A Survey on Audio Segmentation and Classification Methods

Manasi Choche
PG Student
Department of Computer Engineering
Pillai Institute of Information Technology, Engineering, Media Studies and Research, Panvel, India

Dr. Satishkumar Varma
Associate Professor
Department of Information Technology
Pillai Institute of Information Technology, Engineering, Media Studies and Research, Panvel, India

Abstract

An audio is a sound within audible range available to human. This range is study of all mechanical waves in gases, liquids and solids including vibration, and sound. Audio segmentation is most important preprocessing step in several audio processing applications with a significant impact e.g. on speech recognition system. This paper introduces a framework which analyzes the advantages of different segmentation and classification methods. The different audio segmentation methods like Bayesian Information Criterion (BIC), two phase audio segmentation method, metric-based method are studied. Audio segmentation plays important role in audio classification. After segmenting audio it is classified into different classes. The classification model consists of rule-based and model based.

Keywords: Audio Segmentation, Audio Classification, BIC, HMM, Metric-Based, SVM

I. INTRODUCTION

Segmentation is the process of splitting (segmenting) the entire thing into smaller groups. Segmentation is process of partitioning a heterogeneous audio input stream into its composite segments, known as audio clips. The implementation of this segmentation method is the first module of the audio visualization system. The module separates an input audio stream into individual sounds (sound clips) and indicates the silence periods between the sound clips (which are not involved in the subsequent classification step).

Segmentation plays an important role in audio processing applications such as content- based audio retrieval and recognition, audio classification, and audio database management. For general mixed-sound audio files, segmentation is an especially important step because it is much easier and more accurate to undertake further analysis, such as query or classification, using segmented homogeneous audio clips than using a raw audio input stream. Audio segmentation plays important role in audio classification. In which audio data is classified into different classes. There are different audio classification methods which are rule-based and model based.

Audio segmentation methods are divided into audio classification and change detection. Change detection methods recognize homogeneous audio segments by performing different algorithms like event detection. Audio classification module classifies audio data into several classes. And it is implemented with the help of different classification methods. The process of audio segmentation and classification modules is as shown in fig.1,

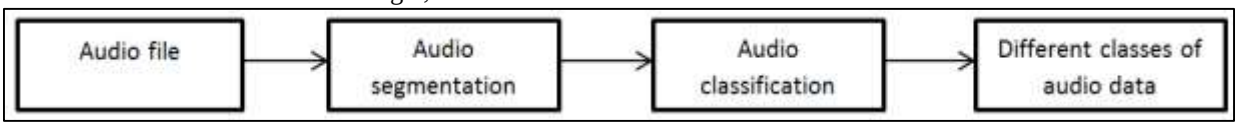

Fig. 1: The process of audio segmentation and classification modules

II. AUDIO SEGMENTATION PROCESS

An audio segmentation process [1] takes an audio clip as an input and segments it into sub-segments (Ex. Silence, non-silence) as sown in fig. 2. This process consists of following two steps,

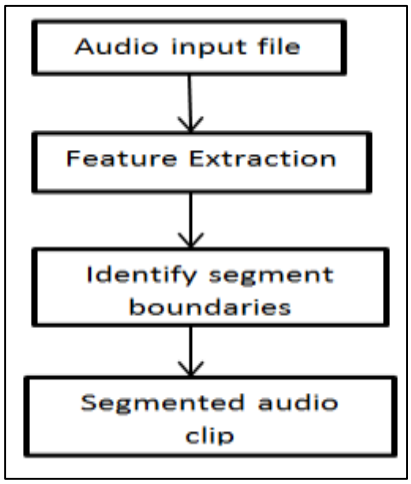

Fig. 2: Audio segmentation process

A. *Feature Extraction*

An audio data contains mixed sound. In this audio data is extracted into different audio features. So that audio data contains single type of sound.

1) *Time Domain*

The first step of audio segmentation discovers the start and end of audio data based on silence. Audio features like Spectral Power (SP) and Root Mean Square (RMS) are used for silence detection.

2) *Frequency Domain*

If there is any minor change in features after getting results of audio data from first step, some process is needed. So that segmentation is carried out and it is used to identify that changes.

B. *Identify Segment Boundaries*

It is an easy task to find the segment boundaries. It is identified by determining the points which are above threshold. The approach used here is to locate maximum wherever the score exceeds the threshold for maximum time precision. The score is normalized such that maximum score is one and minimum is zero. A useful way of organizing the index points is in a binary tree structure constructed by ranking all index points by score.

III. AUDIO SEGMENTATION METHODS

There are several methods of audio segmentation as discussed above. Here those methods are discussed in detail,

A. *Bayesian Information Criterion (BIC) method*

The Bayesian Information Criterion (BIC) is a widely used method for audio segmentation, and there are different algorithms for audio segmentation. BIC is applied within a sliding variable-size analysis window where single changes of the audio are identified [2].

The main steps of this algorithm are:

1) Search start

BIC values are computed only for the first N_{min} observations. N_{min} is the minimum size of the window. Values are computed with low resolution, e.g., dl ¼ 30. In order to have enough observations for computing both R1 and R2, BIC are not computed for the N_{margin} indexes close to the left and right boundaries of the window.

2) Window growth

The window is enlarged by including N_{grow} input observations until a change is detected, or a maximum size N_{max} is reached.

3) Window shift

The N_{max}-sized window is shifted forward by N_{shift} observations.

4) Change confirmation

If in one of the three previous steps a change is detected, DBIC values are re-computed with the high resolution, e.g., dh _ dl=5, centering the window at the hypothesized change. The current size of the window is kept, unless it is larger than N_{second} observations, in which case it is narrowed to that value. If a change is detected again, it is output by the algorithm.

5) Window reset

After the change confirmation step, the algorithm has to go on resizing the analysis window to the minimum value N_{min} and locating it in a position dependent on the result of the confirmation step.

B. *Metric based method*

In metric-based method, various acoustic distance measures have been defined to evaluate the similarity between two adjacent windows shifted along the audio stream to form a distance curve [3]. This distance curve was often low-pass filtered and the locations of peaks were chosen to be acoustic changing points by heuristic thresholds. Most of the distance measure criterions come from the statistical modeling framework. The metric based methods have the advantage of low computation cost and, thus, are suitable for real time applications. The metric based method with the help of BIC is proposed as follows,

1) Detecting the first changing point

Here, the first short window (typically 2 seconds) of the audio stream as a template is used, and the *BIC* value between the template and a sliding window with the same size as the template is calculated.

2) Detecting multiple changing points

Based on the above first changing point detection algorithm, a procedure that can sequentially detect the multiple changing points in an audio stream is designed. To speed up the detection process, the first changing point detection algorithm in a fixed-size window (12 seconds in this study) is applied. If there is no changing point detected, the window will be shifted by 2 seconds. Otherwise, the window will be shifted to the location of the changing point. The details of an algorithm are described as follows:

1) Set the initial window; i.e. a=0,b=12, W = [a,b];
2) detect the first changing point in W by using a template with a length of 2 seconds.
3) if no changing point is detected in W detect the first changing point in W by using a template with a length of 3 seconds. End if no changing point is detected in W shift the window by 2 seconds.

i.e. a = a + 2;b = a + 12;W = [a,b];
else
let *t* be the detected changing point in W, shifts the
window to *t*.
i.e. a = *t*; b = a+12;W = [a,b];
end
4) Again go to step 2.

C. *Two Phase Method*

This section presents a two-phase CIS method for general audio files [4]. It can detect audio clip boundaries from the information contained within the audio file itself. Its implementation and audio feature selection abilities are discussed in each of the corresponding procedures within the method. It consists of three phases silence detection, mapping self-similarity, and edge detection as shown in fig. 3,

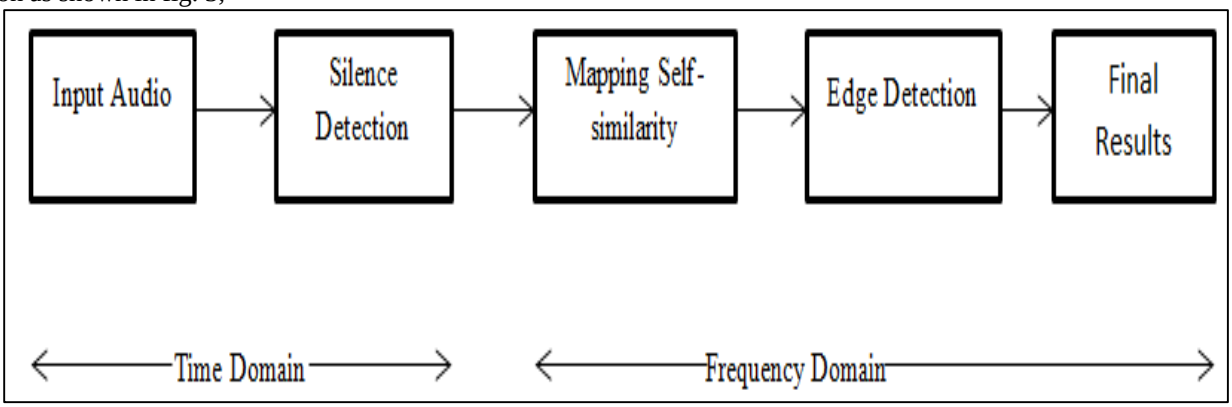

Fig. 3: Framework for two phase audio segmentation method

1) Silence Detection

General audio files may have a short period of silence between sounds where the sound clips do not overlap. Therefore silence detection can often be used to separate sounds. The first phase of the two-phase segmentation finds the start and end of audio clips based on silence periods. Root Mean Square (RMS) [60] and Spectral Power (SP) are widely used audio features for silence detection. If the selected audio feature of a frame is less than a pre-determined threshold, it is regarded as a silence frame. But neither RMS nor SP is suitable when the sound's duration is very short, say less than 10ms, e.g. hand claps.

2) Edge Detection

For any audio clips detected in the first phase, if two sounds are seamlessly connected with no or very short silence between them, such as audio clip seven in the telephone/violin audio example, silence detection does not work. When the presence of silence cannot be totally relied on, a subsequent, more sophisticated refinement phase is required for the separation of contiguous clips. Because a sound's audio features tend to be homogeneous, any abrupt changes in audio features may indicate the start of a new sound clip.

Table-1 shows the average accuracy of different segmentation methods and Table-2 shows comparative study,

Table-1
Accuracy of different segmentation methods

Group	*Average accuracy*
2-phase Euclidean	*91.9%*
2-phase Cosine	*89.1%*
BIC	*89.6%*
Metric Based	*87.6%*
Silence detection	*81.0%*

Table-2
Comparative Study of segmentation methods

Method	*Description*
Method-1 BIC	*-BIC uses three approaches in which last approach which is Cumulative Sum approach is more efficient than other two.* *-Third approach gives low cost and requires less execution time.*
Method-2 Metric-based	*-It has advantage of low computation cost and high accuracy.* *-By introducing the BIC into distance computation of two windows, we can deal with thresholding issue more easily.*
Method-3 Two phase	*-It provides good solution to audio segmentation problems.* *-This method is more efficient than other methods.* *-It gives better average accuracy.*

IV. AUDIO CLASSIFICATION METHODS

The rule based and model-based are the two types of approaches used in classification. Sometimes, rule based method fails to define all rules if the given dataset is incomplete. Hence, rule based method cannot give better result for incomplete dataset and cannot produce any class. There are various model-based methods such as Neural Network (NN), Hidden Markov Model (HMM), Support Vector Machine (SVM), and Gaussian Mixture Model (GMM) are used for classifying audio segments.

A. Hidden Markov Model

HMM model processes with time varying characteristics. K-means algorithm for clustering [6] is used to estimate the values. In this algorithm the center of different clusters are initialized first. The nearer cluster to every data is determined. And then the location of every cluster is set to the mean of data points which belongs to same cluster.

B. Support Vector Machine

SVM is a classifier consists of optimal hyper-plane also known as binary classification model [5]. In this model, the waveform of audio file is converted into individual frame segments. Every frame is then changed into features in terms of vectors. These feature vectors are given to the SVM classification model. Every frame is labeled as +1 or -1 based on decision function. And then the sum of all labels or tags is calculated. If the of audio data is greater than zero, then it is classified as +1 class. And if it is less than zero, then it is classified as -1 class.

V. CONCLUSION AND FUTURE SCOPE

In audio segmentation, audio is segmented into silence and non-silence segments using silence detection on extracted data. Three segmentation methods are studied. The accuracy of different audio segmentation methods is evaluated. Two phase segmentation method gives high accuracy than other methods. After segmentation audio data is classified into different classes. There are two classifiers HMM and SVM for classifying an audio data. The HMM is preferred in MPEG-7 audio classification. The performance of classification methods is analyzed with the help of the combinations of these features. It is used in indexing of video, retrieval of information, and content analysis.

REFERENCES

[1] S. Nomura, T. Shiose, H. Kawakami, O. Katai and K. Yamanaka, "A Novel on Sound Visualization", Process in Virtual 3D Space: the Human Auditory Perception Analysis by Ecological Psychology Approach, Proceedings in 8th Asia Pacific Symp. Intelligent and Evolutionary Systems, Cairns, Australia, pp.137-149, 2004.
[2] Mauro Cettoloa, Michele Vescovib, Romeo Rizzib, "Evaluation of BIC-based algorithms for audio segmentation", Universita degli Studi di Trento, Facolta di Scienze MM.FF.NN. I-38050 Povo, Trento, Italy, 2005.
[3] Shi-sian Cheng and Hsin-min Wang, "A Sequential Metric-based Audio Segmentation Method via The Bayesian Information Criterion", Institute of Information Science, Academia Sinica Taipei, Taiwan, Republic of China, 2008.
[4] Jessie Xin Zhang1, Jacqueline Whalley1, Stephen Brooks2, "A Two Phase Method For General Audio Segmentation", School of Computing and Mathematical Sciences, Auckland University of Technology, Auckland, pp. 626-629, 2009.
[5] Lie Lu, Stan Z. Li and Hong-Jiang Zhang, "Content-based Audio Segmentation Using Support Vector Machines", Microsoft Research China, pp. 749 – 752, 2003.
[6] L. R. Rabiner, "A tutorial on hidden markov models," Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.