IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 3, June-July, 2014
ISSN: 2320 – 8791 (Impact Factor: 1.479)
www.ijreat.org

A New Approach for Automatic Audio Segmentation, And Reconstruction

Mr. Vinayak D. Chavan[1], Dr. Sanjay L. Nalbalwar[2]

[1]M.Tech student, [2]Professor & Head
[1,2]Dept. of Electronics & Telecommunication Engg., Dr. B. A. T. U. Lonere, M.S., India

Abstract

Automatic Audio Segmentation aims at extracting information about type of audio i.e. silence, clean speech or speech with noise, music etc. The aim of this thesis is to find and extract different features of audio, segment the audio and combine them to form single type of audio. In this work, the audio signal is initially decomposed into non-overlapping frames. Then these frames are decomposed into a set of band-limited functions known as Intrinsic Mode Functions (IMF) using Empirical Mode Decomposition (EMD). Temporal and Spectral features then extracted from these IMFs and thereafter classification is done using k-nearest neighbour (kNN) classifier. Depending upon classification accuracy of the features the segmentation is done using k-means clustering algorithm. Then the segments were combined to form new audio signal.

KEYWORDS--*Intrinsic Mode Function, Empirical Mode Decomposition, Spectral And Temporal Features, k-NN classifier, K Means Clustering.*

1. Introduction

In the past decade a huge amount of multimedia data in the form of text, images, audio and video has become available. With the increasing use of such audio data and challenges faced in different multimedia application, it has become essential to put effort into audio signal analysis. An audio signal segmentation system should be able to identify different segments and categorize them according to different audio types i.e. silence, clean speech, speech with noise etc. Before segmentation it is important to identify the features which can discriminate between different segments. So, first we classified audio signal into clean speech, noisy speech and music.

In this paper, we have proposed a method for audio segment identification as clean speech, speech with noise and music which relies on temporal and spectral shape features. Different background noise sources classified in [1,2,3,4] by using spectral features like spectral centroid, spectral roll-off, sub-band energy ratio and zero-crossing rate as temporal feature. In this experiment, the input signal is hierarchically decomposed by empirical mode decomposition (EMD) [6], [7]. A given signal is decomposed into a number of Intrinsic Mode Functions (IMFs) and a residual. The features are computed from IMFs and the residue. The performance of various features extracted from the IMFs is evaluated using K nearest neighbour classifier. Using these feature the audio is segmented according to its type using k means clustering and then the segments are joined together to reconstruct the clean speech signal and music signal.

We organize the remainder of this paper as follows: In section 2 overview of EMD is presented. Proposed segmentation scheme is presented in section 3. Section 4 describes the commonly used features for classification and the basic concepts of k nearest neighbour and k means clustering classifier. The database used for the experiment and experimental results is presented in Section 5. We conclude in Section 6 with conclusions and future work in the field.

2. Empirical mode decomposition

Empirical Mode Decomposition (EMD) is one of the best methods of feature extraction. It is advantageous compared to the other methods as EMD is an adaptive data analysis method that is based on local characteristics of the data, and hence, it catches nonlinear, non-stationary oscillations more effectively. EMD method is able to decompose a complex signal into a series of intrinsic mode functions (IMF) and a residue. [6], [7].

2.1 Intrinsic Mode Function

EMD decomposes the original signal into a definable set of adaptive basis of functions called the intrinsic mode functions. Each IMF must satisfy two basic conditions: i) In the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one. ii) At any point, the mean value of the envelope, one defined by the local maxima and the other by the local minima is zero.

2.1.1. Sifting process

The purpose of sifting is to subtract the large-scale features of the signal repeatedly until only the fine-scale features remain. First, the original signal, x(t), should be enclosed by the upper and lower envelope in the time domain. Using a cubic spline, the local maxima is connected forming the upper envelope $u_+(t)$ and the local minima is connected forming the lower envelope $l_-(t)$. The two envelopes cover all the data points. The envelope mean m(t) is determined as follows,

$$m(t) = (u_+(t) + l_-(t))/2 \quad (1)$$

The first component is described as,

$$h1(t)=x(t) - m(t) \quad (2)$$

The component h1(t) is now examined to see if it satisfies the conditions to be an IMF. If h1(t) doesn't satisfy the conditions, h1(t) is regarded as the original data, the sifting process would repeat, obtaining the mean of the upper and lower envelopes, which is designated as m_{11}; therefore,

$$h_{11}(t) = h_1(t) – m_{11}(t) \quad (3)$$

Then, repeat the procedure until $_{h1k}$ is an IMF,

$$\text{i.e.,} \quad h_{1k}(t)=h_{1(k-1)}(t)–m_{1k}(t) \quad (4)$$

After k siftings, we explained the first intrinsic mode to be,

$$C_1 = h_{1k} \quad (5)$$

Finally, C_1 revealed the higher frequency component of IMF. To obtain enough physical definitions of IMF, the sifting stop criteria, are of great importance found by finding SD. The typical values of SD are 0.2 and 0.3. To obtain the second and subsequent intrinsic mode functions, the residual signal can be calculated as

$$x(t) – C_1(t) = r_1(t) \quad (6)$$

r_1 considers the original data, and by repeating the above procedures, x(t) could be obtained by the second IMF component C_2. The procedure as described above is repeated for n times, then the n-IMFs of signal x(t) could be obtained

$$r_{n-1}(t) – c_n(t) = r_n(t) \quad (7)$$

The decomposition procedure can be stopped when the residue, r_n, becomes a constant, a monotonic function, or a function containing only a single extrema, from which no more IMF can be extracted. By summing Eqns. and the original signal can be reconstructed by adding all IMFs and residue.

3. Proposed Audio Segmentation Scheme

This thesis is divided into two stages
i) Classification: Identifying the featues, that can discriminate between different audio categories
ii) Segmentation: using these features for segmentation and reconstruction

3.1 Classification

The aim of this stage is to identify the capability of features to discriminate between the given audio classes. In [3] four types of noise sources are classified whose accuracy is 75-85% for first three IMFs. In our method, audio signal is first segmented into overlapping frames and each frame is decomposed into IMFs which are used to extract the audio features. From these features, feature vector is computed.

$$Fv, m = [F_{1,m} \ F_{2,m} \ F_{3,m} \ldots . \ F_{r,m}] \quad (8)$$

Where 'v' is signal no., 'm' is IMF no. and 'r' is feature no. in this way, for each IMF feature vector is computed and applied to k-NN classifier for classification. The classification results showed that the accuracy of 100% can be achieved using different feature combinations for first IMF.

2.2 Segmentation

In this phase, initially, silence part is removed from the audio and then it is decomposed into non-overlapping frames. These frames are decomposed into IMFs and selected features are computed from the IMFs. The features determined in classification stage, which are having good discrimination accuracy, are used to form the feature vector. These feature vectors are then applied to k means clustering classifier for identification and reconstruction of the audio segments.

4. Features and Classifiers

Features extraction is the main step for classifying any audio signal into a given class [5]. These features will decide the class of the signal. Feature extraction involves the analysis of the input signal. The feature extraction techniques can be classified as temporal analysis and spectral analysis technique. Temporal features include Short Time Autocorrelation Function (ACF), Short Time Energy (STE), and Zero Crossing Rate (ZCR). Spectral features include Spectral Centroid (SC), Spectral Roll off (SR) and Spectral Flux (SF) [9]. Using these features, feature vector is formed which is then applied to the

classifier for classification. In our experiment we used k nn classifier, as it's a simple one.

.

3.1 K- nearest neighbour classifier

An instance based learning method called the K-Nearest Neighbor or K-NN algorithm has been used in many applications in areas such as data mining, statistical pattern recognition, image processing. k-nn is supervised classifier means it used the training data during testing phase. In the classification scheme, the n feature values of a sample are taken to represent a location in an n-dimensional feature space. The distance between a new data point and every training data point is calculated and a vote is taken among the k closest training data points (k neighbors) to determine the classification of the new data sample.

3.2 K-means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid.

5. Experiments and Results

In classification stage of our experiment, we have classified audio signal into three major classes: speech, speech with background noise and music. For this experiment we selected dataset consisting of 30 clean speech signals consisting of male and female voices, 30 noisy speech signals and 30 music signals of 2-5 sec duration long. The noisy speech consists of one type of noise mixed with speech with different utterances. The given signal is first segmented into frames of 50 ms duration and 25 ms overlap between each frame. Then each frame is decomposed into IMFs using Empirical Mode Decomposition. These IMFs are then used for extracting different features as mentioned in section 3. If there is 'N' no. of frames for one signal then mean of particular feature of all the frames is considered as one of the feature for that signal.

$$Fr = (1/N)\sum_{i=1}^{N} fr(i) \qquad (9)$$

where 'r' denotes feature no. and 'i' denotes the frame no. of one particular signal. From these features we formed the feature vector which is then applied for classification.

$$Fv, m = [F_{1,1}\ \ F_{2,1}\ \ F_{3,1} \ldots .\ F_{r,m}] \qquad (10)$$

Where 'v' is signal no., 'm' in IMF no. and 'r' is feature. Feature vectors are formed by combining different set of features for each audio category and then Classification is performed using k-NN classifier with k=3. Results of classification which are having very good accuracy are presented below.

Table -1: Classification accuracy for audio classes IMF1 (%)

FEATURES	SPEECH	SPECH + NOISE	MUSIC	OVERALL ACCURACY
SR,SC,SF	100	94	100	98
ZCR,STE	100	100	100	100
ZCR,SR, SC,SF	100	100	100	100

From the above results it is clear that, different feature combinations can achieve the accuracy up to 100%. Also it is observed that first IMF can discriminate between the specified categories with good accuracy. Other IMFs also tested for classification but the accuracy is less than the accuracy of IMF1. Hence results for other IMFs are not mentioned.

In segmentation stage, we have combined known audio signals to form a complete audio of duration 1min. So that performance of the segmentation can be easily evaluated. First of all, the silence period is removed from the audio. For this the audio signal is divided into frames of 8ms duration and absolute value is computed for each frame. The frames having maximum value less than 0.001 are considered as silence part and are removed. The remaining audio signal is then decomposed into non-overlapping frames of duration 50ms. These frames are decomposed using empirical mode decomposition into no. of intrinsic mode functions (IMFs). From first IMF the features having high discrimination capability are computed and feature vector is formed. These feature vectors are applied to k means clustering classifier for two clusters. According to the cluster no. the IMFs of that frames are added together to form complete signal. After clustering, from correctly identified frames, the accuracy is determined. Between clean speech and noisy speech, the accuracy is around 70% and between music and noisy speech the accuracy is around 70-75% is achieved. It is observed that

proposed system gives good accuracy for segmentation and reconstruction.

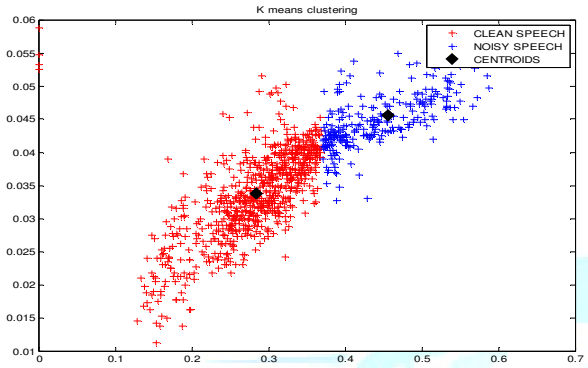

Fig 1: clustering between clean speech and noisy speech

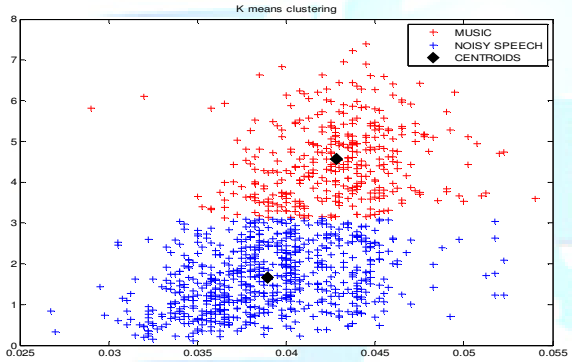

Fig 2: clustering between music and noisy speech

6. Conclusion

Automatic audio classification and segmentation is a complicated and problematic task. In this paper, we introduced use of EMD for audio segmentation. In this audio signal is decomposed into IMFs and features are extracted from these IMFs for identification of the segments. From classification results it is clear that feature combinations can identify the audio segments with very good accuracy. Using these feature combinations, segmentation and reconstruction is done. Experimental results show that EMD proved superior for audio segmentation. For future work, more features can be evaluated for improving the accuracy. Also more audio types can be used for segmentation to improve the segmentation capability.

References

[1] N. Nitanda, M. Haseyama, and H. Kitajima, "Accurate audio segment classification using feature extraction matrix," *in Proc. ICASSP, 2005*

[2] M. A. S. Seoane, A. R. Molares, and J. L. A. Castro, "Automatic classification of traffic noise," *in Proc. Acoustics'08, Paris, June 29–July 4, 2008*

[3] Deepak Jhanwar, Kamlesh K. Sharma and S. G. Modani, "Classification of Environmental Background Noise Sources Using Hilbert-Huang Transform", *International Journal of Signal Processing Systems Vol. 1, No. 1 June 2013*

[4] B.Han and E. Hwang, "Environmental sound classification based on feature collaboration," *in Proc. ICME, 2009*

[5] T. Lidy, R. Mayer, A. Rauber, P. J. Ponce de Leon, A. Pe rtusa, and J. M. Inesta, "A Cartesian Ensemble of Feature Subspace Classifiers for Music Categorization", *11th International Society for Music Information Retrieval Conference (ISMIR 2010), 2010, pp. 279-284.*

[6] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen,C. Tung, and H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Of the Royal Society of London , vol. 454, pp. 903–995, 1998.*

[7] Arijit Ghosal, Bibhas Chandra Dhara, Sanjoy Kumar Saha, "Speech/Music Classification Using Empirical Mode Decomposition", *Second International Conference on Emerging Applications of Information Technology, 2011.*

[8] Zhaohua Wu, Norden E. Huang, "Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method", *Advances In Adaptive Data Analysisvol. 1, No. 1 (2009) 1–41*

[9] George Tzanetakis, Perry Cook, "Musical Genre Classification of Audio Signals", *Ieee Transactions On Speech And Audio Processing, Vol. 10, No. 5, July 2002*

[10] Sujeet Kini, Sankalp Gulati, and Preeti Rao, "Automatic Genre Classification of North Indian Devotional Music", *National Conference on Communications, 2011, pp. 16-20.*

[11] M.J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," *in ICASSP, April 1999*

[12] Megha Agarwal, R.C.Jain, "Ensemble Empirical Mode Decomposition: An adaptive method for noise reduction ", *IOSR-JECE e-ISSN: 2278-2834,p- ISSN: 2278-8735. Volume 5, Issue 5 (Mar. - Apr. 2013), PP 60-65*

[13] T.Meera Devi, Dr.N.Kasthuri, Dr.A.M.Natarajan, "Environmental Noise Classification and Cancellation using Fuzzy Classifier and Fuzzy Adaptive Filters", *IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 2, March 2012*

[14] Martin Kermit, Age J. Eide, "Audio signal identi®cation via pattern capture and template matching", *Pattern Recognition Letters 21 (2000) 269±275*

[15] Matthew Cooper and Jonathan Foote, "SUMMARIZING POPULAR MUSIC VIA STRUCTURAL SIMILARITY ANALYSIS", *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*

[16] David Rybach, Christian Gollan, Ralf Schl¨uter, Hermann Ney, "AUDIO SEGMENTATION FOR SPEECH RECOGNITION USING SEGMENT FEATURES",