Asian Journal of Information Technology 9 (2): 54-61, 2010
ISSN: 1682-3915

Automatic Segmentation and Classification of Audio Broadcast Data

P. Dhanalakshmi, S. Palanivel and V. Ramalingam
Department of Computer Science and Engineering, Annamalai University,
Chidambaram, India

Abstract: In this study, we describe automatic segmentation and classification methods for audio broadcast data. Today, digital audio applications are part of our everyday lives. Popular examples include audio CDs, MP3 audio players, radio broadcasts, TV or video DVDs, telephones, telephone answering machines and telephone enquiries. Efficient algorithms for segmenting the audio broadcast data and classifying the audio data into predefined categories are proposed. Audio features namely Linear Prediction Coefficients (LPC), Linear prediction cepstral coefficients and Mel Frequency Cepstral Coefficients (MFCC) are used for segmenting and classifying the audio data. Experimental results indicate that the proposed algorithms can produce satisfactory results.

Key words: Linear prediction coefficients, linear prediction cepstral coefficients, mel frequency cepstral coefficients, autoassociative neural networks, audio segmentation, classification

INTRODUCTION

Today, digital audio applications are part of the everyday lives. Popular examples include audio CDs, MP3 audio players, radio broadcasts, TV or video DVDs, video games, digital cameras with sound track, digital camcorders, telephones, telephone answering machines and telephone enquiries using speech or word recognition. Audio which includes voice, music and various kinds of environmental sounds is an important type of media and also a significant part of video. Compared to research done on content-based image and video database management very little research has been done on the audio part of the multimedia stream. However, since there are more and more digital audio databases in place these days, people begin to realize the importance of effective management for audio databases relying on audio content an audio classification and segmentation can provide powerful tools for content management.

If an audio clip automatically can be classified it can be stored in an organized database, which can improve the management of audio dramatically. An audio clip can consist of several classes. It can consist of music followed by speech, which is typical in radio broadcasting. Hence, segmentation of the audio clip can be used to find where the various categories begin. This is practical for applications as audio browsers, where the user may browse for particular classes in recorded audio.

Segmentation can improve classification, when classification is coarse at points where the audio content type changes in an audio clip analysis. Audio segmentation in general is the task of segmenting a continuous audio stream in terms of acoustically homogenous regions, where the rule of homogeneity depends on the task. Audio signals which include speech, music and environmental sounds are important types of media. 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. The process of detecting the boundaries in an audio signal when there is any change in the characteristics is referred as segmentation.

Changes in audio signal characteristics such as the entrance of a guitar solo or a change from spoken words to music are some examples of segmentation boundaries. Systems that are designed for classifying audio signals usually take segmented audios rather than raw audio data as input. In order to get segmented audios from a given audio stream that contains different types of sounds, boundaries between the different audio types have to be marked. The primary and important task in audio segmentation, classification and indexing is to extract features representing the audio information in the audio signal. Feature extraction is the process of converting an audio signal into a sequence of feature vectors carrying

Corresponding Author: P. Dhanalakshmi, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India

characteristic information about the signal. These vectors are used as basis for various types of audio analysis algorithms. Feature vectors representing the audio characteristics are extracted and used for building reference models. The performance of an audio classification system depends primarily on the effectiveness of the models in capturing the audio information and hence this plays a major role in determining the performance of the audio classification system.

Audio content analysis and description has been a very active research and development topic. During the early 1990s with the advent of digital audio and video, research on audio and video retrieval become equally important. A very popular means of audio retrieval is to annotate the media with text and use text-based database management systems to perform the retrieval. However, text-based annotation has significant drawbacks when confronted with large volumes of audio data. Annotation can then become significantly labor intensive.

Furthermore, since audio data is rich in content, text may not be rich enough in many applications to describe the data. To overcome these difficulties in the early 1990s content-based audio retrieval emerged as a promising means of describing and retrieving audio data. Content-based retrieval systems describe audio data by their content rather than text. That is based on audio analysis it is possible to describe sound or music energy by its spectral energy distribution, harmonic ratio or fundamental frequency. This allows a comparison with other sound events based on these features and in some cases even a classification of sound into general sound categories.

MATERIALS AND METHODS

During the recent years, there have been many studies on 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 (Lin *et al*., 2005), 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 multigroup classification scheme that focuses on identifying discriminatory time-frequency subspaces using the Local Discriminant Bases (LDB) technique has been described by Umapathy *et al*. (2007). For pure music and vocal music, a number of features such as LPC and LPCC are extracted by Xu *et al*. (2005) 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 (Abu-El-Quran *et al*., 2006), where the researchers 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 by Ajmera *et al*. (2003) in which an Artificial Neural Network (ANN) (Haykin, 2001; Yegnanarayana, 1999) and Hidden Markov Model (HMM) are used. According to Kiranyaz *et al*. (2006), a generic audio classification and segmentation approach for multimedia indexing and retrieval is described.

A method is proposed by Panagiotakis and Tziritas (2005) for Speech/Music Discrimination based on Root mean square and zero-crossings. The method proposed by Eronen *et al*. (2006), 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 by Li *et al*. (2001), 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 by Umapathy *et al*. (2005) 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 (Esmaili *et al*., 2004) 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.

An approach given by Jiang *et al*. (2005) uses Support Vector Machine (SVM) for audio scene classification, which classifies audio clips into one of five classes: pure speech, non-pure speech, music, environment sound and silence. Radial Basis Function Neural Networks (RBFNN) are used (McConaghy *et al*.,

2003) to classify real-life audio radar signals that are collected by ground surveillance radar mounted on a tank. For audio retrieval, a new metric has been proposed by Guo and Li (2003) called Distance-From-Boundary (DFB). When a query audio is given, the system first finds a boundary inside which the query pattern is located. Then, all the audio patterns in the database are sorted by their distances to this boundary. All boundaries are learned by the SVMs and stored together with the audio database. A speech/music discrimination system was proposed based on Mel Frequency Cepstral Coefficient (MFCC) and GMM classifier (Mubarak *et al*., 2005). This system can be used to select the optimum co ding scheme for the current frame of an input signal without knowing a priori whether it contains speech-like or music-like characteristics. A hybrid model comprised of Gaussian Mixtures Models (GMMs) and Hidden Markov Models (HMMs) is used to model generic sounds with large intra class perceptual variations (Rajapakse and Wyse, 2005). The number of mixture components in the GMM was derived using the Minimum Description Length (MDL) criterion.

A new pattern classification method called the Nearest Feature Line (NFL) is proposed by Li (2000) where the NFL explores the information provided by multiple prototypes per class. Audio features like MFCC, ZCR, brightness and bandwidth, spectrum flux were extracted (Lu *et al*., 2003) and the performance using SVM, K-Nearest Neighbor (KNN) and Gaussian Mixture Model (GMM) were compared. Audio classification techniques for speech recognition and audio segmentation for unsupervised multispeaker change detection are proposed by Huang and Hansen (2006). Two new extended-time features: Variance of the Spectrum Flux (VSF) and Variance of the Zero-Crossing Rate (VZCR) are used to preclassify the audio and supply weights to the output probabilities of the GMM networks. The classification is then implemented using weighted GMM networks.

Outline of the research: In this study, automatic audio feature extraction, segmentation and classification approaches are presented. In order to discriminate 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 is used to capture the distribution of the audio feature vectors. 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 and classification accuracy of AANN with Mel cepstral features can provide a better result.

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 (10-30 msec) 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 Cepstrum 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 msec with a shift of 5 msec. A pth order LP analysis is used to capture the properties of the signal spectrum. In the LP analysis of speech each sample (Rabiner and Juang, 2003) is predicted as linear weighted sum of the past p samples, where p represents the order of prediction (Lu *et al*., 2003; Mubarak *et al*., 2005). If s (n) is the present sample, then it is predicted by the past p samples as:

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

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

$$c_0 = \ln \sigma^2 \quad (2)$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \; 1 \le m \le p \quad (3)$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k \alpha_{m-k} \quad (4)$$

where, m>p and D is the number of LP cepstral coefficients. The cepstral coefficients are linearly weighted to get the Weighted Linear Prediction Cepstral Coefficients (WLPCC). In this research, a 19 dimensional WLPCC is obtained from the 14th order LP analysis for each frame. Linear channel effects are compensated to

some extent by removing the mean of the trajectory of each cepstral coefficient. The 19 dimensional WLPCC (mean subtracted) for each frame is used as an acoustic feature vector.

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:

$$F_{mel} = \frac{c\log\left(1+\frac{f}{c}\right)}{\log(2)} \quad (5)$$

where, F_{mel} is the logarithmic scale of f normal frequency scale. The mel-cepstral features (Yegnanarayana *et al*., 2002) 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 Eq. 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.

Auto associative neural network model: 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 (Yegnanarayana *et al*., 2002). The distribution capturing ability of the AANN (Yegnanarayana and Kishore, 2002) 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 (Yegnanarayana and Kishore, 2002; Yegnanarayana, 1999). One can say that the AANN captures the distribution of the input data depending on the constraints imposed by the structure of the network, just as the number of mixtures and Gaussian functions do in the case of Gaussian mixture models.

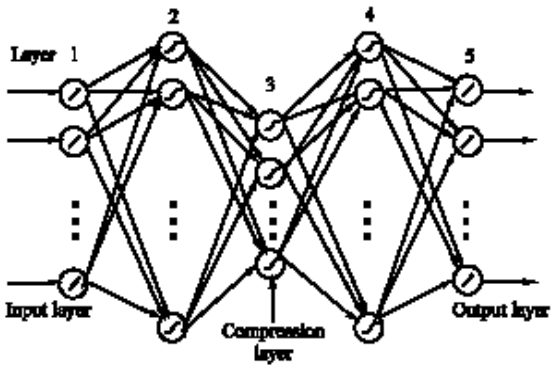

Fig. 1: A five layer AANN model

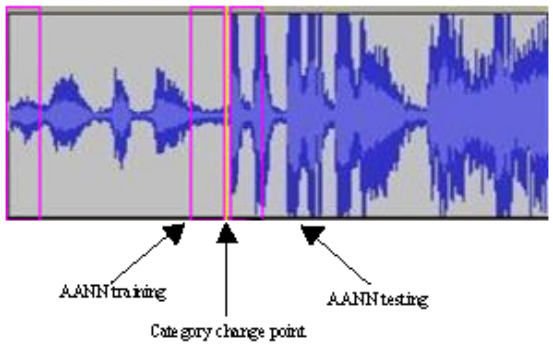

Fig. 2: Segmentation in an audio broadcast data

Audio segmentation: Audio segmentation techniques detect the acoustic change point between categories such as news, movie, advertisement, song, sports and cartoon. Figure 2 shows an audio signal which consists of 3 categories of broadcast data Use a sliding window and compute the feature within the window. 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:

- Extract Audio features (LPC, LPCC, MFCC)
- Select a window of frames from the first frame
- Train AANN for the frames to the left of the centre frame
- Test AANN for the frames to the right of the centre frame

- Find the average confidence score
- Shift the window to the right by 10 frames
- Repeat this for the entire frames
- Identify the frames for which the confidence score is smaller than the threshold
- Category change point is detected

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 AANN training AANN testing Category change point 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. 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 (Jothilakshmi *et al*., 2009). 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. The threshold (t_s) is calculated from the confidence score as follows:

$$t_s = s_{min} + as_{min}, 0 < a < 1 \quad (6)$$

where, s_{min} is the global minimum confidence score and a is the adjustable parameter.

RESULTS AND DISCUSSION

The database: 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: 100 clips of advertisement in different languages, 100 clips of songs (sung by male and female) in different languages, 100 cart on clips, 100 clips of movie from different languages, 100 clips of sports and 100 clips of news (both Tamil and English). 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. The signal duration was slightly increased using the following the rationale that the longer the audio signal analyzed, the better the extracted feature which exhibits more accurate audio characteristics. The training data should be sufficient to be statistically significant. The training data is segmented into fixed-length and overlapping frames (in the experiments we used 20 msec frames with 10 msec overlapping). When neighboring frames are overlapped the temporal characteristics of the audio content can be taken into consideration in the training process. Due to radiation effects of the sound from lips, high-frequency components have relatively low amplitude, which will influence the capture of the features at the high end of the spectrum. One simple solution is to augment the energy of the high-frequency spectrum. This procedure can be implemented via a pre-emphasizing filter that is defined as:

$$S'(n) = s(n) - 0.96s(n-1),\ n = 1,2,3.....,N-1 \quad (7)$$

where, s (n) is the nth sample of the frame s and s' (0) = s (0). Then the pre-emphasized frame is Hamming-windowed by:

$$h(n) = 0.54 - 0.46\cos(2^{\pi} n/N-1), 0 \le n \le N-1 \quad (8)$$

Preprocessing: 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.

Feature selection from non-silent frames: Feature selection is important for audio content analysis. The selected features should reflect the significant characteristics of different kinds of audio signals. The selected features include Linear Prediction coefficients, Linear Prediction derived Cepstrum Coefficients (LPCC) and Mel-Frequency Cepstrum coefficients. LPC and LPCC are two linear prediction methods and they are highly correlated to each other. LPC-based algorithms (Lin *et al*., 2005), measure three values from the audio segment to be classified. These values are the change of the energy of the signal, speech duration and the change of the pitch value. The audio signals are recorded for 60 sec at 8000 samples per second and divided into frames of 20 msec, with a shift of 10 sec. A 14th order LP analysis is used to capture the properties of the signal spectrum. The recursive relation (4) between the predictor coefficients and cepstral coefficients is used to convert the 14 LP

coefficients into 19 cepstral coefficients. The LP coefficients for each frame are linearly weighted to form the WLPCC.

In order to evaluate the relative performance of the proposed work, we compared it with the well-known MFCC features. MFCCs are short-term spectral features are widely used in the area of audio and speech processing. To obtain MFCCs (Umapathy *et al*., 2007), the audio signals were segmented and windowed into short frames of 256 samples. Magnitude spectrum was computed for each of these frames using Fast Fourier Transform (FFT) and converted into a set of mel scale filter bank outputs. Logarithm was applied to the filter bank outputs followed by discrete cosine transformation to obtain the MFCCs. For each audio signal we arrived at 39 features. This number, 39 is computed from the length of the parameterized static vector 13, plus the delta coefficients (13) plus the acceleration coefficients (13).

Audio segmentation: From n frames, m number of frames are selected such that m mod 2 = 1 and considered as analysis window W_k. W_k is the kth analysis window which is given by:

$$W_k = \{S_j\},\ k \le j \le m + k \tag{9}$$

It is assumed that the category change point occurs at the middle frame (C) of the analysis window.

$$C = k + [m^{-2}] \tag{10}$$

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 normalized squared error (e_k) for the feature vector y is given by:

$$e_k = \frac{\| y - o \|^2}{\| y \|^2} \tag{11}$$

where, o is the output vector given by the model. The error e_k is transformed into a confidence score using:

$$s = \exp(-e_k) \tag{12}$$

The average confidence score is calculated by summing the confidence score of the individual frames

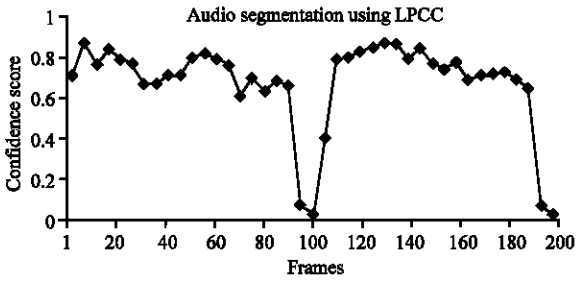

Fig. 3: Performance of audio segmentation for an audio clip of one second music and one second movie using LPCC

and the result is divided by the number of frames in the block. If a category change point occurs at c , then 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. The threshold is calculated from the confidence score as given in Eq. 6. Figure 3 shows the performance of audio segmentation for an audio clip of one second music and one second movie using LPCC.

Modelling using AANN: A is used to capture the distribution of the acoustic feature vectors. 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 tan h (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. The audio signals are recorded for 60 sec at 8000 samples per second and divided into frames of 20 msec with a shift of 10 msec.

The recursive relation between the predictor coefficients and cepstral coefficients is used to convert the 14 LP coefficients into 19 cepstral coefficients. The LP coefficients for each frame are linearly weighted to form the WLPCC. The distribution of the 14 dimensional LPC feature vectors, 19 dimensional LPCC feature vectors and 39 dimensional MFCC feature vectors in the feature space is captured using an AANN model. Separate AANN models are used to capture the distribution of feature vectors of each class. The acoustic feature vectors are given as input to the AANN model and the network is trained for 500 epochs. One epoch of training is a single

presentation of all the training vectors to the network. The training takes about 2 min on a PC with 2.3 GHz CPU. For evaluating the performance of the system, the feature vector is given as input to each of the model. The output of the model is compared with the input to compute the normalized squared error. The error (e) is transformed into a confidence score (c) using c = exp (-e). The average confidence score is calculated for each model.

The class is decided based on the highest confidence score. The performance of the system is evaluated and the method achieves 93.0% classification rate. The structure of AANN model plays an important role in capturing the distribution of the feature vectors. The number of units in the third layer (compression layer) determines the number of components captured by the network.

The AANN model projects the input vectors onto the subspace spanned by the number of units in the compression layer. If there are N_c units in the compression layer, then the acoustic feature vectors are projected onto the subspace spanned by N_c components to realize them at the output layer. The effect of changing the value of N_c on the performance of audio classification is studied. There is no major change in the performance if N_c is between 2-6 and the performance of the system decreases if it is <2 or >6.

The decrease in the performance for N_c<2 indicates that there may not be a boundary between the components representing the acoustic information. The decrease in the performance for N_c>6 indicates that the training audio data may not be sufficient for capturing the distribution of feature vectors. Similarly, the performance is studied by varying the number of units in the second layer (expansion layer) keeping the number of units in the compression layer to 4.

The experimental results show that the performance of the system decreases if the number of units in the expansion layer (N_e) is decreased to 28 but it remains the same when the number of units in the expansion layer (N_e) is increased to 48. After some trial and error, the network structure 19L 38N 4N 38N 19L is obtained.

The structure seems to give good performance in terms of computation time and EER. For testing the feature vectors extracted from the various classes are given as input to the model and the corresponding class has the maximum confidence score. The performance of audio classification in terms of number of units in the expansion layer is shown in Table 1. The duration of training data was slightly increased from 1-5 sec. The average confidence score was calculated for each model. The class is decided based on the highest confidence score. The performance of the system was evaluated and the method achieves 93.0% classification rate.

Table 1: Performance of audio classification in terms of number of units in the expansion layer

Classification rate (%)	No. of units (Ne)
89	28
93	38
93.1	48

CONCLUSION

In this study, 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 before. Experimental results show that the proposed audio segmentation and classification scheme is very effective and the accuracy rate is 93.1%. In the future, the audio classification scheme will be improved to discriminate more audio classes. We will also focus on developing an effective scheme to apply audio content analysis to assist audio indexing.

REFERENCES

Abu-El-Quran, A.R., R.A. Goubran and A.D.C. Chan, 2006. Security monitoring using microphone arrays and audio classification. IEEE Trans. Instrum. Measur., 55: 1025-1032.

Ajmera, J., I. McCowan and H. Bourlard, 2003. Speech/music segmentation using entropy and dynamism features in a HMM classification framework. Speech Commun., 40: 351-363.

Eronen, A.J., V.T. Peltonen, J.T. Tuomi, A.P. Klapuri and S. Fagerlund *et al*., 2006. Audio-based context recognition. Audio Speech Lang. Process., 14: 321-329.

Esmaili, S., S. Krishnan and K. Raahemifar, 2004. Content based audio classification and retrieval using joint time-frequency analysis. IEEE Int. Conf. Acoust. Speech Signal Process., 5: 665-668.

Guo, G. and S.Z. Li, 2003. Content-based audio classification and retrieval by supportvector machines. IEEE Trans. Neural Networks, 14: 308-315.

Haykin, S., 2001. Neural Networks a Comprehensive Foundation. Pearson Education, Asia.

Huang, R. and J.H.L. Hansen, 2006. Advances in unsupervised audio classification and segmentation for the Broadcast news and NGSW corpora. IEEE Trans. Audio Speech Lang. Process., 14: 907-919.

Jiang, H., J. Bai, S. Zhang and B. Xu, 2005. SVM-based audio scene classification. Proceeding of the IEEE, pp: 131-136.

Jothilakshmi, S., V. Ramalingam and S. Palanivel, 2009. Speaker diarization using auto associative neuralnetworks. Eng. Appl. Artif. Intell., 22: 667-675.

Kiranyaz, S., A.F. Qureshi and M. Gabbouj, 2006. A generic audio classification and segmentation approach for multimedia indexing and retrieval. IEEE Trans. Speech Audio Proces., 14: 1062-1081.

Li, D., I.K. Sethi, N. Dimitrova and T. McGee, 2001. Classification of general audio data for content-based retrieval. Pattern Recognit. Lett., 22: 533-544.

Li, S.Z., 2000. Content-based audio classification and retrieval using the nearest feature line method. IEEE Trans. Speech Audio Process., 8: 619-625.

Lin, C.C., S.H. T.K. Chen and Y.C. Truong, 2005. Audio classification and categorization based on wavelets and support vector machine. IEEE Trans. Speech Audio Process., 13: 644-651.

Lu, L., H.J. Zhang and S.Z. Li, 2003. Content-based audio classification and segmentation by using support vector machines. Multimed. Syst., 8: 482-492.

McConaghy, T., H. Leung, E. Bosse and V. Varadan, 2003. Classification of audio radar signals using radial basis function neural networks. IEEE Trans. Instrum. Measur., 52: 1771-1779.

Mubarak, O.M., E. Ambikairajah and J. Epps, 2005. Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources. IEEE Int. Conf. Acoustics Speech Signal Process., 2: 619-622.

Panagiotakis, C. and G. Tziritas, 2005. A speech/music discriminator based on RMS and zero-crossings. IEEE Trans. Multimed., 7: 155-156.

Rabiner, L. and B. Juang, 2003. Fundamentals of Speech Recognition. Pearson Education, Singapore.

Rajapakse, M. and L. Wyse, 2005. Generic audio classification using a hybrid model based on GMMs and HMM. Proceedings of IEEE 11th International Multimedia Modelling Conference, Jan. 12-14, Melbourne, Australia, pp: 53-58.

Umapathy, K., S. Krishnan and R.K. Rao, 2007. Audio signal feature extraction and classification using local discriminant bases. IEEE Trans. Audio Speech Lang. Process., 15: 1236-1246.

Umapathy, K., S. Krishnan and S. Jimaa, 2005. Multigroup classification of audio signals using time frequency parameters. IEEE Trans. Multimed., 7: 308-315.

Xu, C., N.C. Maddage and X. Shao, 2005. Automatic music classification and summarization. IEEE Trans. Speech Audio Process., 13: 441-450.

Yegnanarayana, B. and S. Kishore, 2002. AANN: An alternative to GMM for pattern recognition. Neural Networks, 15: 459-469.

Yegnanarayana, B., 1999. Artificial Neural Networks. Prentice Hall of India, New Delhi.

Yegnanarayana, B., S. Gangashetty and S. Palanivel, 2002. Autoassociative Neural Network Models for Pattern Recognition Tasks in Speech and Image. In: Soft Computing Approach to Pattern Recognition and Image Processing, Ghosh, A. and S.K. Pal (Eds.). World Scientific Publishing Co. Pvt. Ltd., Singapore, pp: 283-305.