ijecs open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 8 Issue 2 February 2019, Page No. 24469-24472
Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v8i2.01

Speech/music classification using PLP and SVM

R. Thiruvengatanadhan

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

Abstract:
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 Support Vector Model (SVM) model. The proposed feature extraction and classification models results in better accuracy in speech/music classification.

Keywords: Speech, Music, Feature Extraction, PLP, SVM

I. Introduction

Classification of audio signal by humans is an implicit task, but a little challenging for computer systems. The work presented in this thesis consists of three dimensional motivations. The first one is to develop a system that identifies change points which occur in audio, a second motivation intends to develop a system that distinguishes various types of audio. The final motivation is to develop a system to find a particular audio of interest from a group. For a better comprehension, it is necessary to understand the science that works behind the human listening system.

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.

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 [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.

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.

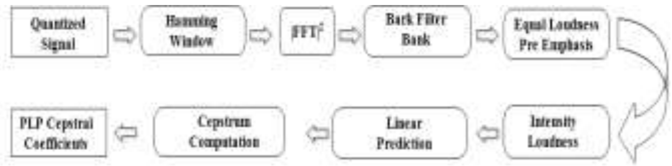

Fig. 1 PLP Parameter Computations.

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 [5]. 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.

III. Support Vector Machine

A machine learning technique which is based on the principle of structure risk minimization is support vector machines. It has numerous applications in the area of pattern recognition [6]. SVM constructs linear model based upon support vectors in order to estimate decision function. If the training data are linearly separable, then SVM finds the optimal hyper plane that separates the data without error [7].

Fig. 2 shows an example of a non-linear mapping of SVM to construct an optimal hyper plane of separation. SVM maps the input patterns through a non-linear mapping into higher dimension feature space. For linearly separable data, a linear SVM is used to classify the data sets [8]. The patterns lying on the margins which are maximized are the support vectors.

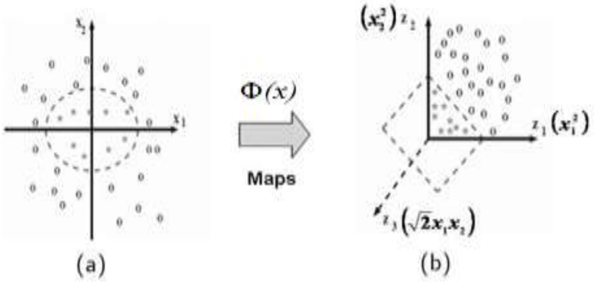

Fig. 2 Example for SVM Kernel Function Φ(x) Maps 2-Dimensional Input Space to Higher 3-Dimensional Feature Space. (a) Nonlinear Problem. (b) Linear Problem.

The support vectors are the (transformed) training patterns and are equally close to hyperplane of separation. The support vectors are the training samples that define the optimal hyperplane and are the most difficult patterns to classify [9]. Informally speaking, they are the patterns most informative of the classification task. The kernel function generates the inner products to construct machines with different types of non-linear decision surfaces in the input space [10].

IV. Results And Discussion

A. Signal Pre-processing

Audio signal has to be pre processed 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. 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.

B. *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.

C. *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. By using the feature values with appended value SVM training is carried out.

As a result of the training data two model files will be created one for speech and the other for music. The SVM trains the audio data and create two models one for speech and the other for music. 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. Each second consists of 100 frames, and each frame is assigned a class by a SVM classifier. Then, a global decision is made based on the most frequently appearing class within that second.

For classification, the audio files other than the files used for training are tested. 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 is then 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. Each second consists of 100 frames, and each frame is assigned a class by a SVM classifier. The SVM will train the audio data and create two modules correspondingly. The training samples are loaded and two classes are created, for each category. The two categories will be trained with two class 0 and class 1 with 100 examples. The testing sample is tested using the trained model and create a result. The result will show whether the audio is speech or music.

Table 1: Classification Performance for different kernel function

Kernel function	Speech	Music
Gaussian	87%	89%
Sigmoidal	85%	86%
Polynomial	84%	85%

The choice of a Kernel depends on the problem at hand because it depends on what we are trying to model. The motivation behind the choice of a particular kernel can be very intuitive and straightforward depending on what kind of information we are expecting to extract about the data. The Table.1 shows that the Gaussian kernel classification performance is greater than the other two kernels.

V. 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 PLP feature. After feature extraction classification process is done using the SVM. The SVM classifier trains the feature vectors to create models for classes. The SVM test the input audio data based on the models created by the SVM train and produce the result data. Based on the result data the input audio is classified into speech or music. SVM based speech/music classification gives a better performance of 89%.

VI. 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] 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.

[6] Chungsoo Lim Mokpo, Yeon-Woo Lee, and Joon-Hyuk Chang, “New Techniques for Improving the practicality of a SVM-Based Speech/Music Classifier,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1657-1660, 2012.

[7] Hongchen Jiang, JunmeiBai, Shuwu Zhang, and Bo Xu, “SVM-Based Audio Scene Classification,” IEEE International Conference Natural Language Processing and Knowledge Engineering, Wuhan, China, pp. 131-136, October 2005.

[8] Lim and Chang, “Enhancing Support Vector Machine-Based Speech/Music Classification using Conditional Maximum a Posteriori Criterion,” Signal Processing, IET, vol. 6, no. 4, pp. 335-340, 2012.

[9] Md. Al Mehedi Hasan and Shamim Ahmad. predSucc-Site: Lysine Succinylation Sites Prediction in Proteins by using Support Vector Machine and Resolving Data Imbalance Issue. International Journal of Computer Applications 182(15):8-13, September 2018.

[10] Hend Ab. ELLaban, A A Ewees and Elsaeed E AbdElrazek. A Real-Time System for Facial Expression Recognition using Support Vector Machines and k-Nearest Neighbor Classifier. International Journal of Computer Applications 159(8):23-29, February 2017.