http://dx.doi.org/10.21172/ijiet.112.06

Speech/Music Change Point Detection using AANN

R. Thiruvengatanadhan[1]

[1]*Assistant Professor/Lecturer (on deputation), Department 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. Changes in audio signal characteristics help in detecting the category change point between different categories. Change point detection has been used extensively in various tasks such as audio classification and audio indexing. 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. Auto associative neural network 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.

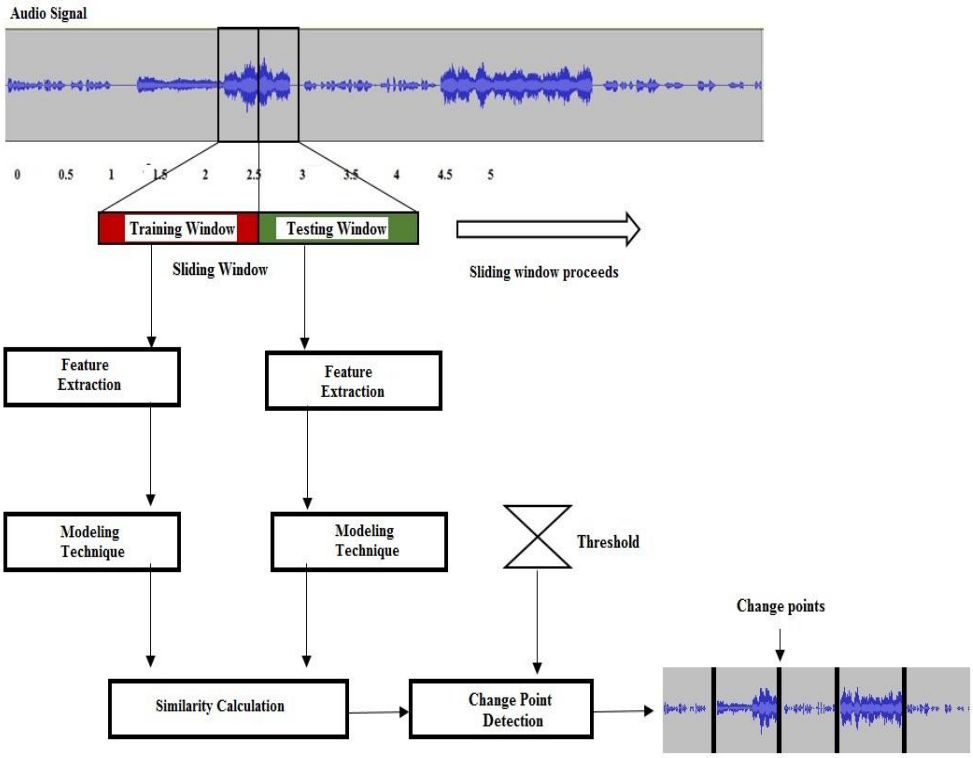
Keywords – Mel-Frequency cepstral coefficients, Auto associative neural network.

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].

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.

A first content characterization could be the categorization of an audio signal as one of speech, music, or silence [3], [4]. Approaches in speech\music change point detection can be categorized into metric-based, model-based, decoder-guided, model-selection-based and hybrid approaches. Metric-based methods simply measure the difference between two consecutive audio clips that are shifted along the audio signal, and speech\music changes are identified at the maxima of the dissimilarity in terms of some distance metric, e.g. vector quantization distortion (VQD), KL distance and divergence shape distance (DSD). Model-based approaches are based on recognizing specific speakers via Gaussian mixture models (GMM) or hidden Markov Models (HMM). Decoder guided approach segments a speech stream into male and female clips via a gender-dependent phone recognizer. In model-selection based methods, the segmentation problem is switched to a model selection problem between two nested competing models. Bayesian information criterion (BIC) is often adopted as the model selection criterion since it has some nice properties such as robustness, threshold-free and optimality. Recently, much effort has been devoted to hybrid methods that combine merits from above different approaches to achieve better performance over single approaches. Figure 1 shows the procedure of the proposed method for change point detection of audio signals.



Proposed Methods for Change Point Detection

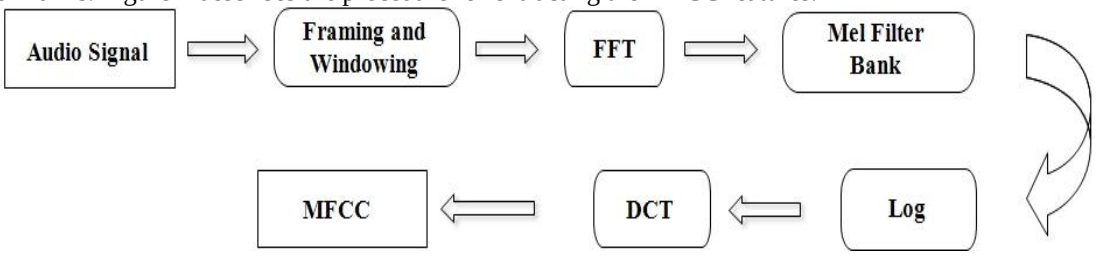
II. ACOUSTIC FEATURE EXTRACTION

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 [5]. Discriminative power of features or feature sets tells how well they can discriminate different classes.

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 [6], [7]. 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 [8].

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. To obtain MFCCs, the audio signals are segmented and windowed into short frames of 20 ms. Figure 2 describes the procedure for extracting the MFCC features.



2.2 Extraction of Mfcc from Audio Signal.

Mel frequency wrapping: 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. The human ear resolves frequencies non-linearly across the audio spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves performance. A popular solution is therefore filter bank analysis since this provides a much more straight forward route to obtain the desired non-linear frequency resolution. However, filter bank amplitudes are highly correlated and hence, the use of a cepstral transformation in this case is virtually mandatory. A simple Fourier transform based filter bank is designed to give approximately equal resolution on a Mel-scale.

To implement this filter bank, the window of audio data is transformed using a Fourier transform and the magnitude is taken. 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.

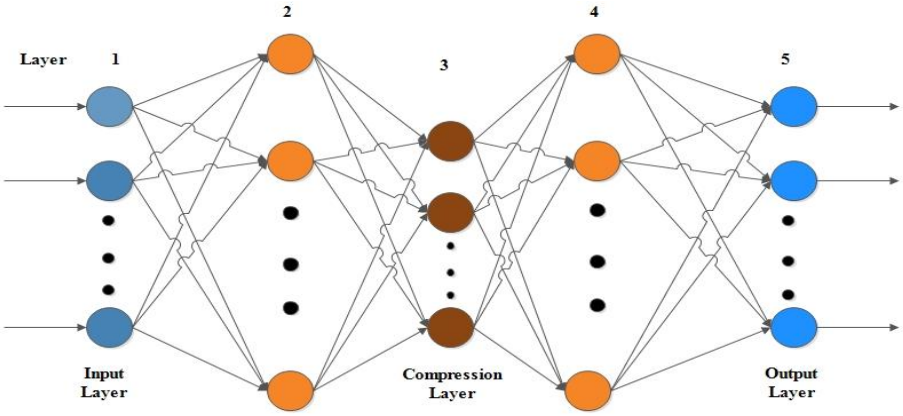
Logarithm is then applied to the filter bank outputs followed by discrete cosine transformation to obtain the MFCCs. Because the Mel spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT). In practice the 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.

Techniques

2.3 Auto associative Neural Network

Autoassociative Neural Network (AANN) model consists of five layer network which captures the distribution of the feature vector as shown in Figure 3. 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 [9]. 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 [10].

The activation functions at the second, third and fourth layer are nonlinear. The structure of the AANN model used in our study is 13L 26N 4N 26N 13L for capturing the distribution of acoustic features, where L denotes a linear unit, and N denotes anon-linear unit. The integer value indicates the number of units used in that layer. The non-linear units use tanh(s) as the activation function, where s is the activation value of the unit. Back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.



Auto associate neural network.

III. EXPERIMENTAL RESULTS

3.1 The database

Performance of the proposed audio change point detection system is evaluated using the Television broadcast audio datacollected 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.

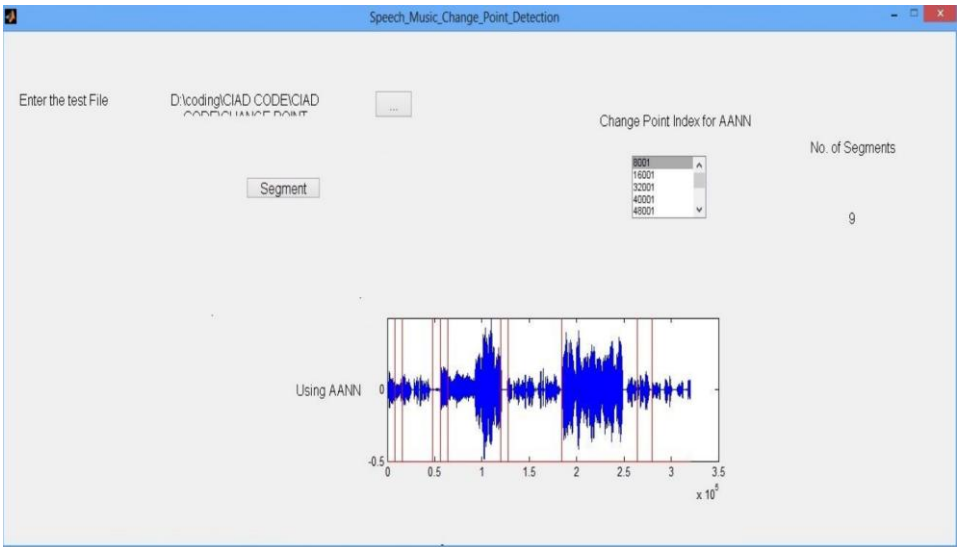
3.2 Acoustic Feature Extraction

13 MFCC 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 × 13 feature vector. AANN model are used to capture the distribution of the acoustic feature vectors.

3.3 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 performance of the proposed speech/music change point detection system. Figure 4 shows a snapshot of Speech/Music Change Point Detection Systems.

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 Table 1.



Snapshot of Speech/Music Change Point Detection Systems.

Table -1 A comparison of the performance of speech/music Change point detection using AANN in terms of the Various measures.

	Precision	Recall	False Alarm Rate	Missed Detection Rate	F-Measure
AANN	92.6%	88.3%	7.3%	12.2%	88.0%

IV. CONCLUSION

In this paper we have proposed a method for detecting the category change point between speech/music using Auto Associative Neural Network (AANN). The performance is studied using 13 dimensional MFCC features. AANN based change point detection gives a better performance of 88.0%.

V. REFERENCE

[1] Francis F. Li, "Using random forests with meta frame and meta features to enable overlapped audio content indexing and segmentation
[2] ," IEEE International Conference on International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, pp. 872-878, 2018.
[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] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content based retrieval", Pattern Recognition
[5] Letters, vol. 22, no. 1, pp. 533–544, 2001.

[6] J. Ajmera, I. McCowan, and H. Bourlard, “Speech/music segmentation using entropy and dynamism features in a hmm classification
[7] framework”, Speech Communication, no. 40, pp. 351–363, April 2003.
[8] Peter M. Grosche, Signal Processing Methods for Beat Tracking, Music Segmentation and Audio Retrieval, Thesis, Universit¨at des Saarlandes, 2012.
[9] B. Yegnanarayana, Artificial neural networks, Prentice Hall of India, New Delhi, 1999.
[10] Ahmad R. Abu-El-Quran, Rafik A. Goubran, and Adrian D. C. Chan, “Security Monitoring using Microphone Arrays and Audio Classification,” IEEE Transaction on Instrumentation and Measurement, vol. 55, no. 4, pp. 1025-1032, August 2006.
[11] 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.
[12] ShaojunRen, Fengqi Si, Jianxin Zhou, Zongliang Qiao, Yuanlin Cheng, “A new reconstruction-based auto-associative neural network for fault diagnosis in nonlinear systems,” Chemometrics and Intelligent Laboratory Systems, Volume 172, 15 January 2018, Pages 118-128N.
[13] 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.
[14] D.Tjondronegoro, Y.Chen, and B.Pham, “The power of play break for automatic detection and browsing of self consumable sport video highlights”, In Proceedings of the ACM Workshop on Multimedia Information Retrieval, pp. 267-274, 2004.