Speech/Music Classification using RBFNN

R. Thiruvengatanadhan
Assistant Professor
Department of Computer Science and Engineering
Annamlai University, Annamalainagar, Tamilnadu, India

Abstract : Audio classification serves as the fundamental step towards the rapid growth in audio data volume. 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 a speech/music discrimination system is developed which utilizes the Discrete Wavelet Transform (DWT) as the acoustic feature. This paper analyses neural networks and their precision when they both stumble upon same targets in similar category. The analysis is done on radial basis function neural network (RBFNN) then a conclusion is formed on the basis of their performance and efficiency.

IndexTerms **– Feature Extraction, Pattern Classification, Discrete Wavelet Transform, Radial Basis Function Neural Network.**

I. INTRODUCTION

Audio refers to speech, music as well as any sound signal and their combination. Audio consists of the fields namely file name, file format, sampling rate, etc. To compare and to classify the audio data effectively, meaningful information is extracted from audio signals which can be stored in a compact way as content descriptors. These descriptors are used in segmentation, storage, classification, reorganization, indexing and retrieval of data. During recent years audio classification is emerging as an important research area because there is a vast need to classify and to categorize the audio data automatically [1]. Audio feature extraction is the process of extracting meaningful information from the audio signal. The features can be more or less complex descriptions and performance of such features depends on the process of extraction [2]. The music signal is a special class in the signal category that has its own characteristics different from the speech signal in many ways. First of all, music normally has a wide range frequency distribution among the audible range of human, from 0 to 20k Hz.

The bandwidth of the speech signal is usually limited into 50 Hz to 7 k Hz and hence, the spectral centroids of music signal are higher than that of the speech. In addition, for considering time-domain characteristics, musical signal usually has a lower silence ratio except that it is sung by a singer or played on a solo instrument only. Compared to an ordinary speech signal, music has lower variability in zero-crossing rate. Besides, music has normally more harmonic than other sound. Therefore, music has higher harmonic than speech. Music usually has regular beats that can be extracted to differentiate it from speech for the sake of the melody and background noise.

The quality of a digital audio recording depends heavily on two factors: the sample rate and the sample format or bit depth. Increasing the sample rate or the number of bits in each sample increases the quality of the recording, but also increases the amount of space used by audio files on a computer or disk. Sample rates are measured in hertz (Hz), or cycles per second. This value simply represents the number of samples captured per second in order to represent the waveform; the more samples per second, the higher the resolution, and thus the more precise the measurement is of the waveform. Fig.1 shows the speech/music change point detection of audio signals.

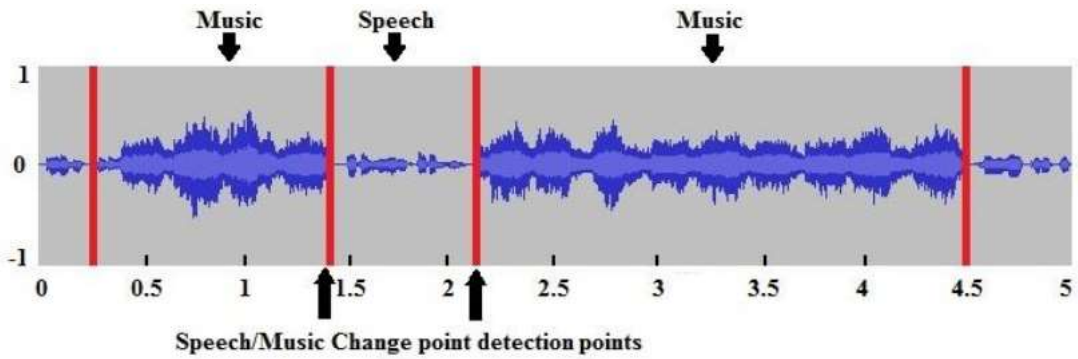

Fig. 1. Speech/Music Change point detection

II. 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. An audio feature extraction and a multi-group classification scheme that focuses on identifying discriminatory time-frequency subspaces using the Local Discriminate Bases(LDB) technique has been described in [3]. For pure music and vocal music,a number of features such as LPC and LPCC are extracted in [4] to characterize the music content. Based on calculated features, a clustering algorithm is applied to structure the music content.

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 anArtificial Neural Network (ANN) and HIDDEN Markov Model (HMM) are used. [6], a generic audio classification and segmentation approach for multimedia indexing and retrieval is described. A method is proposed in [7] for speech/music discrimination based on root mean square and zero-crossings.

The method proposed in [8], 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. [9] a speech/music discrimination system was proposed based on Mel-Frequency Cepstral Coefficient (MFCC) and GMM classifier. This system can be used to select the optimum coding scheme for the current frame of an input signal without knowing a priori whether it contains speech-like or music-like characteristics. The classification of continuous general audio data for content-based retrieval was addressed. The DWT is computed by successive low pass and high pass filtering of the discrete time-domain signal which extracts features that characterize their spectral change over time.

Li et al [10] use the local and global information of music signals to compute the histograms, and a comparative study on many features and several machine learning algorithms including support vector machines, K-Nearest Neighbor, Gaussian Mixture Models and Linear Discriminant Analysis is done. The results show that the proposed wavelet coefficients histogram method can achieve good results. [11] utilized the Gaussian Mixture Models (GMM) to train the MFCCs and achieved a good result.

III. Outline of the Work

In this study, automatic audio feature extraction and classification approaches are presented. In order to discriminate the speech and music features such as Discrete Wavelet Transform are extracted to characterize the audio content. RBFNN is applied to obtain select an optimal RBFN model between the classes by learning from training data. Experimental results show that the classification accuracy of RBFNN with DWT features can provide a better result. Fig. 2 1illustrates the block diagram of Speech/Music classification system.

Fig. 2. Block Diagram for speech/music classification.

IV. ACOUSTIC FEATURE EXTRACTION

The Discrete Wavelet Transform (DWT), which is based on sub-band coding, is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required. The foundations of DWT go back to 1976 when techniques to decompose discrete time signals were devised. Similar work was done in speech signal coding which was named as sub-band coding. In 1983, a technique similar to sub-band coding was developed which was named pyramidal coding. Later many improvements were made to these coding schemes which resulted in efficient multi-resolution analysis schemes. In DWT, a time-scale representation of the digital signal is obtained using digital filtering techniques. The signal to be analyzed is passed through filters with different cutoff frequencies at different scales. Filters are one of the most widely used signal processing functions. The wavelet analysis process is to implement a wavelet prototype function, known as analyzing wavelet or mother wavelet. Coefficients in a linear combination of the wavelet function can be used in order to represent the development of the original signal in terms of a wavelet, data operations can be performed with the appropriate wavelet coefficients. Choose the best wavelets adapted to represent your data, also truncate the coefficients below a threshold [12]. Wavelets can be realized by iteration of filters with rescaling.

The resolution of the signal, which is a measure of the amount of detail information in the signal, is determined by the filtering operations and the scale is determined by up sampling and down sampling (sub sampling) operations [13]. The DWT is computed

by successive low pass and high pass filtering of the discrete time-domain signal as shown in Fig. 3. This is called the Mallat algorithm or Mallat-tree decomposition. Its significance is in the manner it connects the continuous-time multi resolution to discrete-time filters.

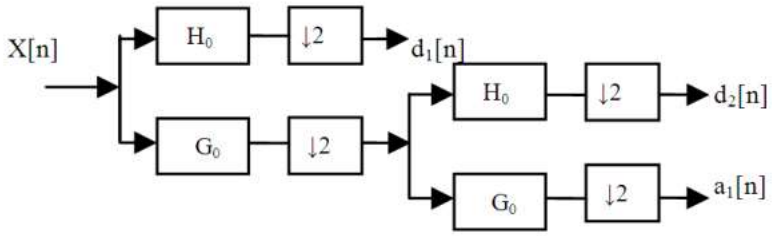

Fig. 3.Two level wavelet decomposition technique.

V. CLASSIFICATION MODEL

Radial basis function neural network (RBFNN) forms a special architecture with several distinctive features. A typical RBF neural network classifier has three layers, namely input, hidden, and output layer. The input layer of the network is made of source nodes that connect the coordinates of the input vector to the nodes in the second layer. The second layer, the only hidden layer in the network, includes processing units called the hidden basis function units which are located on the centers of well chosen clusters. Each hidden layer node adopts a radial activated function, and output nodes implement a weighted sum of hidden unit outputs [14]. The output layer is linear, and it produces the predicted class labels based on there sponse of the hidden units. The structure of multi-input and multi-output RBF neural network is represented by Fig. 4.The parameters of an RBF type neural network consist of the centers spread the basis functions at the hidden layer nodes and the synaptic weights of the output layer nodes. The RBF centers are also points in the input space. It would be ideal to have them at each distinct point on the input space, but for any realistic problem, only a few input points from all available points are selected using clustering.

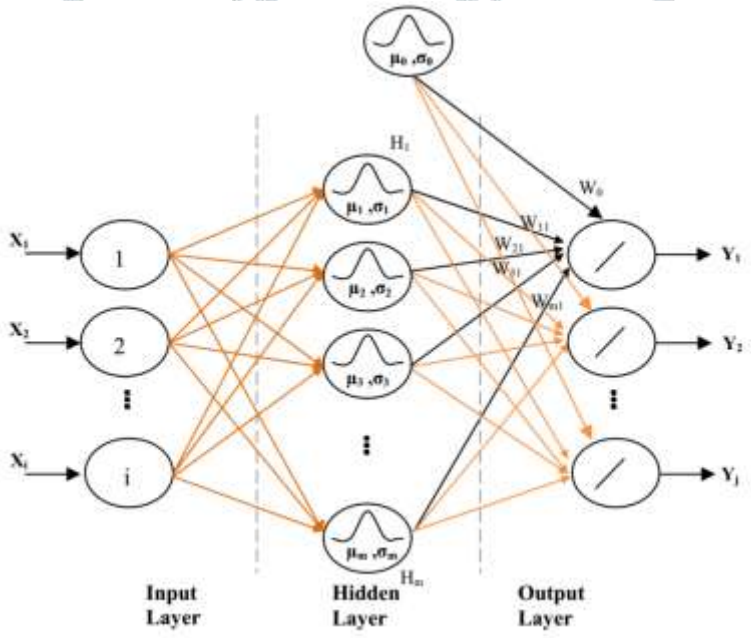

Fig. 4: RBFNN Architecture

VI. RESULTS AND DISCUSSION

6.1 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.

6.2 Acoustic feature extraction

The feature is extracted from each frame of the audio by using the feature extraction techniques. Here the DWT features are taken. An input wav file is given to the feature extraction techniques. The feature values will be calculated for the given wav file. The feature values for all the wav files will be stored separately for speech and music.

6.3 Classification

When the feature extraction process is done the audio should be classified either as speech or music. In a more complex system more classes can be defined, such as silence or speech over music. The latter is often classed as speech in systems with only two basic classes. The extracted feature vector is used to classify whether the audio is speech or music. A method where the classification is based on the output of many frames together is proposed. In this method, based on the output the feature values are extracted from the speech/music wav file and it is appended with two categories. One category is appended for speech wav and the other category is appended for the music wav. By using the feature values with appended value RBFNN training is carried out. For testing the feature extraction is done on different speech and music wav files other than the speech and music wav files used in the training set. All the values would be used for testing, the SVM tests the features based on models created during the training.

The RBFN is trained by adaptively updating the free parameters, i.e. center and width of the basis function, and the weight between the hidden and output neurons of the network. To select an optimal RBFN model, the number of neurons in the hidden layer was varied from 2 to 30, and the learning rate was varied between 0.05 and 0.5. The initial basis function centers were chosen randomly from the input space, and the initial weight values were chosen randomly between ±0.9. Normalized datasets were used for the training, testing, and validation of the RBFN model. The best network was found to be one having 26 basis functions with a learning rate of 0.9 and 0.05 for center and weight respectively. The prediction errors of the validation patterns are larger because these patterns are outside the training space.

Table 6.1: Performance of classification for different Wavelet Transforms

Mother wavelet	Speech (%)	Music (%)	Overall (%)
Haar	90.5	87.6	89.5
Symlets2	89.6	86.4	88.3
Daubechies8	93.0	91.0	92.4

When Table 6.2 is taken into consideration, it can be seen that wavelet based parameters have higher classification results than traditional features. The best performance has been obtained with Daubechies8 wavelet. The Fig. 5 shows the comparison of various means in RBFNN.

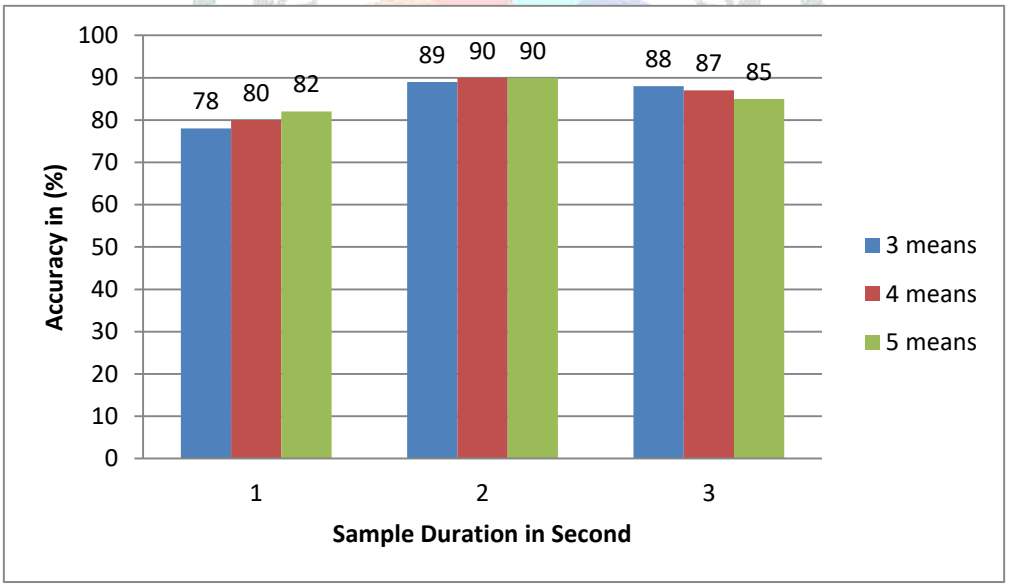

Fig: 5 Comparison graph for various means in RBFNN

VII. CONCLUSION

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 DWT feature. In this paper we have proposed a method for detecting the category change point between speech/music using Radial Basis Function Neural Network (RBFNN). The performance is studied DWT features. RBFNN based change point detection gives a better performance of 92.1%.

REFERENCES

[1] Bhaumik Choksi, Alisha Sawant and Swati Mali. Style Transfer for Audio using Convolutional Neural Networks. *International Journal of Computer Applications* 175(8):17-20, October 2017.

[2] Tayseer M F Taha and Amir Hussain. A Survey on Techniques for Enhancing Speech. *International Journal of Computer Applications* 179(17):1-14, February 2018.

[**3**] K Prakash and Hepzibha Rani D. Article: Blind Source Separation for Speech Music and Speech Mixtures. International Journal of Computer Applications 110(12):40-43, January 2015.

[**4**] Vyankatesh Kharat, Kalpana Thakare and Kishor Sadafale. Article: A Survey on Query by Singing/Humming. International Journal of Computer Applications 111(14):39-42, February 2015.

[5] Frikha, M. and A.B. Hamida, 2012. A comparitive survey of ANN and hybrid HMM/ANN architectures for robust speech recognition. Am. J.Intell. Syst., 2: 1-8. DOI: 10.5923/j.ajis.20120201.01

[**6**] Subashini, K., S. Palanivel and V. Ramaligam, 2012. Audio-video based segmentation and classification using AANN. Int. J. Comput. Applic. Technol., 1: 53-56. DOI: 10.7753/IJCAT0102.1003

[**7**] F. Lu, J. Haung, "An Improved Local Binary Pattern Operator for Texture Classification," International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[**8**] Jiang, Y.G., S. Bhattacharya, S.F. Chang and M. Shah, 2013. High-level event recognition in unconstrained videos. Int. J. Multimed. Inform. Retr., 2: 73-101. DOI: 10.1007/s13735-012-0024-2

[**9**] Feki, I., A.B. Ammar and A.M. Alimi, 2012. New process to identify audio concepts based on binary classifiers framework. Int. J. Comput. Electr. Eng., 4: 515-518.

[**10**] T. Li, M.O. M, Q. Li, "A comparative study on content-based music genre classification," in international ACM SIGIR conference on research and development in information retrieval, 2003, pp. 282–289

[**11**] H. Ezzaidi, J. Rouat, "Automatic musical genre classification using divergence and average information measures," in Research report of the world academy of science, engineering and technology, 2006

[**12**] Rekik, S., D. Guerchi, H. Hamam and S.A. Selouani, 2012. Audio steganography coding using the discrete wavelet transforms. Int. J. Comput. Sci. Security, 6: 79-83.

[**13**] Patil, V.D. and S.D. Ruikar, 2012. Wavelet-based image enhancement using nonlinear anisotropic diffusion. Int. J. Adv. Res. Comput. Sci. Software Eng., 2: 158-162.

[**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.