Speech/Music Classification Using Spectral Features And K-Means

R. Thiruvengatanadhan
Dept of Computer Science and Engineering
Annamalai University, Annamalainagar, Tamil Nadu, India

Abstract- *The objective of a speech/music classification is to classify speech and music data using one or more acoustic characteristics associated with the signal. A speech/music classification system is developed which utilizes the Spectral Features are Zero Crossing Rate (ZCR), Short Time Energy (STE), spectral centroid, spectral flux, spectral entropy and spectral roll-off. as the acoustic feature. Multi resolution analysis is the most significant statistical way to extract the features from the input signal and in this study, a method is deployed to model the extracted wavelet feature. k-means clustering organizes the feature vectors into k number of groups. Classifying is done by minimizing the Euclidean distance between feature vector and corresponding cluster centroid.*

Keywords- Speech, Music, Feature Extraction, Zero Crossing Rate (ZCR), Short Time Energy (STE), spectral centroid, spectral flux, spectral entropy and spectral roll-off, K-means.

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. Since digitalization fosters platform independence, one can create and prototype using a digital processing platform, and then deploy on another platform [1]. Such a development platform would be for ease-of-use and testing, while the criteria for a deployment platform may be totally separate: low power, small size, high speed, low cost, etc.

Digitalized information has shown the problem of automatic audio indexing and classification as essential one for broadcasting process or analysis of stored multimedia data. Numerous researches are investigating these problems nowadays.

Firstly, the audio signal should characterize the audio signal as either one of speech, music or silence [2]. This step can employ any approach like, metric-based, model-based, decoder-guided, model-selection-based and hybrid approaches. 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. Decoder guided approach segments a speech stream into male and female clips via a gender-dependent phone recognizer. In model-selection based methods, the segmentation problem is switched to a model selection problem between two nested competing models. Recently, hybrid methods have been given much effort because it combines all the merits from different approaches for giving a better performance.

II. ACOUSTIC FEATURE EXTRACTION

Acoustic feature extraction plays an important role in constructing an audio classification system. The aim is to select features which have large between class and small within class discriminative power.

A. Zero Crossing Rate

The Zero Crossing Rate (ZCR) is a simple measure of the frequency content of a signal. For narrow band signals, the frequency content of the signal can be estimated using average ZCR [3]. However, a broad band signal such as speech, it is much less accurate. The spectral properties can be roughly estimated using short time average zero crossing rate [4]. Each pair of samples are checked to determine where zero crossings occur and then the average is computed over N consecutive samples.

B. Short Time Energy

Short Time Energy (STE) is used in different audio classification problems. STE provides a basis for distinguishing voiced speech segments from unvoiced ones in speech signal. STE is a useful feature in distinguishing high quality speech from silence [5]. The audio signal amplitude varies with time. A convenient representation that reflects the amplitude variations is known as the short time energy of the signal.

C. Spectral Centroid

A significant measure called spectral centroid is used in digital signal processing to characterize a spectrum. It indicates the "center of mass" of the spectrum and is perceptually connected with the brightness of sound. It is computed as the weighted mean of the frequencies present in the signal, which is calculated using a Fourier transform. Spectral centroid may also be used to refer to the median of the spectrum. It is similar to the difference that exists between unweighted median and mean statistics. Since both median and mean, measures the central tendency they exhibit similar characteristics. However, the audio spectral are not distributed normally and hence the values for two measures strongly defer. Grey and Gordonin [6] analyzed the difference between these two measures and suggested that mean gave a better fit than the median. Since the spectral centroid predicts the brightness of a sound digital audio music processing systems make use of spectral centroid of an acoustic measure of timbre.

D. Spectral Flux

The average variation in value of spectrum between two adjacent frames in a given audio clip is called Spectral Flux (SF). Normally, speech signal consists of alternating voiced and unvoiced sounds in the syllable rate whereas this structure does not exist in music signals. Environmental sounds have the highest variation of spectrum flux than that of a speech and music [7]. Hence, SF is the significant acoustic feature for distinguishing environmental sounds which exhibits strong periodicity. It also discriminates music, speech and environmental sounds effectively.

E. Spectral Roll Off

Spectral roll off is strongly related to spectral centroid and it represents the spectral shape of the sound signal. It defines a frequency level where more than 85% of the spectrum energy is below the typical frequency.

F. Spectral Entropy

A quantitative measure of the spectral disorder is called spectral entropy. Entropy captures the formants of distribution peaks and hence the discriminatory characteristic of this feature can be used in speech recognition, speech tracking and Voice Activity Detection (VAD) [8]. The spectrum is converted into a probability mass function by normalizing the spectrum in each subband.

III. TECHNIQUES

A. k-means

Clustering is an unsupervised learning problem which deals with finding a structure in a collection of unlabeled data [9]. It is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

K-means algorithm is one of the clustering algorithms that groups data with similar characteristics or features together. These groups of data are called clusters [10]. The data in a cluster will have similar features or characteristics which will be dissimilar from the data in other clusters. K-means clustering organizes the feature vectors into k number of groups. Grouping is done by minimizing the Euclidean distance between feature vector and corresponding cluster centroid. The K-means clustering algorithm is described below:

1. Initialize k centroids.
2. Compute the distance between each feature vector and the centroids.
3. Assign the feature vector to the centroid whose distance is minimum.
4. Re-estimate the centroids.
5. Repeat the above three steps until there is no change in centroids or for a fixed number of iterations.

IV. EXPERIMENTAL RESULTS

A. The database

The speech and music audio data are recorded various sources namely 300 clips of speech and 300 clips of music. Each clip consists of audio data ranging from one second to about ten seconds, with a sampling rate of 8 kHz, 16-bits per sample, monophonic, and 128 kbps audio bit rate.

B. Acoustic feature extraction

6 set of Spectral features are extracted from each frame of the audio by using the feature extraction techniques. The above process is continued for 600 wav files. The feature values for all the wav files will be stored separately for speech and music.

Experiments were conducted to test the performance of the system using K-means. In this work, K-means clustered gave better performance. Fig. 1 shows the performance of speech and music classification using K-means for different duration respectively.

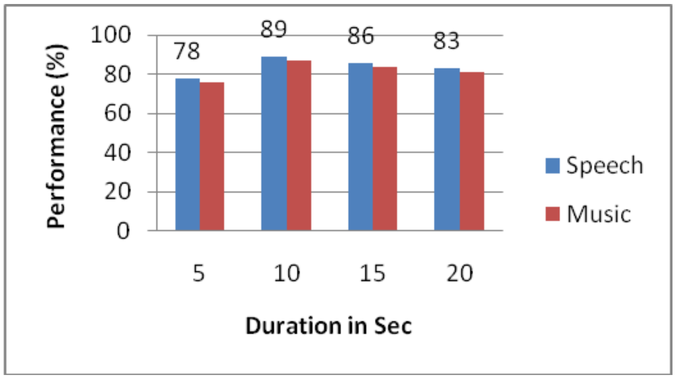

Fig. 1: Performance of audio classification for different duration of speech and music clips using K-means

V. CONCLUSIONS

In this paper, Spectral features for the classification of speech and music files are presented. Further it is possible to improve the classification accuracy by using different types of domain based features together. The proposed classification method is implemented using K-means clustering for classification. The overall accuracy of proposed method K-means using Spectral features is 89%. It shows that the proposed method can achieve better classification accuracy.

REFERENCES

[1] Aiswarya Lakshmi Thakka Ravunniy, Anaswara G K., Apoorva Eliza John, Dhanusha R., Gayathry Mohan and Sreelekshmi P S.. Call Transcription and Text Classification - A Review. International Journal of Computer Applications 179(28):9-15, March 2018.

[2] Francisco Carlos M Souza, Alinne Correa C Souza, Carolina Y V Watanabe, Patricia Pupin Mandrá and Alessandra Alaniz Macedo. An Analysis of Visual Speech Features for Recognition of Non-articulatory Sounds using Machine Learning. International Journal of Computer Applications 177(16):1-9, November 2019.

[3] C. Panagiotakis and G. Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings," IEEE Transactions on Multimedia, vol. 7, no. 1, pp. 155-156, February 2005.

[4] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multi feature Music/Speech Discriminator," International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 1331–1334, April 1997.

[5] G. Peeters, "A Large Set of Audio Features for Sound Description," Technical representation, IRCAM, 2004.

[6] Tin Lay N W E and Haizhou L I, "Broadcast News Segmentation by Audio type Analysis," IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 1065-1068, 2005.

[7] Breebaart J and McKinney M, "Features for Audio Classification," International Conference on Music Information Retrieval, 2003.

[8] Toru Taniguchi, Mikio Tohyama, and Katsuhiko Shirai, "Detection of Speech and Music Based on Spectral Tracking," Speech Communication, vol. 50, pp. 547-563, April 2008.

[9] Nurlela Pandiangan, Rahmat Gerowo and Vicensius Gunawan. K-Means Clustering and Firefly Algorithm for Shortest Route Solution based on Crime Hotspots. International Journal of Computer Applications 180(52):19-24, June 2018.

[10] Konstantin Biatov, "Audio clips retrieval using anchor reference space and latent semantic analysis," in Proc. 11th IEEE Int. Symposium on Multimedia, California,USA, December 2009, pp. 32–37.