Speech/music classification using PLP and AANN

R. Thiruvengatanadhan
Assistant Professor (On Deputation)
Department of Computer Science and Engineering
Annamalai University, Annamalainagar, Tamilnadu, India

***Abstract*— The term audio is used to indicate all kinds of audio signals, such as speech, music as well as more general sound signals and their combinations. This paper deals with the Speech/Music classification problem, starting from a set of features extracted directly from audio data. Automatic audio classification is very useful in audio indexing; content based audio retrieval and online audio distribution. The accuracy of the classification relies on the strength of the features and classification scheme. In this work Perceptual Linear Prediction (PLP) features are extracted from the input signal. After feature extraction, classification is carried out, using Auto associative neural network (AANN) model. The proposed feature extraction and classification models results in better accuracy in speech/music classification.**

***IndexTerms*—Speech, Music, Feature Extraction, PLP, AANN.**

I. INTRODUCTION

Speech is transmitted through sound waves, which follow the basic principles of acoustics. Music is an art form whose medium is sound and silence. Pitch, rhythm, dynamics, and the sonic qualities are common elements of timbre and texture. The term audio is used to indicate various kinds of audio signals, such as speech, music as well as more general sound signals combinations of audio recordings [1]. However, the audio is usually treated as an opaque collection of bytes with only the most primitive fields attached; namely, file format, name, sampling rate, etc. Meaningful information can be extracted from digital audio waveforms in order to compare and classify the data efficiently [2]. Approaches in speech\music change point detection can be categorized into metric-based, model-based, decoder-guided, model-selection-based and hybrid approaches [3]. 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. During the recent years, there have been many studies on automatic audio classification using several features and techniques. A data descriptor is often called a feature vector and the process for extracting such feature vectors from audio is called audio feature extraction. Usually a variety of more or less complex descriptions can be extracted to feature one piece of audio data.

II. PERCEPTUAL LINEAR PREDICTION (PLP)

Hermansky developed a model known as PLP. It is based on the concept of psychophysics theory and discards unwanted information from the human pitch [4]. It resembles the procedure to extract LPC parameters except that the spectral characteristics of the speech signal are transformed to match the human auditory system. PLP is the approximation of three aspects related to perceptron namely resolution curves of the critical band, curve for equal loudness and the power law relation of intensity loudness.

The process of PLP computation is shown in Fig 1. The audio signal is hamming windowed to reduce discontinuities. The Fast Fourier Transform (FFT) transforms the windowed speech segment into the frequency domain [5]. The auditory warped spectrum is convolved with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing. Critical band is the frequency bandwidth created by the cochlea, which acts as an auditory filter. The cochlea is the hearing sense organ in the inner ear. Bark scale corresponds to 1 to 24 critical bands. The power spectrum of the critical band masking curve and auditory warped spectrum are convoluted to simulate the human hearing resolution. The equal loudness pre-emphasis needs to compensate the unequal perception of loudness at varying frequencies.

A weight function is added to the sampled values using an equal loudness curve to simulate the human hearing sensitivity at varying frequencies. The intensity loudness power law is an approximation of the power law of hearing, which relates sound intensity and perceived loudness of the sound [6]. Each intensity is raised to the power of 0.33 as stated by the power law and thus the equalized values are transformed. An all pole model normally applied in Linear Prediction (LP) analysis is used to approximate the spectral samples. Either the coefficients can be used as such for representing the signal or they can further be transformed to Cepstral coefficients. In this work, a 9th order LP analysis is used to approximate the spectral samples and hence obtained a 9-dimensional feature vector for a speech signal of frame size of 20 milliseconds is obtained.

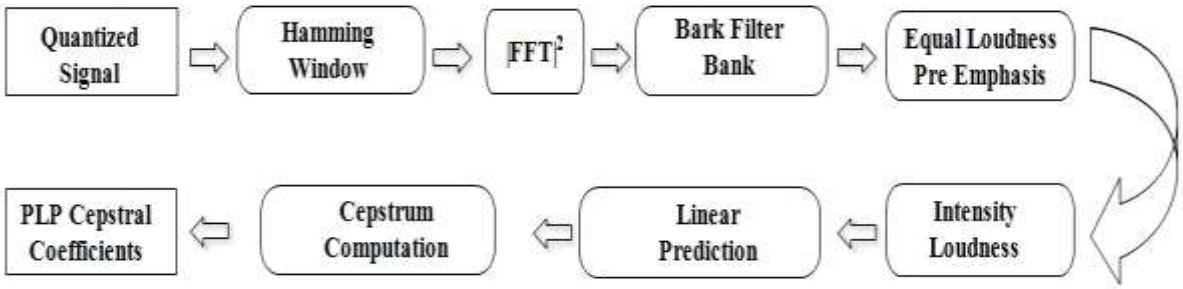

Fig. 1 PLP Parameter Computations.

III. AUTOASSOCIATIVE NEURAL NETWORK

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 [7]. 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 [8].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.

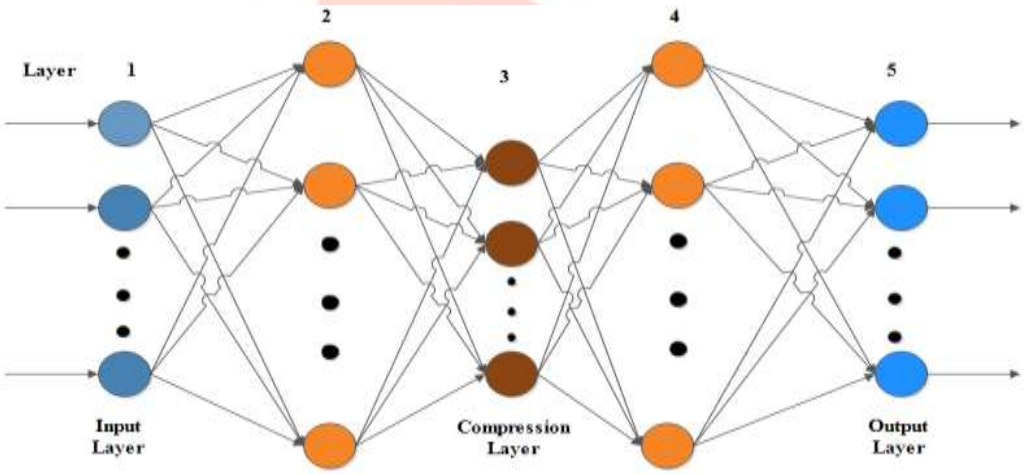

Fig. 2 The Five Layer AANN Model.

The activation function used by the non-linear units is tanh(s), where s is the activation value of the unit. The number of units used in each layer is indicated by the integer value. The weights of the network are adjusted to minimize the mean square error for every feature vector using back propagation learning algorithm. A new measure called a low average error can be used to achieve the best probability surface [9]. 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 9L 18N 3N 18N 9L for PLP, for capturing the distribution of the acoustic features of a class.

IV. RESULTS AND DISCUSSION

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.

Acoustic feature extraction

An input wav file is given to the feature extraction techniques. PLP 9 dimensional feature values will be calculated for the given wav file. The above process is continued for 100 number of wav files.

Classification

An AANN model is used to capture the distribution of six dimensional spectral and six dimensional of DWT features respectively. The feature vectors are given as input and compared with the output to calculate the error. In this experiment the network is trained for 500 epochs. The confidence score is calculated from the normalized squared error and the category is decided based on highest confidence score. The network structures 9L 18N 3N 18N 9L gives a good performance and this structure is obtained after some trial and error. Fig. 3 shows the performance of AANN for speech/music classification for various durations of training data.

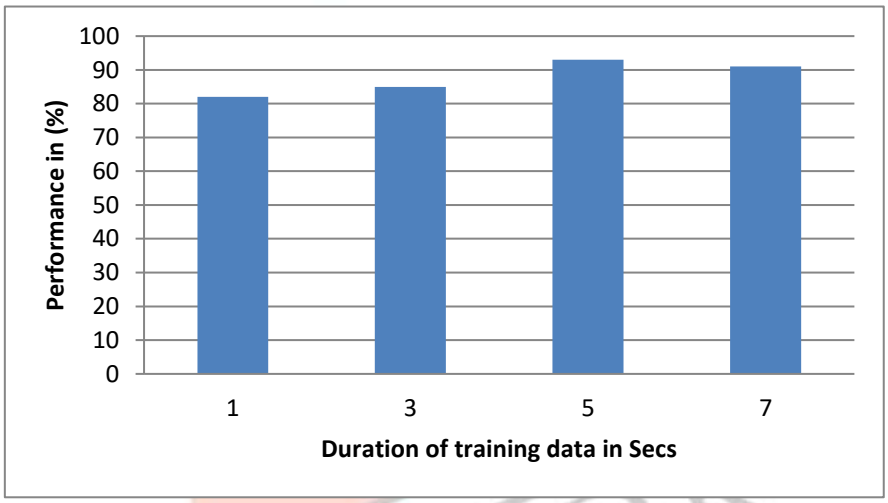

Fig.3 Performance of AANN for Speech/Music Classification.

V. CONCLUSIONS

The system classifies the audio data into speech or music. It is currently the state of the art approach for categorization. In order to classify the audio first the feature extraction is done using PLP feature. After feature extraction classification process is done using the AANN. The AANN classifier trains the feature vectors to create models for classes. The AANN test the input audio data based on the models created by the AANN train and produce the result data. Based on the result data the input audio is classified into speech or music. AANN based speech/music classification gives a better performance of 93%.

REFERENCES

[1] Lim C, J-H(2012) ' Enhancing support vector machine-based speech/music classification using conditional maximum a posteriori criterion.',Signal Processing, IET ,vol. 64, pp 335-340.

[2] Chungsoo Lim Mokpo YWL, Chang JH (2012) ' New techniques for improving the practicality of an svm-based speech/music classifier.',Acoustics, Speech and Signal Processing (ICASSP) , pp 1657-1660.

[3] Changsheng Xu NCM, Shao X(2005) ' Automatic music classification and summarization. ',IEEE Trans Speech and Audio Processing , vol. 13, pp 441450.

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

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

[6] Poonam Sharma and Anjali Garg. Feature Extraction and Recognition of Hindi Spoken Words using Neural Networks. International Journal of Computer Applications 142(7):12-17, May 2016.

[7] N J Nalini and S Palanivel. Article: Emotion Recognition in Music Signal using AANN and SVM. International Journal of Computer Applications 77(2):7-14, September 2013.

[8] D. Li, I. K. Sethi, N. Dimitrova, and T. Mc Gee, "Classification of General Audio Data for Content Based Retrieval," Pattern Recognition Letters, vol. 22, no. 1, pp. 533-544, 2001.

[9] 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.