Innovative Technique for Audio Segmentation

Borawake Madhuri P.
Lectuer
63,Shree Ganesh Vihar Col.,
Wagoli,Pune Serum Inst.Rd.Near Akashwani

Kawitkar Rameshwar
Professor
SCOE,PUNE

Bhandari G.M
Ass.Professor
J.S.P.M.,BS COE ,
Hadapsar,Pune -411028

ABSTRACT

Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages. The term applies both to the mental processes used by humans, and to artificial processes of processing. Speech segmentation is an important sub problem of speech recognition, and cannot be adequately solved in isolation. The lowest level of speech segmentation is the breakup and classification of the sound signal into a string of phones. The difficulty of this problem is compounded by the phenomenon of co-articulation of speech sounds, where one may be modified in various ways by the adjacent sounds: it may blend smoothly with them, fuse with them, split, or even disappear. This phenomenon may happen between adjacent words just as easily as within a single word. The notion that speech is produced like writing, as a sequence of distinct vowels and consonants. In fact, the way we produce vowels depends on the surrounding consonants and the way we produce consonants depends on the surrounding vowels. Therefore, even with the best algorithms, the result of phonetic segmentation will usually be very distant from the standard written language.

Keywords: audio content analysis, audio database management, audio segmentation .

1. INTRODUCTION

A real-time audio segmentation scheme is presented in this paper. Audio recordings are segmented and classified into basic audio types such as silence, speech, music, song, environmental sound, speech with the music background, environmental sound with the music back- ground, etc. Simple audio features such as the energy function, the average zero-crossing rate, the fundamental frequency, and the spectral peak track are adopted in this sys- tem to ensure on-line processing. Morphological and statistical analysis for temporal curves of these features are performed to show differences among different types of audio. A heuristic rule-based procedure is then developed to segment and classify audio signals by using these features. The proposed approach is generic and model free. It can be applied to almost any content-based audio management system. It is shown that the proposed scheme achieves an accuracy rate: of more than 90% for audio classification. Examples for segmentation and indexing of accompanying audio signals in movies and video programs are also provided.

Audio, which includes voice, music, and various kinds of environmental sounds, is an important type of media, and also a significant part of audiovisual data. Compared to research done on content-based image and video database management, very little work has been done on the audio part of the multimedia stream. However, since there are more and more digital audio databases in place these days, people begin to realize the importance of effective manage- ment for audio databases relying on audio content analysis. Audio segmentation and classification have applications in professional media production, audio archive management, commercial music usage, surveillance, and so on. Further-more, audio content analysis may play a primary role in video annotation. Current approaches for video segmentation and indexing are mostly focused on the visual information. However, visual-based processing often leads to a far too fine segmentation of the audiovisual sequence with respect to the semantic meaning of data. Integration of the diverse multimedia components (audio, visual, and textual information) will be essential in achieving a fully functional system for video parsing. Existing research on content-based audio data management is very limited. There are in general four directions. One direction is audio segmentation and classification. One basic problem is speech/music discrimination [8], [9]. Further classification of audio may take other sounds into consideration , where audio was classified into “music”, “speech” and “others”. It was developed for the parsing of news stories. In [4], audio recordings were classified into speech, silence, laughter, and non-speech sounds for the purpose of segmenting discussion recordings in meetings. The second direction is audio retrieval. One specific technique in content-based audio retrieval is query by-humming, and the work in [3] gives a typical example. Two approaches for generic audio retrieval were presented, respectively, in [2] and [lo]. The third direction is audio analysis for video indexing. Audio analysis was applied to the distinction of five kinds of video scenes: news report, weather report, basketball game, football game, and advertisement in [5]. Audio characterization was performed on MPEG sub-band level data for the purpose of video indexing in [7]. The fourth direction is the integration of audio and visual information for video segmentation and indexing. Two approaches were proposed in [I] and [6], respectively. In this research, we propose a heuristic rule-based approach for the segmentation and annotation of generic audio data. Compared with existing work, there are several distinguishing features of this scheme, as described below.

First, besides the commonly studied audio types such as speech and music, we have included into this scheme hybrid sounds which contain more than one basic audio type. For example, the speech signal with the music background and the singing of a person are two types of hybrid sounds which have characters of both speech and music. We classify these kinds of sounds into additional different categories in our system, because they are very important in characterizing audiovisual segments. For example, in document trieres or commercials, there is usually a musical background with speech of commentary appearing from time to time. It is also common

that clients want to retrieve the segment of video, in which there is singing of one particular song. There are other kinds of hybrid sounds such as speech or music with environmental sounds as the background (where the environmental sounds may be treated as noise), or environmental sounds with music as the background. Second, we put more emphasis on the distinction of environmental audio which is often ignored in previous work. Environmental sounds are an important ingredient in audio recordings, and their analysis is inevitable in many real applications. In this work, we divide environmental sounds into six categories according to their harmony, periodicity and stability properties. Third, feature extraction schemes are investigated based on the nature of audio signals and the problem of interest. For example, the short-time features of energy, the aver- age zero-crossing rate and the fundamental frequency are combined organically in distinguishing silence, speech, music and sounds in the environment . We use not only the feature values, but also their change patterns over the time and the relationship among the three features. We also propose a method to detect the spectral peak track and use this feature specifically for the distinction of sound segments of the song and speech with the music background. Finally, the proposed approach is real-time and model-free. It can be easily applied to any audio or audiovisual data management system. The framework of the proposed scheme is illustrated in Figure 1.

The paper is organized as follows. In Section 2, the computations and characteristics of audio features used in this research are analyzed. The proposed procedures for the segmentation and indexing of generic audio data are described in Section 3. Experimental results are shown in Section 4. Finally, concluding remarks and future research plans are given in Section 5.

2. AUDIO FEATURE ANALYSIS

Short-Time Energy Function

The short-time energy function of an audio signal is defined as

$$E_n = \frac{1}{N}\sum_m [x(m)w(n-m)]^2$$

Where x(m) is the discrete time audio signal , n is time index of the short time energy & w
(m) is rectangle window, i.e.

$$w(n) = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix} 0 \le n \le N-1$$

It provides a convenient representation of the amplitude variation over the time. The main reasons of using the short-time energy feature in our work include the following. First, for speech signals, it provides a basis for distinguishing voiced speech components from unvoiced speech components. This is due to the fact that values of E, for the unvoiced components are in general significantly smaller than those of the voiced components. Second, it can be used as the measurement to distinguish audible sounds from silence when the signal-to-noise ratio is high. Third, its change pattern over the time may reveal the rhythm and periodicity nature of the underlying sound.

Short-Time Average Zero-Crossing Rate

In the context of discrete-time signals, a zero-crossing is said to occur if successive samples have different signs. The rate at which zero-crossings occur is a simple measure of the frequency content of a signal. The short-time averaged zero-crossing rate is defined as

$$z_n = \frac{1}{2}\sum |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]|\, w(n-m)$$

where

$$\operatorname{sgn}[x(m)] = \begin{Bmatrix} 1 \\ -1 \end{Bmatrix} x(n) \ge 0, x(n) \prec 0$$

and w(n) is a rectangle window of length N. Temporal curves of the average zero-crossing rate (ZCR) for several audio samples are shown in Figure 2. The averaged zero-crossing rate can be used as another measure for making distinction between voiced and unvoiced speech signals, because unvoiced speech components normally have much higher ZCR values than voiced ones. As shown in Figure 2(a), the speech ZCR curve haa peaks and troughs from unvoiced and voiced components, respectively. This results in a large variance and a wide range of amplitudes for the ZCR curve. Note also that the ZCR waveform has a relatively low and stable baseline with high peaks above it. Compared to that of speech signals, the ZCR curve of music has a much lower variance and average amplitude as plotted in Figure 2(b). This suggests that the averaged zero-crossing rate of music is normally much more stable during a certain period of time. ZCR curves of music generally have an irregular waveform with a changing baseline and a relatively small range of amplitudes . Since environmental audio consists of sounds from various
origins, their ZCR curves can have very different properties. For example, the ZCR curve of the sound of chime reveals a continuous drop of the frequency centroid over the time while that of the footstep sound is rather irregular. Generally speaking, we can classify environmental sounds according to properties of their ZCR curves such as regularity, periodicity, stability, and the range of amplitudes.

Short-Time Fundamental Frequency :A harmonic sound consists of a series of major frequency components including the fundamental frequency and those which are integer multiples of the fundamental one. With this concept, we may divide sounds into two categories, i.e. harmonic and non-harmonic sounds. The spectra of sounds generated by trumpet and rain are illustrated in Figure 3. It is clear that the former one is harmonic while the latter one is non-harmonic.

Whether an audio segment is harmonic or not depends on its source. Sounds from most musical instruments are harmonic. The speech signal is a harmonic and non-harmonic mixed sound, since voiced components are harmonic while unvoiced components are non-harmonic. Most environmental sounds are non-harmonic, such as the sounds of applause, footstep and explosion. However, there are also examples of sound effects which are harmonic and stable, like the sounds of doorbell and touch-tone; and those which are harmonic and non-harmonic mixed such as the sounds of laughter and dog bark.

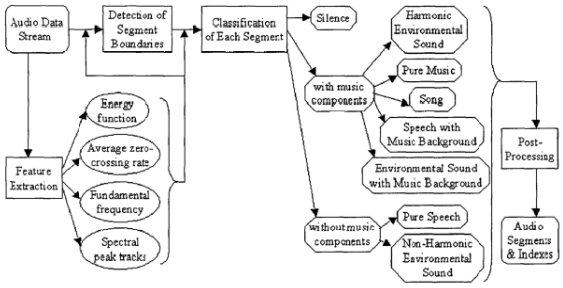


Fig.1 Generic Segmentation of audio data

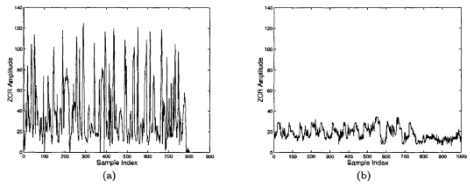


Fig.2 The Short Time Average zero crossing rates curves a) Speech b) Piano

In order to measure the harmony feature of sounds, we define the short-time fundamental frequency as follows. When the sound is harmonic, the value is equal to the fundamental frequency estimated from the audio signal. When the sound is non-harmonic, it is set to zero. In this work, the fundamental frequency is calculated based on peak detection from the spectrum of the sound. The Post- Processing spectrum is generated with autoregressive (AR) model coefficients estimated from the autocorrelation of audio signals. This AR model generated spectrum is a smoothed version of the frequency representation. Moreover, as the AR model is an all-pole expression, peaks are prominent in the spectrum. Detecting peaks associated with harmonic frequencies is much easier in the AR generated spectrum than in the spectrum directly computed with FFT. In Order to keep a good precision of the estimated fundamental frequency, we choose the order of the AR model to be 40. With this order, harmonic peaks are remarkable while there are also non-harmonic peaks appearing. However, compared with harmonic peaks, non-harmonic ones lack a precise harmonic relation among them and usually have local maxima that are less sharp and of a smaller height. To summarize, a sound is classified to be harmonic, if there is a least-common-multiple relation among peaks, and some peaks are sharp and high.

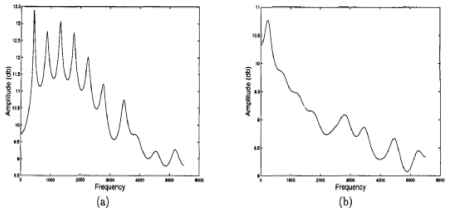


Fig.3 Spectra of harmonic & nonharmonic sounds a) trumpet b) rain

Examples of fundamental freq. curves of sounds are illustrated in Figure 4. Shown on top of each picture is the "zero ratio" of the fundamental freq. curve, which is defined as the ratio between the number of samples with a zero fundamental freq. value (i.e. the non harmonic sound) and the total number of samples in the curve. We can see that music is generally continuously harmonic. Also, the fundamental freq. value tends to concentrate on certain values for a short period of time in music. Harmonic and non-harmonic components appear alternately in the fundamental freq. curve of the speech signal, since voiced components are harmonic and unvoiced components are non-harmonic. The fundamental frequency of voiced components is normally in the range of lOO-300Hz. Most environmental sounds are non-harmonic with zero ratios over 0.9. The sound of rain is an example of them. An instance of the mixed harmonic and non-harmonic sound is the sound of laughing, in which voiced segments are harmonic, while intermissions in between as well as transitional parts are non-harmonic. It has a zero ratio of 0.25 which is similar to that of the speech segment.

Spectral Peak Track

The peak track in the spectrogram of an audio signal often reveals important characteristics of the sound. For example, sounds from musical instruments normally have spectral peak tracks which remain at the same frequency level and last for a certain period of time. Sounds from human voices have harmonic peak tracks in their spectrograms which align tidily in a comb shape. The spectral peak tracks in songs may exist in a broad range of frequency bands, and the fundamental frequency ranges from 87Hz to 784Hz. There are relatively long tracks in songs which are stable because the voice stays at a certain note for a period of time, and they are often in a ripple-like shape due to the vibration of vocal chords. Spectral peak tracks in speech normally lie in the lower frequency bands, and are more close to each other due to the fundamental frequency range of lOO-3OOHz. They also tend to be of a shorter length because there are intermissions between voiced syllables, and may change slowly because the pitch may change during the pronunciation of certain syllables. We extract spectral peak tracks for the purpose of characterizing sounds of songs and speech. Basically, it is done by detecting peaks in the power spectrum generated by the AR model parameters and checking harmonic relations among peaks. Compared to the problem of fundamental frequency estimation where the precision requirement is less strict and slight errors are allowed, the task here is more difficult since the locations of tracks should be determined more accurately. However, by using the fact that only spectral peak tracks in song and speech segments are considered, we are able to derive a set of rules to pick up proper harmonic peaks based on distinct features of such tracks as described above.

Harmonic peaks detected through our developed procedure for two frames of song and speech signals are shown in Figure 5, where each detected peak is marked with a vertical line. Locations of detected peaks are aligned along the temporal direction to form spectral peak tracks. Spectrograms and spectral peak tracks estimated with our method for two segments of song and speech signals are illustrated in Figures 6 and 7. The first segment is female vocal solo without accompanying musical instruments. There are seven notes in the segment as "5-1-6-4-3-1-2". We see that the pitch and the duration of each note are clearly reflected in detected peak tracks. The harmonic tracks range from the fundamental frequency at about 225-400Hz up to 5000Hz, and are in a ripple-like shape. The second segment is female speech with music and other noise in the background. However, the speech signal seems to be dominant in the spectrogram, and the spectral peak tracks are nicely detected despite the interference. These tracks are shorter than those in the song segments and have a pitch level of 150-250Hz.

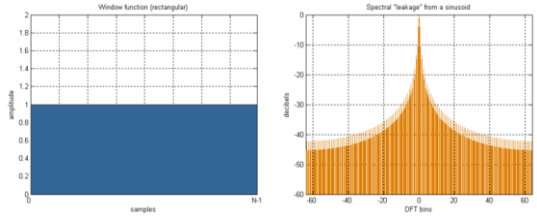
3. WINDOWING

The window determines the portion of the speech signal that is to be processed by zeroing out the signal outside the region of interest. The ideal window frequency response has a very narrow main lobe which increases the resolution and no side lobes (or frequency leakage). Since such a window is not possible in practice, a compromise is usually selected foe each specific application**.**

1. Rectangular Window:

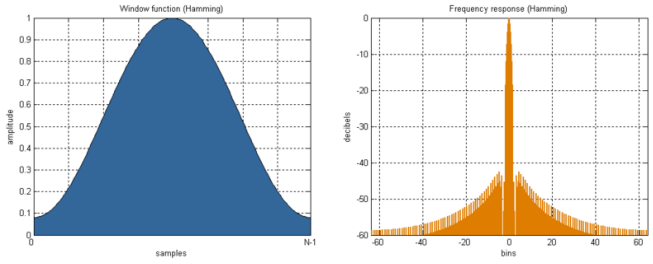
$$w(n) = 1$$

Rectangular window; B=1.00



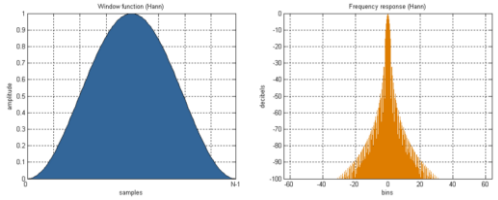
2. Hamming Window:

$$w(n) = 0.53836 - 0.46164\ \cos\left(\frac{2\pi n}{N-1}\right)$$



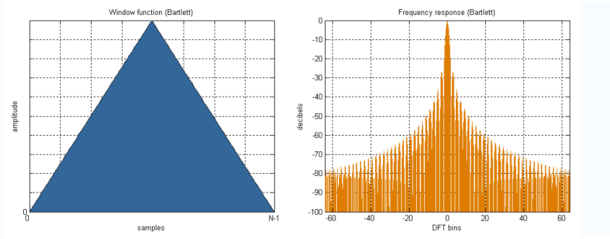
3. Hanning Window:

$$w(n) = 0.5\ \left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right)$$



4. Bartlett Window:

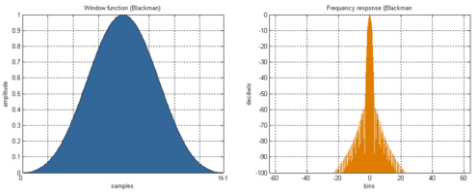
$$w(n) = \frac{2}{N-1} \cdot \left(\frac{N-1}{2} - \left|n - \frac{N-1}{2}\right|\right)$$



5. Blackman Window:

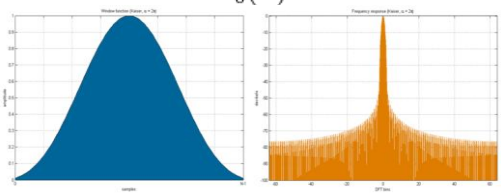
$$w(n) = a_0 - a_1 \cos\left(\frac{2\pi n}{N-1}\right) + a_2 \cos\left(\frac{4\pi n}{N-1}\right)$$

$$a_0 = 0.42; \quad a_1 = 0.5; \quad a_2 = 0.08$$



6. Kaiser Window:

$$w(n) = \frac{I_0\left(\alpha\sqrt{1 - (\frac{2n}{N-1} - 1)^2}\right)}{I_0(\alpha)}$$



change detected in any of these three features, a segment boundary is set. In the temporal curve of each feature, there are two adjoining sliding windows installed with the average amplitude computed within each window. The sliding windows proceed together with newly computed feature values, and the average amplitude within each window is updated. We compare these two values. Whenever there is a significant difference between them, an abrupt change is claimed to be detected at the common edge of the two windows. Examples of boundary detection from temporal curves of short-time energy function and short-time fundamental frequency are shown in Figure 8. We see that because the temporal evolution pattern and the range of amplitudes of short-time features are different for speech, music, environmental sound, etc., dramatic changes can be detected from these features at boundaries of different audio types.

4. CONCLUSION

We presented in this research a heuristic approach for the parsing and annotation of audio signals based on the analysis of audio features and a rule-based procedure. It was shown that an on-line segmentation and classification of audio data into twelve basic types were accomplished with this approach. The segmentation boundaries were set accurately, and a correct classification rate higher than 90% was achieved. Further research can be done in two areas. One is audio

feature extraction in the compressed domain such as .wav file

6. REFERENCES

[1] Boreczky, J. S. and Wilcox, L. D.: A hidden Markov model framework for video segmentation using audio and image features, in Proceedings of ICASSP'98 pp.3741-3744, Seattle, May 1998.

[2] Foote, J.: Content-based retrieval of music and audio ,in Proceedings of SPIE'97, Dallas, 1997.

[3] Ghias, A., Logan, J. and Chamberlin, D.: Query by humming - musical information retrieval in an audio database, in Proceedings of ACM Multimedia Conference, ~~~231-235, 1995.

[4] Kimber, D. and Wilcox, L.: Acoustic segmentation for audio browsers, in Proceedings of Interface Conference, Sydney, Australia, July 1996.

[5] Liu, Z., Huang, J., Wang, Y. et al.: Audio feature extraction and analysis for scene classification, in Proceedings of IEEE 1st Multimedia Workshop, 1997.

[6] Naphade, M. R., Kristjansson, T., Frey, B. et al.Probabilistic multimedia objects (MULTIJECTS): a

novel approach to video indexing and retrieval in multimedia systems, in Proceedings of IEEE Conference

[7] Patel, N. and Sethi, I.: Audio characterization for video indexing, in Proceedings of SPIE Conference

on Storage and Retrieval for Still Image and Video Databases, ~01.2670, ~~~373-384, San Jose, 1996.

[8] Saunders, J.: Real-time discrimination of broadcast speech/music, in Proceedings of ICASSP'96.

[9] Scheirer, E. and Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator, in Proceedings of ICASSP'97, Munich, Germany, Apr. 1997.

[1o] Wold, E., Blum, T. and Keislar, D. et al.: Contentbased classification, search, and retrieval of audio,

IEEE Multimedia, pp.27-36, Fall, 1996.

[ll] Wyse, L. and Smoliar, S.: Toward content-based audio indexing and retrieval and a new speaker discrimination technique, in http://www.iss.nus.sg/People/lwyse/lwyse.html, Dec.

[12]Haykin Simon," *Communication System* " , Fourth Edition, Wiley Student Edition(2001), Reprint (2005)

[13] Taub Herbert, Donald Schilling," *Principles Of Communication Systems*", Second Edition, Tata Mcgraw-Hill (1991) Edition, Thirty Second Reprint(2005)

[14]Hayes Monson, "*Digital Signal Processing*", Third Edition, Tata McGraw-Hill Edition, New Delhi, (2004)

[15]Ifeachor Emmanuel C. & Barrie W. Jervis, "*Digital Signal Processing*", Second Edition, First Indian Reprint, Pearson Education, New Delhi (2002)

[16]Ingle K. Vinay & John J. Proakis, "*Digital Signal Processing Using Matlab*", International Student Edition, Thomson Books, Vikas Publishing House, Fifth Reprint Bangalore (2004)

[17]Ludeman Lonnie C., "*Fundamentals of Digital Signal Processing*", First Edition, Wiley publication, Singapore (1986)

[18]Proakis John G. & Dimitrise G. Manolakis, "*Digital Signal Processing, Principle Algorithms & Applications*", Third Edition, Sixth Indian Reprint, Pearson Education, New Delhi (2005)

[19]Duda Richard O., Peter E. Hart, David G. Stork, "Pattern Classification", Second Edition, John Willey & Sons, Singapore (2004)

[20]"Window function", on http://en.wikipedia.org/wiki/Window_function