Different Audio Feature Extraction using Segmentation

Gayatri M. Bhandari
Research Scholar
JJT University, Rajasthan

Abstract

Today's internet world consists of tremendous amount of audio and video data which is consuming more space on clouds and servers. Even streaming of such audio and video data requires lot of efforts from hardware as well as software point of view, which increase the cost of service. So today there is need to convert this audio/video into some digital form which can be easily accessible and the downloaded over the internet. Segmentation gives the best approach for dividing the multimedia data into digital data by extracting different features of multimedia data. This paper introduces new technique to segment the audio signals and extract the different features in feature vector so that this data will be reproduced after transmission over the internet. With feature extraction from audio, it is possible to recognize the content of a piece of audio. This paper explains the need for audio feature extraction system, and also describes the most important attributes such as Mel Frequency Cepstral Coefficient, Zero Crossing Ratio, Linear Predictive Coefficient, Signal to Noise Ratio, Spectrum Flux, Power Spectrum, RMS etc.

Keywords: Audio segmentation, feature extraction, FFT, Magnitude Spectrum, Power Spectrum, ZCR, SNR, Constant, MFCC etc

I. INTRODUCTION

Audio data is available at everywhere, but it is often in an unstructured order. So, it is necessary to arrange them into regularized features in order to use them more easily. It is also useful to segment an audio stream from a video clip according to audio types [1].

In many applications it is interested to segment the audio stream into homogeneous areas. Thus audio segmentation is the process of partitioning a continuous audio stream in terms of acoustically homogenous or equivalent category regions. [8]

The goal of audio segmentation is to find acoustic changes in an audio signal. This segmentation produces useful information such as division into speaker signals and speaker identities, allowing for automatic indexing and information retrieval of all occurrences of a particular speaker. If we collect together all segments produced by the same speaker we can perform automatic online speech recognition acoustic models to improve overall system performance [2].

II. SEGMENTATION APPROACHES

There are segmentation methods which are categorized into three groups, viz energy-based, metric-based, and model-based. The energy-based algorithm uses sound power in time domain. On the other hand, both the metric-based and the model-based method are based on statistical models. There are different types of segmentation approaches i) Energy Based ii) Model Based and iii) Metric Based.

Fig 1 shows the basic architecture of feature extraction. It simply shown how features are separated into Speech and Non-speech features. According to these here six different classes of audio are defined.

A. Speech:

It is the pure speech recorded in the studio without background music.

B. Speech over music:

It includes all studio speech with music in the background which may add instrumental audio also.

C. Telephone speech:

In this some part of the conversation have telephonic speech from the viewers. These interventions are get mixed in the program's main audio as a wide band stream.

D. Telephone speech:

The same as previous class but additionally there is music in the background.

E. Music:

It is the pure music which is recorded in the studio without any speech.

F. Silence:
No Sound

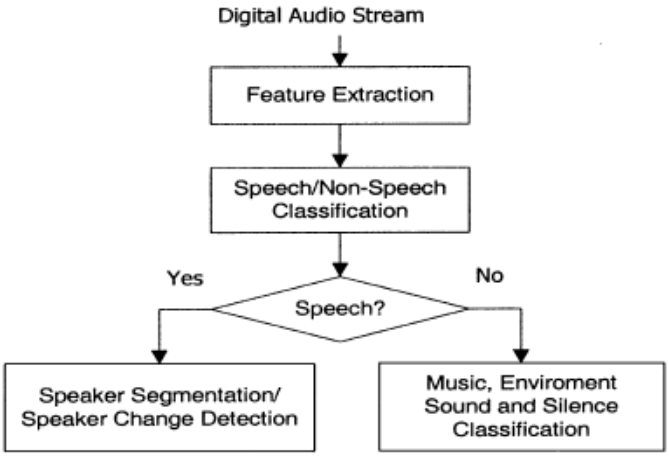

Fig. 1: Basic Architecture of Feature Extraction

III. FEATURES OF AUDIO

A. Magnitude Spectrum:
The magnitude spectrum is calculated by initially calculating the Fast Fourier Transform (FFT) with a Hamming code window. The magnitude spectrum value for each bin is calculated by first adding the squares of the real and imaginary parts of audio components. The square root of this is then found and the result is divided by the number of bins [20].

B. FFT Bin Frequencies:
This feature is calculated using bin labels, in Hz, which is used in power spectrum or magnitude spectrum which would be calculated by the FFT of a window of the size of that provided to the feature extractor. Even though it is not a useful feature for classification, it can be used for calculating other depended features [15].

C. Power Spectrum:
The power spectrum of a signal provides the distribution of the audio signal power among various frequencies. It is the Fourier transform of the correlation function, and generates the information on the correlation structure of the signal. [15]

D. Spectral Centroid:
This feature is a measure of the "centre of mass" of the power spectrum and is found by calculating the mean bin of the power spectrum. The result returned is a number from 0 to 1 which represents a fraction of the total number of bins. [15,23]

E. Spectral Flux:
Spectral flux is evaluated by calculating the difference between the current values of each magnitude spectrum bin in the current window and previous window. These differences are squared, and then added to get the result. [22,23]

F. Compactness:
This feature represents how important a regular beat is played in a piece of music. This is calculated by finding the sum of all values in the beat histogram [22].

G. ZCR:
This feature is calculated as a measure of the pitch and noisiness sound of a signal. It is also calculated by finding the number of times the signal changes sign from one sample to another (or touches the zero axis) [23].

H. Bit Histogram:
This is the histogram which shows the strength of different rhythmic periodicities in a signal. This is calculated by taking the Root Mean Square of 256 windows and then taking the FFT of the result [23].

I. RMS:

It extracts the Root Mean Square (RMS) from a set of samples. RMS is calculated by adding the squares of each sample, dividing this by the total number of samples in the window, and finding the square root of the result. [23]

J. MFCC:

Mel-frequency Cepstral coefficients (MFCCs) are the signal coefficients that are collected to forms MFC. The Mel-frequency cepstrum different from cepstrum in the frequency bands which are equally divided on the Mel scale. MFCCs are rigorously used as features in speech recognition systems.

K. Relative Difference Function:

This feature evaluates the log of the derivative of the RMS. This is useful for onset detection. [20]

L. Bit Histogram:

Normally histogram shows the strength of different rhythmic periodicities in a signal. This is calculated by taking the Root Mean Square of 256 windows and then taking the FFT of the result. [8]

IV. BASIC SEGMENTATION ARCHITECTURE

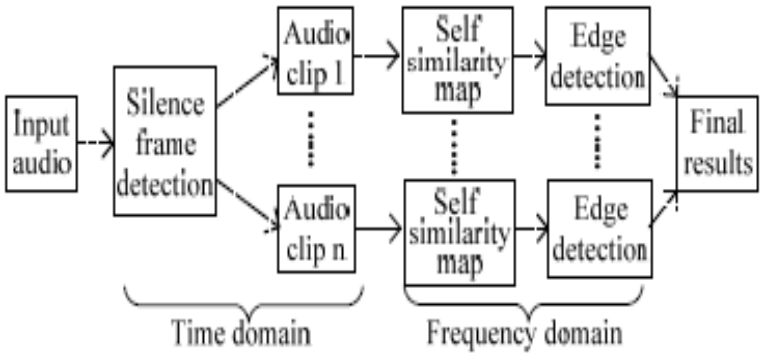

Fig. 2: Frame work of audio segmentation

Figure 2 shows the basic architecture of the proposed approach that performs audio segmentation and speaker segmentation. Non speech segments are again categorized into music, environmental signal and silence [2].

The proposed algorithm for audio segmentation segments the audio into different parameters as described before also algorithm used for feature extraction divides the different audio features such as MFCC, LPC, SF, SNR, Power spectrum, Magnitude spectrum, Relative Difference Function, ZCR etc.

Figure 3 shows the flowchart of proposed audio segmentation and classification algorithm. It is a hierarchical structure. In the initial part of this level, a long audio stream can be segmented into some audio clips according to the change of background sound in histogram modeling. After feature extraction, the input digital audio stream is classified into speech and non-speech. Non speech segments are further classified into music, environmental signals and silence signal, while speech segments are further segmented by speaker identity

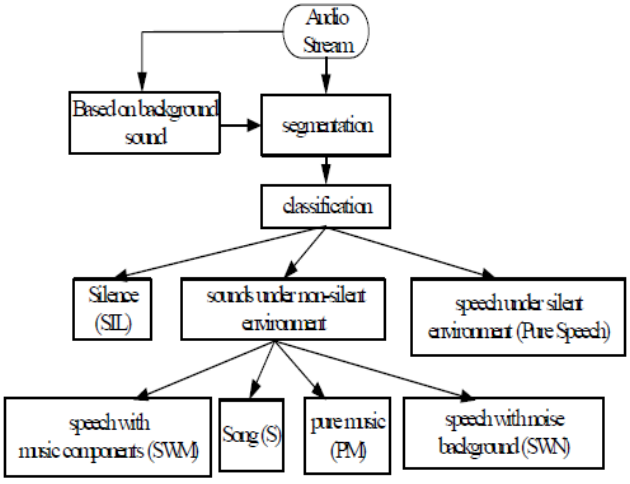

Fig. 3: The flowchart of segmentation and classification algorithm [24]

In this algorithm the audio into different parameters as described before also feature extraction algorithm separates out the different audio features [14]

Usually a chosen set of features is extracted from each frame of the audio. The features are then often normalized by the calculated mean value and the standard variation over a larger time unit and then stored in a feature vector. [17]

Features are used in two ways, either by using the extracted value or by using changes over time. When using the changes over time it is possible to calculate statistical features like variance and standard deviation. In [5] it is shown how using changes over time are more accurate than only using the absolute values of the features. Only one feature is used in [1], [4] and [7], although they all use advanced special features described earlier. Others have chosen to have a set of standard features. In [3] five different features are used, energy, ZCR, Spectral Entropy and the two first MFCCs. RMS and ZCR are used in [2].

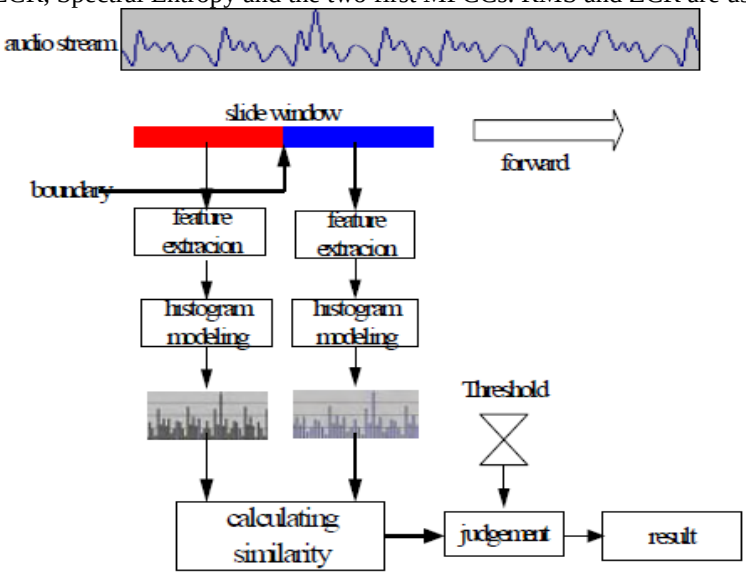

Fig. 4: Segmentation Algorithm [25]

Fig 4 shows the details about how the segmentation and feature extraction will be implemented. Initially open any kind of audio data file in buffer and determine the frame size using sampling rate. Apply windowing techniques for filtering data with window size of 1024 bytes in overlapping method. Apply formulae for FFT, LPC, ZCR, Magnitude spectrum, Power spectrum, Spectral Centroid [18].

V. RESULTS

Table - 1
Different Audio Features from Samples

MS	*PS*	*LPC*	*FFT Bin Frequencies*	*ConstantQ*	*RMS*	*Compactness*
0.042	*0.036*	*-0.035*	*2000*	*0.0001*	*0.29*	*5.96*
0.027	*0.015*	*-0.046*	*2000*	*0.0002*	*0.16*	*1.34*
0.047	*0.037*	*-0.035*	*2000*	*0.0001*	*0.21*	*4.28*
0.034	*0.016*	*0.0005*	*2000*	*-0.0008*	*0.01*	*4.9*
0.029	*0.027*	*-0.037*	*2000*	*-0.0008*	*0.18*	*-4.8*
0.003	*0.001*	*0.005*	*2000*	*-0.0008*	*0.10*	*3.14*

Table - 2
Result of First Level of Classification

Audio type	*Accuracy*	*Recall*	*Precision*
Pure Speech (PS)	*85.15%*	*85.63%*	*87.62%*
Silence (SIL)	*97.10%*	*86.14%*	*91.29%*
Others	*77.95%*	*95.08%*	*85.67%*
Pure Speech (PS)	*91.33%*	*93.65%*	*92.47%*
Silence (SIL)	*98.22%*	*92.97%*	*95.52%*
Others	*85.68%*	*95.45%*	*90.3%*

The above results are calculated by applying some audio samples on the code implemented on java and the test results are generated mentioned in above table. From this it has been observed that FFT Bin Frequencies parameter remains constant for any audio sample applied for testing and other parameters are generated. Also this is tested for any kind of audio file formats such as.avi, mp3, mp4, voc, .wav etc.

VI. CONCLUSION

In above method, we have presented comparative analysis of on feature extraction using segmentation techniques. Different parameters such as audio type, accuracy, recall factor and precision have been observed for pure speech, silence etc at primary level.

The above classification can be extended for other feature such as Spectrum, Spectral Centroid, MFCC, LPC, ZCR, SNR, Moments, Beat Histogram, Beat Sum, RMS etc in order to precisely segment all these features has been calculated and its comparison is made. The above algorithms can be to all types of audio files such as.wav, pcm, .avi , .mp3 etc also in future from video clip an audio can be segmented and features can be calculated.

ACKNOWLEDGMENT

The completion of this paper would not have been possible without the help and guidance of Dr. R. S. Kawitkar and MANAGEMENT of JSPM Engineering College whose support was always there with us to correct at every step. We would like to thank my family members and seniors for their motivation and encouragement. Last but definitely not the least we would thank the almighty god without whose grace this paper would not have achieved success.

REFERENCES

[1] Ewald Peiszer, Thomas Lidy, Andreas Rauber,: "Automatic Audio Segmentation: Segment Boundary and Structure Detection in Popular Music", 2008.
[2] George Tzanetakis Perry Cook: "Multifeature Audio Segmentation for Browsing and Annotation", Proc.1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, pp W99-1 – W99-4
[3] Guojun Lu:" Indexing and Retrieval of Audio: A Survey", pp 269-290, 2001I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III,
[4] Jessie Xin Zhang1, Jacqueline Whalley1, Stephen Brooks: "A Two Phase Method for general audio segmentation"
[5] Jonthan Foote: "Automatic Audio Segmentation Using A Measure of Audio Novelty"
[6] Julien P., Jose´ A. and Re´gine A.:" Audio classification by search of primary components", PP 1-12
[7] Lie Lu, Hong-Jiang Zhang and HaoJiang,:"Content based Analysis for Audio Classification and Segmentation",IEEE Transaction on Speech and Audio Processing, pp 504-516, ,2002.
[8] Lie Lu, Stan Z. Li and Hong-Jiang Zhang:,"Content based audio segmentation using Support Vector Machines", 2008.
[9] Lateu Aguilo, Taras Butko, Andrey Temko, Climent Nadeu:" A Hierarchical Architecture for Audio Segmentation in a Broadcast News Task" ,PP 17-20, 2009.
[10] Mauro Cettolo , Michele Vescovi , Romeo Rizzi, "Evaluation of BIC-based algorithms for audio segmentation" Elsevier pp 147–170,2005
[11] Michael M. Goodwin and Jean Laroche, "Audio Segmentation by feature space clustering using linear discriminant analysis and dynamic programming", 2003
[12] Mohammad A. Haque& Jong-Myon Kim, "An analysis of content-based classification of audio signals using a fuzzy c- means algorithm", 2012
[13] Nima Mesgarani, Malcolm Slaney, and Shihab A. Shamma, "Discrimination of Speech from Nonspeech".
[14] Panagiotis S., Vasileios M., Ioannis K., Hugo M., Miguel B., Isabel T., "On the use of audio events for improving video scene segmentation."
[15] P. Krishnamoorthy · Sarvesh Kumar, "Hierarchical audio content classification system using an optimal feature selection algorithm" , pp 415–444, 2010.
[16] [16] Samer Abdallah· Mark Sandler ·Christophe Rhodes, Michael Casey, "Using duration Models to reduce fragmentation in audio segmentation, pp 65:485–515, 2006.
[17] Shih-Sian Cheng, Hsin-Min Wang, and Hsin-Chia Fu, "BIC-BASED Audio Segmentation by divide and conquer".
[18] Shi Yong, "Audio Segmentation", PP 1-4, 2007.
[19] ShoichiMatsunaga, Osamu Mizuno, Katsutoshi Ohtsuki, and Yoshihiko Hayashi, "Audio source segmentation using spectral correlation features for automatic indexing of broadcast news ", PP 2103 to 2106
[20] Theodoros Giannakopoulos,AggelosPikrakis , Sergios Theodoridis , "A Novel Efficient Approach for Audio Segmentation ", 2008
[21] Yibin Zhang and Jie Zhou,:" Audio Segmentation based on Multiscale audio classification" PP IV – 349 to IV – 352, 2004.
[22] YuxinPeng, Chong-Wah Ngo, Cuihua Fang, Xiaoou Chen, and Jianguo Xiao,:"Audio Similarity Measure by Graph modeling and Matching", pp603-606.
[23] Zaid Harchaoui, F´elicienVallet, Alexandre Lung-Yut-Fong, Olivier Cap,"Regularized Kernel-Based ApproachToUnsupervised Audio Segmentation"
[24] G.M.Bhandari , Dr. R.S.Kawitkar , "Audio Segmentation for speech recognition using feature extraction", IJCTA , Vol 4 (2),182-186 ISSN:2229-6093
[25] G.M.Bhandari , Dr. R.S.Kawitkar , "Audio Segmentation for speech recognition using feature extraction", CSI 2013 Vol 2 Springer , ISBN 978-3-319-03095 pp 209- 217
[26] G.M.Bhandari , Dr. R.S.Kawitkar , "Audio Segmentation for speech recognition using feature extraction",IPASJ , Volume 2, Issue 3, March 2014, ISSN 2321-59923