International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 6, Number 1 (2016), pp. 45-49

http://www. irphouse.com

Speech/Music Change Point Detection using Sonogram and AANN

R. Thiruvengatanadhan

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

Abstract

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. In this paper, Sonogram features are extracted which are used to characterize the audio data. Auto associative Neural Network (AANN) 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 and AANN.

I. INTRODUCTION

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 [1].

The category change point in broadcast audio data consisting of speech and music categories [2]. The category change point detection is made using Sonogram features extracted from the broadcast audio data and techniques such as AANN are used [3].

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.

A. 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 [4]. 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 [5]. 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.

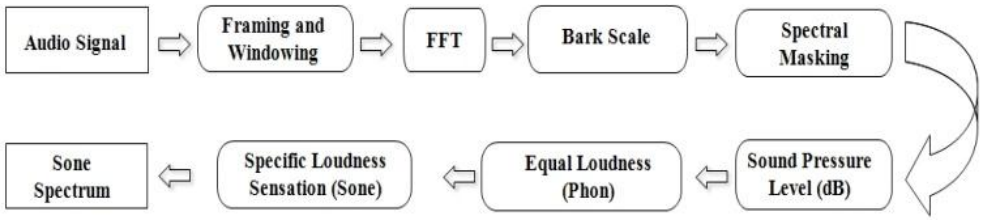

Fig. 1 Sonogram Feature Extractions.

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 [6]. 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 [7].

III. TECHNIQUES

A. Auto associative Neural Network (AANN)

Autoassociative Neural Network (AANN) model consists of five layer network which captures the distribution of the feature vector as shown in Fig. 2. The input layer in the network has less number of units than the second and the fourth layers. The first and the fifth layers have more number of units than the third layer [170]. The number of processing units in the second layer can be either linear or non-linear. But the

processing units in the first and third layer are non-linear. Back propagation algorithm is used to train the network [173].

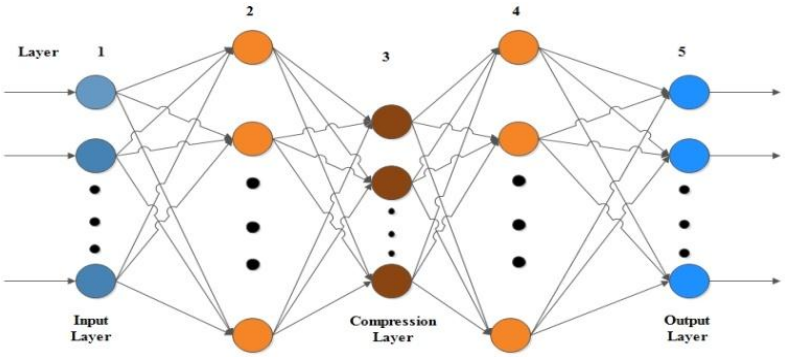

Fig. 2 The Five Layer AANN Model.

The shape of the hyper surface is determined by projecting the cluster of feature vectors in the input space onto the lower dimensional space simultaneously, as the error between the actual and the desired output gets minimized.

A new measure called a low average error can be used to achieve the best probability surface. The weights of the network are trained using back propagation algorithm to minimize the mean square error for every feature vector. The network is set to be trained for 1 epoch if the weight adjustment is done for all feature vectors in one go. An average of the mean square error is computed for successive epochs.

During testing the acoustic features extracted are given to the trained model of AANN and the average error is obtained. The structure of the AANN model used in our study is *22L 38N 8N 38N 22L* for Sonogram, for capturing the distribution of the acoustic features of a class.

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.

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. AANN models are used to capture the distribution of the acoustic feature vectors.

C. Category change point detection

AANN model is trained to capture the distribution of the feature vectors in the left half of the window. The feature vectors in the right half of the window are used for testing. The output of the model is compared with the input to compute the normalized squared error. Average confidence score is obtained for the feature vectors in the right half of the window. The above process is repeated by moving the window with a shift of 10 ms until it reaches the right end of the signal. The category change points are detected from the average confidence scores by applying a threshold. A low average confidence score 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 AANN.

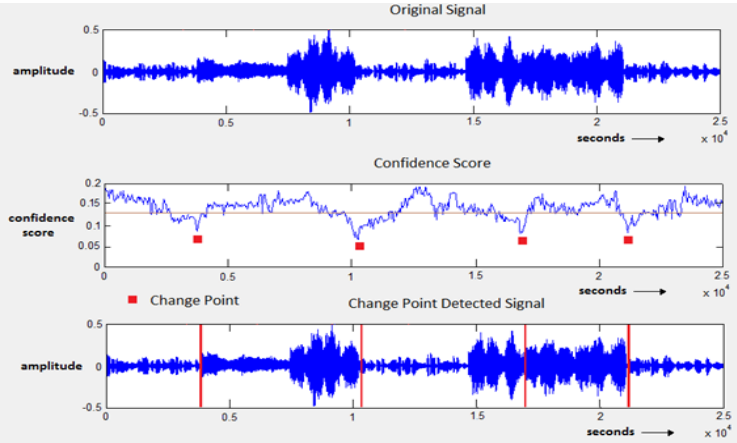

Fig. 3 Snapshot of Speech/Music Change Point Detection Systems Using AANN.

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

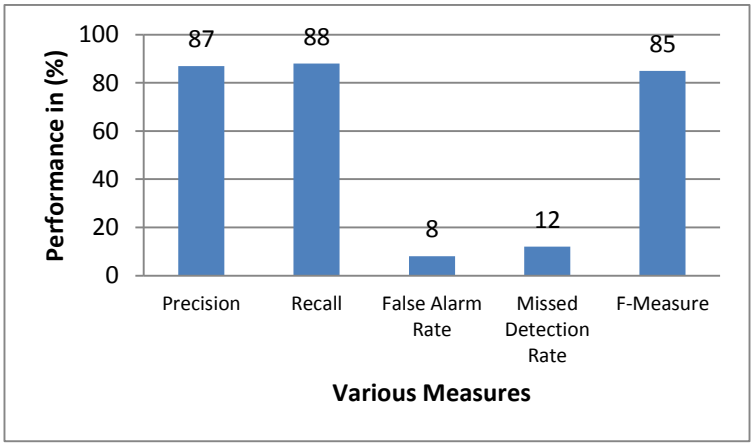

Fig. 4 Performance of to detect the change point in terms of the various measures using AANN.

V. CONCLUSIONS

In this paper we have proposed a method for detecting the category change point between speech/music using AANN. The performance is studied using 22 dimensional Sonogram features. AANN based change point detection gives a better performance of 85% F-measure is achieved.

REFERENCES

[1] 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.

[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] Xiaowen Cheng, Jarod V. Hart, and James S. Walker, "Time-frequency Analysis of Musical Rhythm," Notices of AMS, vol. 56, no. 3, 2008.

[5] Ausgef¨uhrt, Evaluation of New Audio Features and Their Utilization in Novel Music Retrieval Applications, Master's thesis, Vienna University of Technology, December 2006.

[6] Eberhard Zwicker and Hugo Fastl, "Psychoacoustics-Facts and Models," Springer Series of Information Sciences, Berlin, 1999.

[7] M. R. Schroder, B. S. Atal, and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," Journal of the Acoustical Society of America, vol. 66, pp. 1647-1652, 1979.