Audio Segmentation for Speech Recognition Using Segment Features

Gayatri M. Bhandari[1], Rameshwar S. Kawitkar[2], and Madhuri P. Borawake[3]

[1] J.J.T. University and
JSPM's Bhivarabai Sawant Institute of Tech. & Research(W),
Pune, India
gayatri.bhandari1980@gmail.com
[2] Sinhgad Institute of Technology, Pune
rskawitkar@rediffmail.com
[3] J.J.T. University and
PDEA, College of Engg., Pune
madhuri.borawake@gmail.com

Abstract. The amount of audio available in different databases on the Internet today is immense. Even systems that do allow searches for multimedia content, like AltaVista and Lycos, only allow queries based on the multimedia filename, nearby text on the web page containing the file, and metadata embedded in the file such as title and author. This might yield some useful results if the metadata provided by the distributor is extensive. Producing this data is a tedious manual task, and therefore automatic means for creating this information is needed. In this paper an algorithm to segment the given audio and extract the features such as MFCC, SF, SNR, ZCR is proposed and the experimental results shown for the given algorithm.

Keywords: Audio segmentation, Feature extraction, MFCC, LPC, SNR, ZCR.

1 Introduction

Audio exists at everywhere, but is often out-of-order. It is necessary to arrange them into regularized classes in order to use them more easily. It is also useful, especially in video content analysis, to segment an audio stream according to audio types.

In many applications we are interested in segmenting the audio stream into homogeneous regions. Thus audio segmentation is the task of segmenting a continuous audio stream in terms of acoustically homogenous regions [4]. The goal of audio segmentation is to detect acoustic changes in an audio stream. This segmentation can provide useful information such as division into speaker turns and speaker identities, allowing for automatic indexing and retrieval of all occurrences of a particular speaker. If we group together all segments produced by the same speaker we can perform an automatic online adaption of the speech recognition acoustic models to improve overall system performance.

S.C. Satapathy et al. (eds.), *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of CSI - Volume II*, Advances in Intelligent Systems and Computing 249,
DOI: 10.1007/978-3-319-03095-1_23, © Springer International Publishing Switzerland 2014

1.1 Segmentation Approaches

In typical segmentation methods are categorized into three groups, namely energy-based, metric-based, and model-based. The energy-based algorithm only makes use of the running power in time domain. On the other hand, both the metric-based and the model-based method are based on statistical models, say, multivariate Gaussian distributions. That means, rather than using the feature values directly, the running means and variances of them are modeled by a multidimensional Gaussian distribution.

1.1.1 Energy-Based Algorithm

The energy-based algorithm can be very easily implemented. Silence periods, that measured by the energy value and a predefined threshold, are assumed to be the segment boundaries. However, since there is no direct connection between the segment boundaries and the acoustic changes, this method can be problematic for many applications, such as gender detection, and speaker identification, etc.

1.1.2 Model-Based Algorithm

In the model-based algorithm, statistical distribution models are used for each acoustic class (e.g., speech, music background, noise background, etc.) The boundaries between classes are used as the segment boundaries. Typically, Bayesian Information Criterion (BIC) is used to make the decision if the changing point turns out, which is essentially a hypothesis testing problem.

1.1.3 Metric-Based Algorithm

In the metric-based algorithm, statistical distribution models are also used for modeling the feature space. Gaussian model is a typical choice, but some other distributions can also be used. For example, Chi-squared distribution are found to be appropriate and with less computational cost in. The sound transition is measured by the distance between the distributions of two adjacent windows. The local maximum of distance value suggests a changing point.

According to this here six different classes of audio are defined.

1. Speech: This is pure speech recorded in the studio without background such as music.

2. Speech over Music: This category includes all studio speech with music in the background.

3. Telephone Speech: Some sections of the program have telephonic interventions from the viewers. These interventions are mixed in the program's main audio stream as a wide band stream.

4. Telephone Speech over Music: The same as previous class but additionally there is music in the background.

5. Music: Pure music recorded in the studio without any speech on top of it.

6. Silence

Real-time speaker segmentation is required in many applications, such as speaker tracking in real-time news-video segmentation and classification, or real-time speaker adapted speech recognition. Here real-time, yet effective and robust speaker segmentation algorithm based on LSP correlation analysis can be done. Both the speaker identities and speaker number are assumed unknown. The proposed incremental speaker updating and segmental clustering schemes ensure this method can be processed in real-time with limited delay.

1.2 Related Work

Several methods have been developed for audio segmentation. Chen identifies two types of segmentation approaches namely, classification-dependent segmentation (CDS) and classification-independent segmentation (CIS) [1]. CDS methods are problematical because it is difficult to control the performance [1].

CIS approaches can be further separated into time-domain and frequency-domain depending upon which audio features they use, or supervised and unsupervised approaches depending on whether the approach requires a training set to learn from prior audio segmentation results. CIS may also be defined as model-based or non-model based methods.

In model-based approaches, Gaussian mixture models (GMM) [4], [5], Hidden Markov Models (HMM) [6], Bayesian [7], and Artificial Neural Networks (ANN) [8] have all been applied to the task of segmentation. Examples of an unsupervised audio segmentation approach can be found in [7] and [9]. These unsupervised approaches test the likelihood ratio between two hypotheses of change and no change for a given observation sequence. On the other hand, the systems developed by Ramabhadran et al. [6] and Spina, and Zue [4] must be trained before segmentation. These existing methods are limited because they deal with limited and narrow classes such as speech/music/noise/ silence.

Audio segmentation methods based on a similarity matrix, have been employed for broadcasting news, which is relatively acoustic dissimilar, and for music to extract structures or music summarizations. The accuracy evaluation of these methods was undertaken with specific input audios and has not been previously reported for use

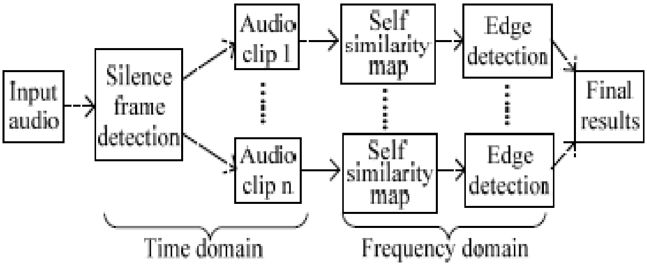

Fig. 1. Frame work for audio segmentation method

with audio files in a non-music/non-speech database. This paper introduces a two phase unsupervised model-free segmentation method that works for general audio files. In this paper, we discuss the process by which we developed and evaluated an efficient CIS method that can determine segment boundaries without being supplied with any information other than the audio file itself.

2 Feature Analysis

Fig. 2 shows the basic processing flow of the proposed approach that integrates audio segmentation and speaker segmentation. After feature extraction, the input digital audio stream is classified into speech and nonspeech. Nonspeech segments are further classified into music, environmental sound, and silence, while speech segments are further segmented by speaker identity [2].

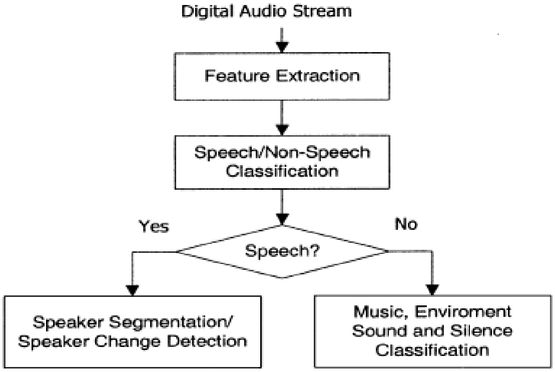
Digital Audio Stream
Feature Extraction
Speech/Non-Speech Classification
Yes
No
Speech?
Speaker Segmentation/ Speaker Change Detection
Music, Enviroment Sound and Silence Classification

Fig. 2. Basic Processing flow of audio content analysis

3 Objectives of the Proposed Research

1. The objective of audio segmentation for classifying the audio components into Speech Music , Non-speech , noise , silence along with the other major features such as MFCC , SF , LPC , SNR , HZCRR etc and can be transferred over the network and by analyzing these audio features the reconstruction of audio signal should be more accurate.
2. We proposed to use thirteen features in time, frequency, and cepstrum domains and model-based (MAP, GMM, KNN, etc.) classifier, which achieved an accuracy rate over 90% on real-time discrimination between speech and music. As in general, speech and music have quite different spectral distribution and temporal changing patterns, it is not very difficult to reach a relatively high level of discrimination accuracy.
3. Further classification of audio data may take other sounds into consideration besides speech and music.

4. We also proposed an approach to detect and classify audio that consists of mixed classes such as combinations of speech and music together with environment sounds. The accuracy of classification is more than 80%.
5. An acoustic segmentation approach was also proposed where audio recordings to be segmented into speech, silence, laughter and non-speech sounds.

We have to use cepstral coefficients as features and the Hidden Markov model (HMM) as the classifier. We propose a MGM-based (Modified Gaussian Modelling) hierarchical classifier for audio stream classification. Compared to traditional classifiers, MGM can automatically optimize the weights of different kinds of features based on training data. It can raise the discriminative capability of audio classes with lower computing cost.

4 System Flow

Fig. 3 shows the flowchart of proposed audio segmentation and classification algorithm. It is a hierarchical structure. In the first level, a long audio stream can be segmented into some audio clips according to the change of background sound by MBCR based histogram modeling. Then a two level MGM (Modified Gaussian modeling) classifier is adopted to hierarchically put the segmented audio clips into six pre-defined categories in terms of discriminative background sounds, which is pure speech (PS), pure music (PM), song (S),speech with music (SWM), speech with noise (SWN)and silence (SIL).

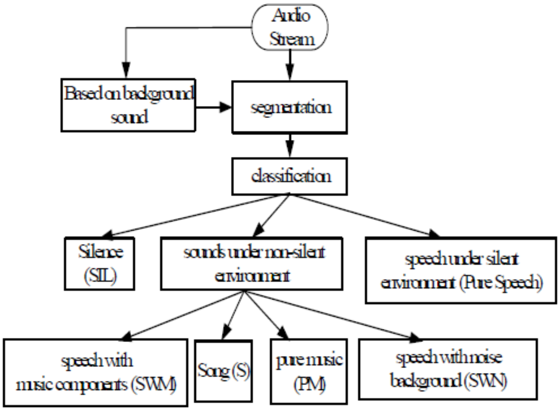
Audio Stream
Based on background sound
segmentation
classification
Silence (SIL)
sounds under non-silent environment
speech under silent environment (Pure Speech)
speech with music components (SWM)
Song (S)
pure music (PM)
speech with noise background (SWN)

Fig. 3. The flowchart of segmentation and classification algorithm

5 Segmentation Algorithm

Since background sounds always change with the change of scenes, the acoustic skip point of an audio stream may be checked by background sounds. As shown in Fig. 2, the MBCR feature vectors are firstly extracted from the audio stream. We set a sliding window which consists of two sub-windows with equal time length. The window on input signal is shifted with a range of overlapping. Then two histograms are created from each sliding sub-windows. The similarity between two sub-windows can be measured by histogram matching. The skip point can thus be detected by searching the local lowest similarity below a threshold.

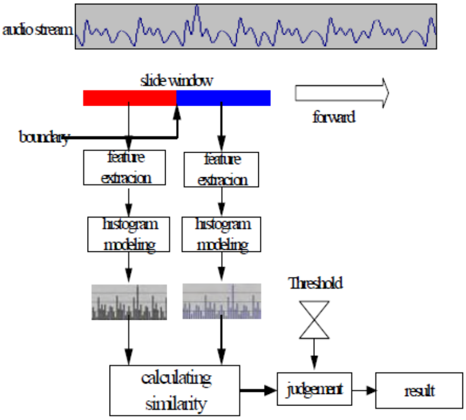

Fig. 4. Block diagram of Segmentation algorithm

The proposed algorithm for audio segmentation segment the audio into different parameters as described before also feature extraction algorithm separates out the different audio features such as MFCC, LPC, SF, SNR, and HZCRR

6 Feature Extraction

Considering the lower frequency spectrum is too sensitive to even a bit of changes of the scenes and speakers, it could cause segmented clips too small. It will have effects on succeeding audio classification. We, thus, use Multiple sub-Bands spectrum Centroid relative Ratio (MBCR) [5] over 800Hz as basic feature. This feature may depict centroid movement trend in a time frequency-intensity space. Its mathematical description can be described as follows.

$$SCR(i,j) = \frac{SC(i,j)}{\underset{J = 1:N}{\mathrm{Max}}\ (SC(i,j))} \qquad (1)$$

$$SC(i,j) = \frac{f(j) * FrmEn(i,j\ )}{\sum_{k=1}^{N} FrmEn(i,k)} \qquad (2)$$

where *SCR*(*i*, *j*) is MBCR of the ***ith*** frame and the **_jth_sub**-band, *SC*(*i*, *j*) is the frequency Centroid of the ***ith***frame and the ***jth***sub-band, and *N* denotes the number of frequency sub-bands. The element of *f* (*j*) is the normalized central frequency.

$$FrmEn(i,j) = \log(\int_{\omega_L(j)}^{\omega_H(j)} \left|F(i,\omega)\right| d\omega) \qquad (3)$$

where ω *L* (*j*) and ω *H* (*j*) are lower and upper bound of sub-band *j* respectively, *F*(*i*,ω) represent denotes the Fast Fourier Transform (FFT) at the frequency ω and frame *i*, and | *F*(*i*,ω) | is square root of the power at the frequency ω and frame *i*.

7 Results

We conducted a series of experiments based on proposed audio segmentation and classification approach. The performance was evaluated on the recordings of real TV program. The segmentation and classification results were evaluated by the recall rateδ, accuracy rateξ, and average precisionη. These are defined as

$$\delta = \frac{\text{the number of correctly objects}}{\text{the number of objects that should be correct}}$$

$$\xi = \frac{\text{the number of correctly objects}}{\text{the number of all get objects}}$$

$$\eta = \frac{\delta * \xi}{0.5*(\delta + \xi)}$$

We pre-defined six categories as audio classes, which is pure speech (PS), pure music (PM), song (S), speech with music (SWM), speech with noise (SWN) and silence (SIL).

Table 1. The results of first level classification

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

8 Conclusions

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 has been evaluated for pure speech, silence etc.

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 in order to reduce the storage capacity which is under process. The above algorithms can be modified to extract other than above features which are not mentioned here.

References

[1] Peiszer, E., Lidy, T., Rauber, A.: Automatic Audio Segmentation: Segment Boundary and Structure Detection in Popular Music (2008)

[2] Cook, G.T.P.: Multifeature Audio Segmentation for Browsing and Annotation. In: Proc.1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, pp. W99-1–W99-4 (1999)

[3] Lu, G.: Indexing and Retrieval of Audio: A Survey, pp. 269–290 (2001)

[4] Zhang, J.X., Whalley, J., Brooks, S.: A Two Phase Method for general audio segmentation (2004)

[5] Foote, J.: Automatic Audio Segmentation Using A Measure of Audio Novelty

[6] Julien, P., José, A., Régine, A.: Audio classi_cation by search of primary components, pp. 1–12

[7] Lu, L., Zhang, H.-J., Jiang, H.: Content Analysis for Audio Classification and Segmentation. IEEE Transaction on Speech and Audio Processing, 504–516 (2002)

[8] Lu, L., Li, S.Z., Zhang, H.-J.: Content based audio segmentation using Support Vector Machines (2008)

[9] Aguilo, M., Butko, T., Temko, A., Nadeu, C.: A Hierarchical Architecture for Audio Segmentation in a Broadcast News Task, pp. 17–20 (2009)

[10] Cettolo, M., Vescovi, M., Rizzi, R.: Evaluation of BIC-based algorithms for audio segmentation, pp. 147–170. Elsevier (2005)

[11] Goodwin, M.M., Laroche, J.: Audio Segmentation by feature space clustering using linear discriminant analysis and dynamic programming (2003)

[12] Haque, M.A., Kim, J.-M.: An analysis of content-based classification of audio signals using a fuzzy c-means algorithm (2012)

[13] Mesgarani, N., Slaney, M., Shamma, S.A.: Discrimination of Speech From Nonspeech Based on Multiscale Spectro-Temporal Modulations, pp. 920–930 (2006)

[14] Krishnamoorthy, P., Kumar, S.: Hierarchical audio content classification system using an optimal feature selection algorithm, pp. 415–444 (2010)

[15] Panagiotis, S., Vasileios, M., Ioannis, K., Hugo, M., Miguel, B., Isabel, T.: On the use of audio events for improving video scene segmentation

[16] Abdallah, S., Sandler, M., Rhodes, C., Casey, M.: Using duration Models to reduce fragmentation in audio segmentation 65, 485–515 (2006)

[17] Cheng, S.-S., Wang, H.-M., Fu, H.-C.: BIC-BASED Audio Segmentation by divide and conquer

[18] Yong, S.: Audio Segmentation, pp. 1–4 (2007)

[19] Matsunaga, S., Mizuno, O., Ohtsuki, K., Hayashi, Y.: Audio source segmentation using spectral correlation features for automatic indexing of broadcast news, pp. 2103–2106

[20] Sainath, T.N., Kanevsky, D., Iyengar, G.: Uusupervised audio segmentation using extended Baum-Welch Transformations, I 209-I 212 (2007)

[21] Giannakopoulos, T., Pikrakis, A., Theodoridis, S.: A Novel Efficient Approach for Audio Segmentation (2008)

[22] Zhang, Y., Zhou, J.: Audio Segmentation based on Multiscale audio classification, pp. IV-349–IV-352 (2004)

[23] Peng, Y., Ngo, C.-W., Fang, C., Chen, X., Xiao, J.: Audio Similarity Measure by Graph Modeling and Matching, pp. 603–606

[24] Harchaoui, Z., Vallet, F., Lung-Yut-Fong, A., Cap, O.: Regularized Kernel-Based ApproachToUnsupervised Audio Segmentation