function Φ(x) Maps 2-Dimensional Input Space to Higher 3-Dimensional Feature Space. (a) Nonlinear Problem. (b) Linear Problem.

Fig. 2 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. The patterns lying on the margins which are maximized are the support vectors. The support vectors are the (transformed) training patterns and are equally close to hyper plane of separation. The support vectors are the training samples that define the optimal hyperplane and are the most difficult patterns to classify. Informally speaking, they are the patterns most informative of the classification task [14]. The kernel function generates the inner products to construct machines with different types of non-linear decision surfaces in the input space.

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

22 Sonogram 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 × 22 feature vector. SVM models are used to capture the distribution of the acoustic feature vectors.

C. Category change point detection

The sliding window is initially placed at the left end of the signal. 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, then the misclassification rate of the left and right half feature vectors of the window are used for testing. 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. The performance of the proposed speech/music change point detection system is shown in Fig. 3 for SVM.

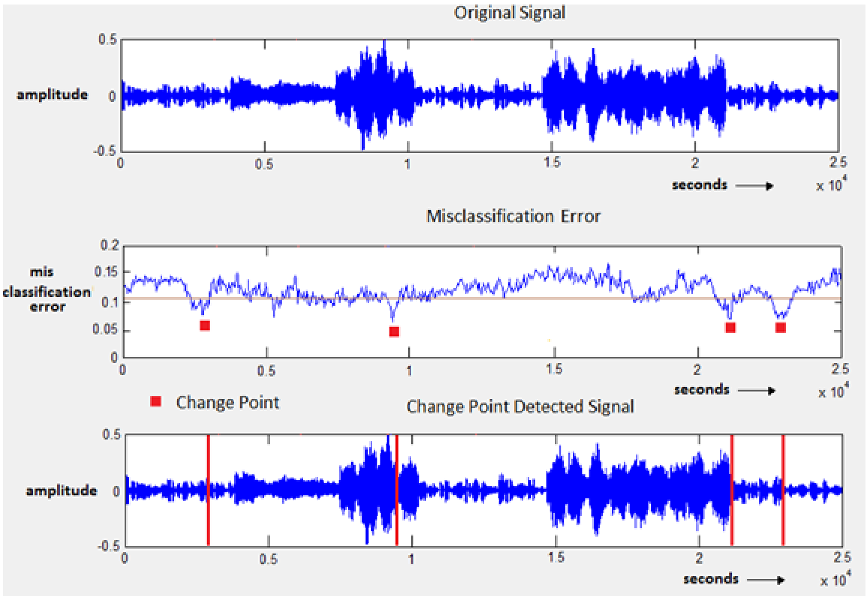

Fig. 3 Snapshot of Speech/Music Change Point Detection Systems Using SVM.

The performance of the speech/music change point detection system using SVM to detect the change point in terms of the various measures is shown in Fig. 4.

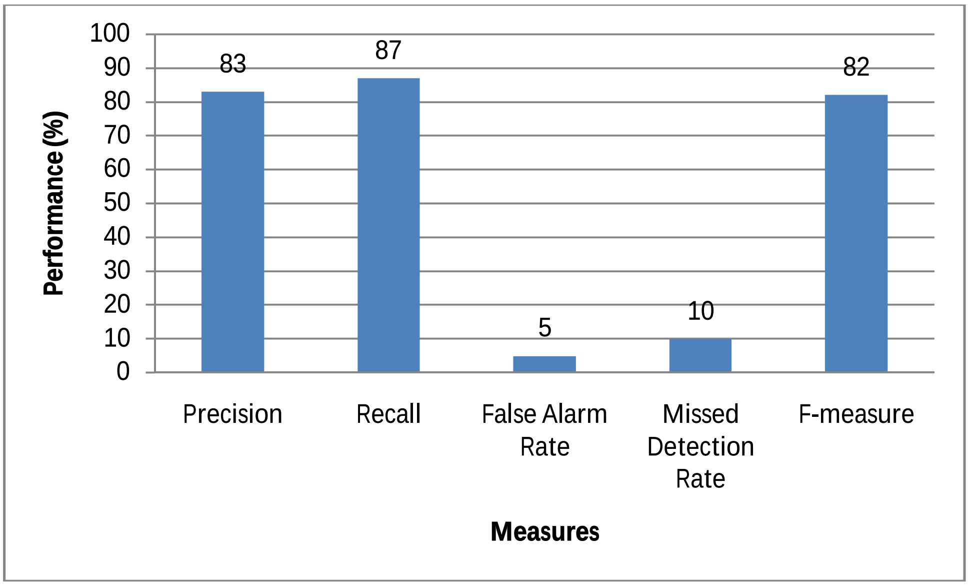

Fig. 4: Performance of detect the change point in terms of the various measures using SVM.

V.CONCLUSION

In this paper we have proposed a method for detecting the category change point between speech/music using SVM. The performance is studied using 22 dimensional Sonogram features. The performance of the system is evaluated as a large dataset collected from Television broadcast speech/music of various channels. SVM based change point detection gives a better performance of 82% F-measure is achieved.

REFERENCES

[1] 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.

[2] N. Nitananda, M. Haseyama, and H. Kitajima, "Accurate Audio-Segment Classification using Feature Extraction Matrix," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 261-264, 2005.

[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] 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.

[5] 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.

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

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

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] Theodoros Theodorou, Iosif Mporas and Nikos Fakotakis, "An Overview of Automatic Audio Segmentation," *International Journal of Information Technology and Computer Science*, vol.11, pp. 1-9, 2014.