ISSN: 2229– 6948 (ONLINE)
DOI: 10.21917/ijct.2010.0027

AUTOMATIC SEGMENTATION OF BROADCAST AUDIO SIGNALS USING AUTO ASSOCIATIVE NEURAL NETWORKS

P. Dhanalakshmi[1], S. Palanivel[2] and M. Arul[3]
Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India
Email: [1]abi_dhana@rediffmail.com, [2]spal_yughu@yahoo.com
[3]Department of Business Administration, Annamalai University, Tamil Nadu, India
Email: aruldhana@yahoo.com

Abstract
In this paper, we describe automatic segmentation methods for audio broadcast data. Today, digital audio applications are part of our everyday lives. Since there are more and more digital audio databases in place these days, the importance of effective management for audio databases have become prominent. Broadcast audio data is recorded from the Television which comprises of various categories of audio signals. Efficient algorithms for segmenting the audio broadcast data into predefined categories are proposed. Audio features namely Linear prediction coefficients (LPC), Linear prediction cepstral coefficients, and Mel frequency cepstral coefficients (MFCC) are extracted to characterize the audio data. Auto Associative Neural Networks are used to segment the audio data into predefined categories using the extracted features. Experimental results indicate that the proposed algorithms can produce satisfactory results.

Keywords:
Linear Prediction Cepstral Coefficients, Mel Frequency Cepstral Coefficients, Auto Associative Neural Networks, Audio Segmentation, Audio Classification

1. RELATED WORK

During the recent years, there have been many studies on automatic audio classification and segmentation using several features and techniques. The most common problem in audio classification is speech/music classification, in which the highest accuracy has been achieved, especially when the segmentation information is known beforehand. In [1], wavelets are first applied to extract acoustical features such as sub band power and pitch information. The method uses a bottom-up SVM over these acoustic features and additional parameters, such as frequency Cepstral coefficients, to accomplish audio classification and categorization. An audio feature extraction and a multi group classification scheme that focuses on identifying discriminatory time-frequency subspaces using the Local Discriminant Bases (LDB) technique has been described in [2]. For pure music and vocal music, a number of features such as LPC and LPCC are extracted in [3], to characterize the music content. Based on calculated features, a clustering algorithm is applied to structure the music content. Audio classification is also used in the field of surveillance [4], where the authors propose a security monitoring system that can detect and classify the location and nature of different sounds within a room. This system is reliable and robust even in the presence of reverberation and in low signal-to-noise (SNR) environments.

A new approach towards high performance speech/music discrimination on realistic tasks related to the automatic transcription of broadcast news is described in [5], in which an artificial neural network (ANN) [6] [23] and hidden Markov model (HMM) are used. In [7], a generic audio classification and segmentation approach for multimedia indexing and retrieval is described. A method is proposed in [8] for Speech/Music Discrimination based on Root mean square and zero-crossings. The method proposed in [9], investigates the feasibility of an audio-based context recognition system where simplistic low-dimensional feature vectors are evaluated against more standard spectral features. Using discriminative training, competitive recognition accuracies are achieved with very low-order hidden Markov models.

The classification of continuous general audio data for content-based retrieval was addressed in [10], where the audio segments where classified based on MFCC and LPC. They also showed that cepstral-based features gave better classification accuracy. The method described in[11] content based audio classification and retrieval using joint time-frequency analysis exploits the non- stationary behavior of music signals and extracts features that characterize their spectral change over time . The audio signals were decomposed in[12], using an adaptive Time Frequency decomposition algorithm, and the signal decomposition parameter based on octave (scaling) was used to generate a set of 42 features over three frequency bands within the auditory range. These features were analyzed using linear discriminant functions and classified into six music groups.

2. OUTLINE OF THE WORK

In this paper, automatic audio feature extraction, and segmentation approaches are presented. In order to segment the six categories of broadcast audio namely music, news, sports, advertisement, cartoon and movie, a number of features such as LPC, LPCC, and MFCC are extracted to characterize the audio content. The five layer auto associative neural network model as described in Section 5 is used to capture the distribution of the audio feature vectors and find the category change point in the audio stream. The AANN model is used for capturing the distribution of the acoustic features of a class, and the back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector. Experimental results show that the segmentation accuracy of AANN with Mel cepstral features can provide a better result.

The paper is organized as follows. The acoustic feature extraction is presented in Section 2. A segmentation technique for detecting the category change point is described in section 3. Experimental results are reported in Section 4. Finally, conclusions and future work are given in Section 5.

3. ACOUSTIC FEATURE EXTRACTION

Acoustic features representing the audio information can be extracted from the speech signal at the segmental level. The segmental features are the features extracted from short (10 to 30 ms) segments of the speech signal. These features represent the short-time spectrum of the speech signal. The short-time spectrum envelope of the speech signal is attributed primarily to the shape of the vocal tract. The spectral information of the same sound uttered by two persons may differ due to change in the shape of the individual's vocal tract system, and the manner of speech production.

The selected features include linear prediction coefficients (LPC), Linear prediction derived cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC).

Linear Prediction Analysis

For acoustic feature extraction, the differenced speech signal is divided into frames of 20 ms, with a shift of 5 ms. A pth order LP analysis is used to capture the properties of the signal spectrum. In the LP analysis of speech each sample [22] is predicted as linear weighted sum of the past p samples, where p represents the order of prediction [19], [16]. If s (n) is the present sample, then it is predicted by the past p samples as

$$\hat{s}(n)= -\sum_{k=1}^{p} a_k\, s(n-k) \qquad (1)$$

The recursive relation (2) between the predictor coefficients and cepstral coefficients is used to convert the LP coefficients into LP cepstral coefficients.

$$c_0=\ln\sigma^2$$

$$c_m=a_m+\sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \qquad 1\leq m\leq p$$

$$c_m=\sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \qquad m>p \qquad (2)$$

where $c_1, c_2, \ldots, c_d$ are the cepstral coefficients, m > p and D is the number of LP cepstral coefficients . In this work, a 19 dimensional LPCC is obtained from the 14th order LP analysis for each frame.

Mel frequency cepstral coefficients

The mel-frequency cepstrum has proven to be highly effective in recognizing structure of music signals and in modeling the subjective pitch and frequency content of audio signals. Psychophysical studies have found the phenomena of the mel pitch scale and the critical band, and the frequency scale-warping to the mel scale has led to the cepstrum domain representation. The mel scale is defined as

$$F_{mel}=\frac{c\log\left(1+\frac{f}{c}\right)}{\log(2)} \qquad (3)$$

where Fmel is the logarithmic scale of f normal frequency scale. The mel-cepstral features can be illustrated by the MFCCs, which are computed from the fast Fourier transform (FFT) power coefficients. The power coefficients are filtered by a triangular band pass filter bank. When c in (5) is in the range of 250 - 350, the number of triangular filters that fall in the frequency range 200 - 1200 Hz (i.e. the frequency range of dominant audio information is higher than the other values of c. Therefore, it is efficient to set the value of c in that range for calculating MFCCs.

4. AUDIO SEGMENTATION USING AANN

Auto associative neural network models are feed forward neural networks performing an identity mapping of the input space, and are used to capture the distribution of the input data. The distribution capturing ability of the AANN model is described in this section. Let us consider the five layer AANN model shown in Fig. 1, which has three hidden layers. In this network, the second and fourth layers have more units than the input layer. The third layer has fewer units than the first or fifth. The processing units in the first and third hidden layer are nonlinear, and the units in the second compression/hidden layer can be linear or nonlinear. As the error between the actual and the desired output vectors is minimized, the cluster of points in the input space determines the shape of the hyper surface obtained by the projection onto the lower dimensional space. The nonlinear output function for each unit is tan h(s), where s is the activation value of the unit. The network is trained using back propagation algorithm. The structure of the AANN model used in our study is 14L 38N 4N 38N 14L for LPC, 19L 38N 4N 38N 19L for LPCC, 39L 38N 4N 38N 39L for MFCC, for capturing the distribution of the acoustic features of a class, where L denotes a linear unit, and N denotes a nonlinear unit. The nonlinear units use tanh(s) as the activation function, where s is the activation value of the unit. The back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.

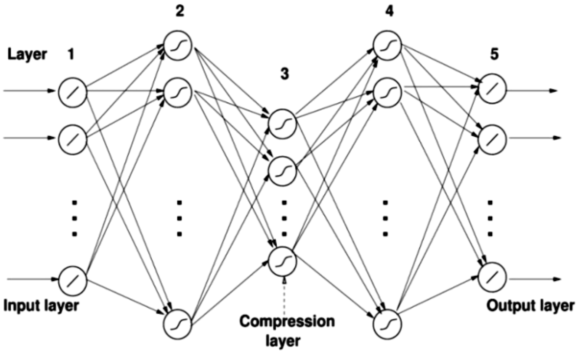

Fig.1. A five layer AANN model

4.1 AUDIO SEGMENTATION

Audio segmentation techniques detect the acoustic change point between categories such as news, movie, advertisement, song, sports, and cartoon. A sliding window is used and the feature within the window is computed. Sliding window is proceeded through the entire frames. The feature of the current window is compared with that of the previous window. A major change in feature represents a category change point.

The proposed audio segmentation algorithm:

a) Extract Audio features (LPC, LPCC, and MFCC)
b) Select a window of frames from the first frame.
c) Train AANN for the frames to the left of the centre frame

d) Test AANN for the frames to the right of the centre frame

e) Find the average confidence score.

f) Shift the window to the right by 10 frames

g) Repeat this for the entire frames

h) Identify the frames for which the confidence score is smaller than the threshold

i) Category change point is detected

4.2 CATEGORY CHANGE POINT DETECTION USING AANN

We begin with the assumption that there is a category change located at the center of the analysis window. If the audio signal of this window comes from different categories of audio broadcast data, all the feature vectors in the right half of the window may not fall into the distribution of the feature vectors from the left half of the window.

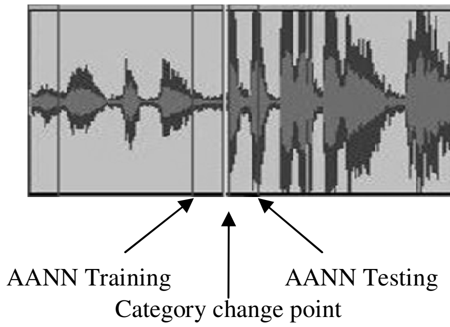

Fig.2. Segmentation in an audio broadcast data

But, if the audio signal comes from the same category, the feature vectors of the right and left half window fall under the same distribution. The average confidence score is calculated by summing the confidence score of the individual frames and the result is divided by the number of frames in the block. The frames are shifted by 10 frames until the last frame is reached. The category change points can be detected by applying a threshold. Fig 2 shows an audio signal which consists of 3 categories of broadcast data.

5. EXPERIMENTAL RESULTS

The evaluation of the proposed audio classification and segmentation algorithms have been performed by using a generic audio database which consists of the following contents: Audio samples are of different length, ranging from one second to about ten seconds, with a sampling rate of 8 kHz and 16-bits per sample. Broadcast audio data from various TV channels comprising of different categories of audio namely advertisement, news, cartoon, movie, music and sports are recorded. The wave files are converted into the raw format and features namely LPC, LPCC, MFCC are extracted. The AANN model is used to train and test the feature vectors.

The preprocessing phase included the removal of the silent frames. Only the non-silent frames were considered for training. The aim of preprocessing is to remove silence from a music sequence. Silence is defined as a segment of imperceptible audio, including unnoticeable noise and very short clicks. We use short-time energy to detect silence. The short-time energy function of a music signal is defined as where x(m) is the discrete time music signal, n is the time index of the short-time energy, and w(m) is a rectangular window, i.e., If the short-time energy function is continuously lower than a certain set of thresholds (there may be durations in which the energy is higher than the threshold, but the durations should be short enough and far apart from each other), the segment is indexed as silence. Silence segments will be removed from the audio sequence. The processed audio sequence will be segmented into fixed length and 10 ms overlapping frames. From n frames, m number of frames are selected such that m mod 2 = 1, and considered as analysis window W_k. It is assumed that the category change point occurs at the middle frame (C) of the analysis window. All the frames in the analysis window that are located to the left of C are considered as left half window and all the frames located to the right of C are considered as right half window. AANN is trained using the frames in the left half window. Then the features in the right half window are given as input to the AANN model and the output of the model is compared with the input to compute the normalized squared error e_k The average confidence score is calculated by summing the confidence score of the individual frames and the result is divided by the number of frames in the block. If a category change point occurs at C, then the average confidence score at C will be very low. Likewise, if C is not the true category change point, then the average confidence score will be very high. The frames are shifted by 10 and the average confidence score is calculated. Fig 3 shows the performance of audio segmentation for an audio clip of one second music and one second movie using LPCC.

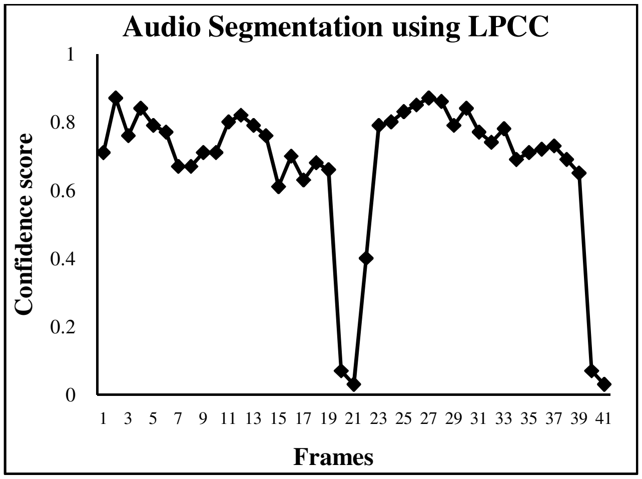

Fig.3. Performance of Audio segmentation for an audio clip of one second music and one second movie using LPCC

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an automatic audio segmentation and classification system using AANN. Linear Prediction Cepstrum coefficients (LPC, LPCC) and Mel Frequency Cepstral coefficients are calculated as features to characterize audio content. The five layer auto associative neural network model is used to capture the distribution of the acoustic feature vectors. The structure of the AANN model used in our study is described in Section 3. Experimental results show that the proposed audio segmentation scheme is very effective and the accuracy rate is 93.1%.

REFERENCES

[1] C.C. Lin, S.H. Chen, T.K. and Truong, Y. Chang, 2005, "Audio classification and categorization based on Wavelets and Support Vector Machine", IEEE Trans. Speech, Audio Processing, Vol.13, No.5, pp. 644–651.

[2] K. Umapathy, S. Krishnan and R. K. Rao, 2007, "Audio signal feature extraction and classification using local discriminant bases", IEEE Trans. Audio, Speech and Lang Processing, Vol.15, No.4, pp. 1236–1246.

[3] C. Xu, N. C. Maddage and X. Shao, 2005, "Automatic Music Classification and Summarization", IEEE Trans. Speech, Audio Processing, Vol.13, No.3, pp. 441–450.

[4] R. Abu-El-Quran, R. A. Goubran and A. D. C. Chan, 2006, "Security monitoring using Microphone arrays and Audio classification", IEEE Trans. Instrumentation and Measurement, Vol.55, No. 4, pp. 1025–1032.

[5] J. Ajmera, I. McCowan and H. Bourlard, 2003, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework", Speech Communication, Vol. 40, No. 3, pp. 351–363.

[6] S. Haykin, 2001, "Neural Networks: A Comprehensive Foundation", Second Edition, Pearson Education.

[7] S. Kiranyaz, A.F. Qureshi and M. Gabbouj, "A Generic Audio Classification and Segmentation approach for Multimedia Indexing and Retrieval", IEEE Trans. on Audio, Speech and Language Processing, Vol.14, No.3, pp. 1062–1081.

[8] Panagiotakis and G. Tziritas, 2005, "A Speech/Music discriminator based on RMS and zero-crossings", IEEE Trans. Multimedia, Vol. 7, No. 1, pp. 155–156.

[9] J. Eronen *et. al.*, 2006, "Audio-based Context Recognition," IEEE Trans. Audio, Speech and Language Processing, Vol. 14, No. 1, pp. 321–329.

[10] Li *et al.*, 2001, "Classification of General Audio data for content-based retrieval", Pattern Recognition Letters, Vol. 22, No. 5, pp. 533–544.

[11] K. Umapathy, S. Krishnan and S. Jimaa, 2005, "Multigroup classification of audio signals using time frequency parameters", IEEE Trans. Multimedia, Vol. 7, No. 2, pp.308-315.

[12] S. Esmaili. S. Krishnan and K. Raahemifar, 2004, "Content based audio classification and retrieval using joint time-frequency analysis", IEEE Int'l Conf. Acoustics, Speech and Signal Processing, pp. 665-68.