Speech/music Classification using Perceptual Linear Prediction

R. Thiruvengatanadhan

Assistant Professor/Lecturer (on Deputation), Department of Computer Science and Engineering
Annamalai University, Annamalainagar, Tamil Nadu, 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. 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 Gaussian Mixture Model (GMM) model. The proposed feature extraction and classification models results in better accuracy in speech/music classification.*

Key Words: Feature Extraction, Perceptual Linear Prediction (PLP) and Gaussian Mixture Model (GMM).

1. 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].

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. The efficiency of a particular feature used for comparison and classification depends greatly on the application, the extraction process and the richness of the description itself [2]. Digital analysis may discriminate whether an audio file contains speech, music or other audio entities.

2. ACOUSTIC FEATURES FOR AUDIO CLASSIFICATION

An important objective of extracting the features is to compress the speech signal to a vector that is representative of the meaningful information it is trying to characterize. In these works, acoustic features namely PLP features are extracted.

2.1 Perceptual Linear Prediction

Hermansky developed a model known as PLP. It is based on the concept of psychophysics theory and discards unwanted information from the human pitch [3]. 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.

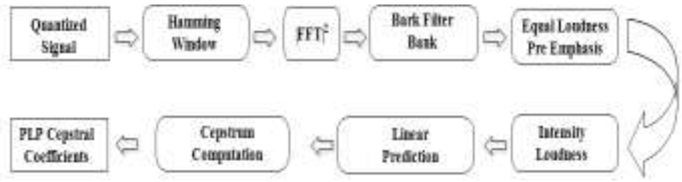

Fig -1: PLP Parameter Computations.

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 [4]. 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 [10]. 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.

3. CLASSIFICATION MODEL

3.1 Gaussian Mixture Models

Parametric or non-parametric methods are used to model the distribution of feature vectors. Parametric models are based on the shape of probability density function [5]. In non-parametric modeling only minimal or no assumption regarding the probability density function of feature vector is made [6]. The Gaussian mixture model (GMM) is used in classifying different audio classes. The Gaussian classifier is an example of a parametric classifier. It is an intuitive approach when the model consists of several Gaussian components, which can be seen to model acoustic features. In classification, each class is represented by a GMM and refers to its model. Once the GMM is trained, it can be used to predict which class a new sample probably belongs to [7].

The probability distribution of feature vectors is modeled by parametric or non-parametric methods. Models which assume the shape of probability density function are termed parametric. In non-parametric modeling, minimal or no assumptions are made regarding the probability distribution of feature vectors. The potential of Gaussian mixture models to represent an underlying set of acoustic classes by individual Gaussian components, in which the spectral shape of the acoustic class is parameterized by the mean vector and the covariance matrix, is significant.

Also, these models have the ability to form a smooth approximation to the arbitrarily-shaped observation densities in the absence of other information [8]. With Gaussian mixture models, each sound is modeled as a mixture of several Gaussian clusters in the feature space. The basis for using GMM is that the distribution of feature vectors extracted from a class can be modeled by a mixture of Gaussian densities.
The motivation for using Gaussian densities as the representation of audio features is the potential of GMMs to represent an underlying set of acoustic classes by individual Gaussian components in which the spectral shape of the acoustic class is parameterized by the mean vector and the covariance matrix[11]. Also, GMMs have the ability to form a smooth approximation to the arbitrarily shaped observation densities in the absence of other information. With GMMs, each sound is modeled as a mixture of several Gaussian clusters in the feature space [12].

A variety of approaches to the problem of mixture decomposition have been proposed, many of which focus on maximum likelihood methods such as expectation maximization (EM) or maximum a posteriori estimation (MAP). Generally these methods consider separately the question of parameter estimation and system identification, that is to say a distinction is made between the determination of the number and functional form of components within a mixture and the estimation of the corresponding parameter values.

4. IMPLEMENTATION

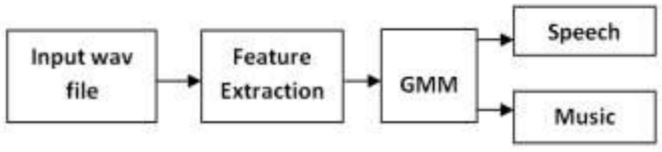

Fig -1: Block Diagram for speech/music classification.

4.1 Signal Pre-processing

Audio signal has to be preprocessed before extracting features. There is no added information in the difference of two channels that can be used for classification or segmentation. Therefore it is desirable to have a mono signal to simplify later processes. The algorithm checks the number of channels of the audio. If the signal has more than one channel, it is mixed down to mono [9]. The amplitude of the signal is then normalized to the maximum amplitude of the whole file to remove any effects the overall amplitude level might have on the feature extraction.

4.2 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.

4.3 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 mean vector is calculated for the whole audio and it is compared either to results from training data or to predefined thresholds. 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.

Gaussian mixtures for the two classes are modeled for the features extracted. For classification the feature vectors are extracted and each of the feature vectors is given as input to the GMM model. The distribution of the acoustic features is captured using GMM. We have chosen a mixture of 2, 4, 5, 10 mixture models. The class to which the audio sample belongs is decided based on the highest output.

The performance of the system for 2, 5 and 10 Gaussian mixtures is shown in Table.1. The distribution of the acoustic features is captured using GMM. The class to which the speech and music sample belongs is decided based on the highest output. Table.1 shows the performance of GMM for speech and music classification based on the number of mixtures.

Table -1: Performance of GMM for different mixtures.

GMM	2	5	10
Speech	94%	93%	94%
Music	89%	87%	87%

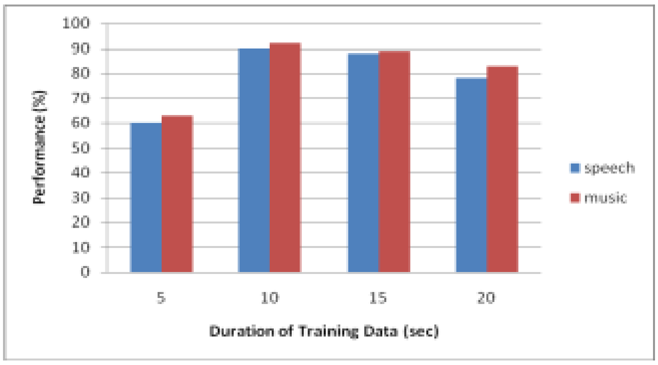

Chart -1: Performance of audio classification for different duration of speech and music clips

Audio classification using GMM gives an accuracy of 94.9%. The performance of GMM for different duration as shown in Chart 1 shows that when the mixtures were increased from 5 to 10 there was no considerable increase in the performance. With GMM, the best performance was achieved with 10 Gaussian mixtures.

5. CONCLUSION

In this paper, PLP feature vectors for the classification of speech and music files are presented. Further it is possible to improve the classification accuracy by using different types of do-main based features together. First of all, we perform feature extraction technique to extract the features from the speech and music files for classification. The proposed classification method is implemented using EM algorithm approach to fit the GMM parameters for classification. The average speech and music classification accuracy rate of the proposed method is over 94%. It shows that the proposed method can achieve better classification accuracy than other approaches. As the classification accuracy is high, this method cans retrieve a data more effectively from a large database.

REFERENCES

[1] R.A. Redner and H.F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," SIAM Review, vol. 26, pp. 195-239, 1984.

[2] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on rms and zero-crossings,.IEEE Trans. Multimedia, 7(5):155–156, February 2005.

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

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

[5] Tang, H., Chu, S. M., Hasegawa-Johnson, M. and Huang, T. S., "Partially Supervised Speaker Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 5, pp. 959-971, 2012.

[6] Chunhui Wang, Qianqian Zhu, Zhenyu Shan, Yingjie Xia and Yuncai Liu, "Fusing Heterogeneous Traffic Data by Kalman Filters and Gaussian Mixture Models," IEEE International Conference on Intelligent Transportation Systems, pp. 276-281, 2014.

[7] Sourabh Ravindran, Kristopher Schlemmer, and David V. Anderson, "A physiologi-cally inspired method for audio classification," Journal on Applied Signal Processing, vol. 9, pp. 1374–1381, 2005.

[8] Menaka Rajapakse and Lonce Wyse, "Generic audio classification using a hybrid model based on GMMs and HMMs," in IEEE Int'l Conf. Multimedia Modeling, February 2005, pp. 1550–1555.

[9] L. Rabiner and R.W. Schafer. Digital processing of speech signals. Pearson Education, 2005.

[10] 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.

[11] Sujay G Kakodkar and Samarth Borkar. Speech Emotion Recognition of Sanskrit Language using Machine Learning. International Journal of Computer Applications 179(51):23-28, June 2018.

[12] Mohammad Masoud Javidi and Nasibeh Emami. Proposing a New Method to Improve Feature Selection with Meta-Heuristic Algorithm and Chaos Theory. International Journal of Computer Applications 181(9):1-9, August 2018