Heuristic Approach for Generic Audio Data Segmentation and Annotation

Tong Zhang and C.-C. Jay Kuo
Integrated Media Systems Center and Department of Electrical Engineering-Systems
University of Southern California, Los Angeles, CA 90089-2564
Tel: +1-213-7404658
Email:{tzhang,cckuo}@sipi.usc.edu

ABSTRACT

A real-time audio segmentation and indexing 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 background, 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 system 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.

Keywords: audio content analysis, audio database management, audio segmentation and indexing, heuristic rules.

1. INTRODUCTION

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 management 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. Furthermore, 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 as done in [11], 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 [10]. 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 [1] 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

ACM Multimedia '99 10/99 Orlando, FL, USA
© 1999 ACM 1-58113-151-8/99/0010...$5.00

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 documentaries 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 enviromental 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 average 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

2.1. Short-Time Energy Function

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

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

where $x(m)$ is the discrete time audio signal, n is time index of the short-time energy, and $w(m)$ is a rectangle window, i.e.

$$w(n) = \begin{cases} 1, & 0 \leq n \leq N-1, \\ 0, & \text{otherwise.} \end{cases}$$

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_n 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.

2.2. 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

$$Z_n = \frac{1}{2}\sum_m |sgn[x(m)] - sgn[x(m-1)]|w(n-m), \tag{2}$$

where

$$sgn[x(n)] = \begin{cases} 1, & x(n) \geq 0, \\ -1, & x(n) < 0, \end{cases}$$

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 has 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.

2.3. 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.

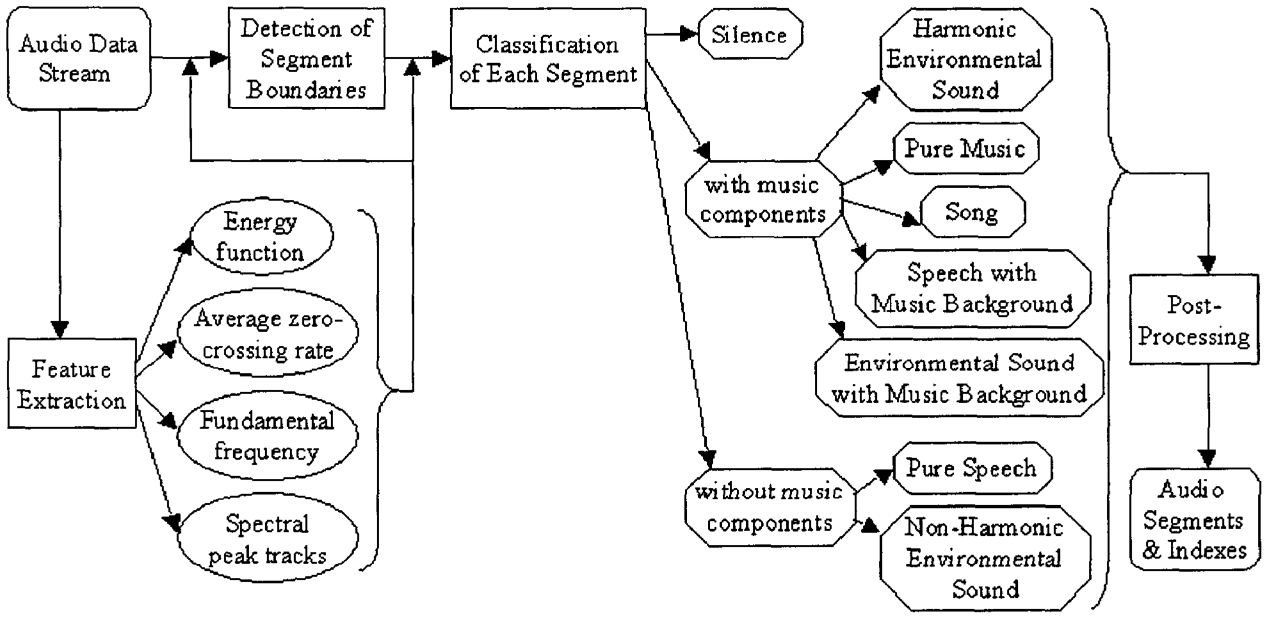
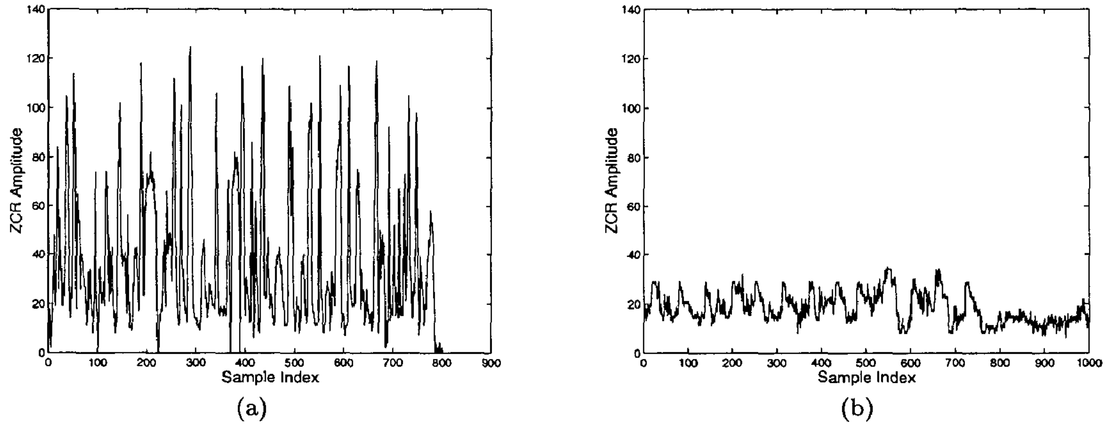

Audio Data Stream
Detection of Segment Boundaries
Classification of Each Segment
Silence
Harmonic Environmental Sound
Pure Music
with music components
Song
Speech with Music Background
Environmental Sound with Music Background
Pure Speech
without music components
Non-Harmonic Environmental Sound
Post-Processing
Audio Segments & Indexes
Energy function
Average zero-crossing rate
Feature Extraction
Fundamental frequency
Spectral peak tracks

Figure 1: Automatic segmentation and indexing of generic audio data.



(a)

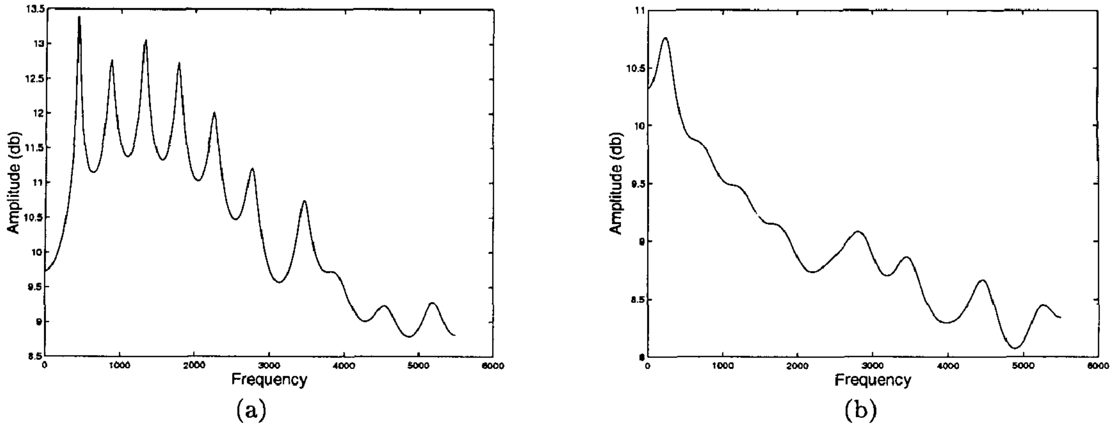
(b)

Figure 2: The short-time averaged zero-crossing rate curves: (a) speech and (b) piano.

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.

In order to measure the harmony feature of sounds, we define the short-time fundamental frequency (FuF) as follows. When the sound is harmonic, the FuF 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 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.



(a) (b)

Figure 3: Spectra of harmonic and non-harmonic sounds: (a) trumpet and (b) rain.

Examples of FuF curves of sounds are illustrated in Figure 4. Shown on top of each picture is the "zero ratio" of the FuF curve, which is defined as the ratio between the number of samples with a zero FuF 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 FuF value tends to concentrate on certain values for a short period of time in music. Harmonic and non-harmonic components appear alternately in the FuF 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 100-300Hz. 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.

2.4. 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 100-300Hz. 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 sung 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. HEURISTIC PROCEDURES FOR SEGMENTATION AND INDEXING OF GENERIC AUDIO DATA

3.1. Detection of Segment Boundaries

For on-line segmentation of audio data, the short-time energy function, short-time averaged zero-crossing rate, and short-time fundamental frequency are computed on the fly with incoming audio data. Whenever there is an abrupt

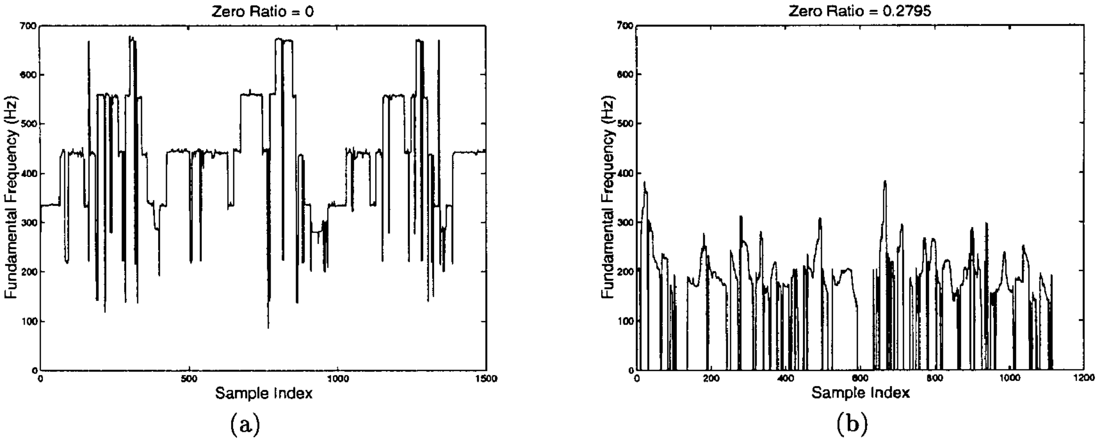


Figure 4: The short-time fundamental frequency curves for (a) trumpet and (b) speech.

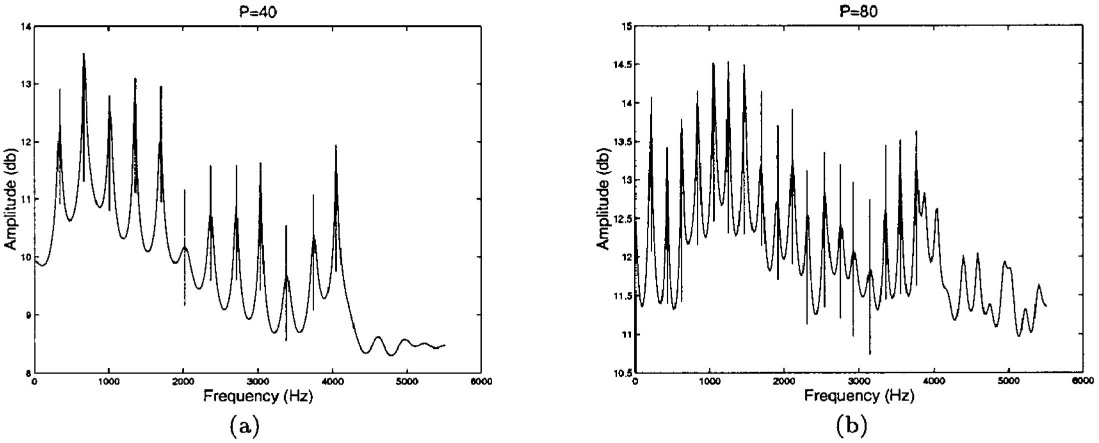


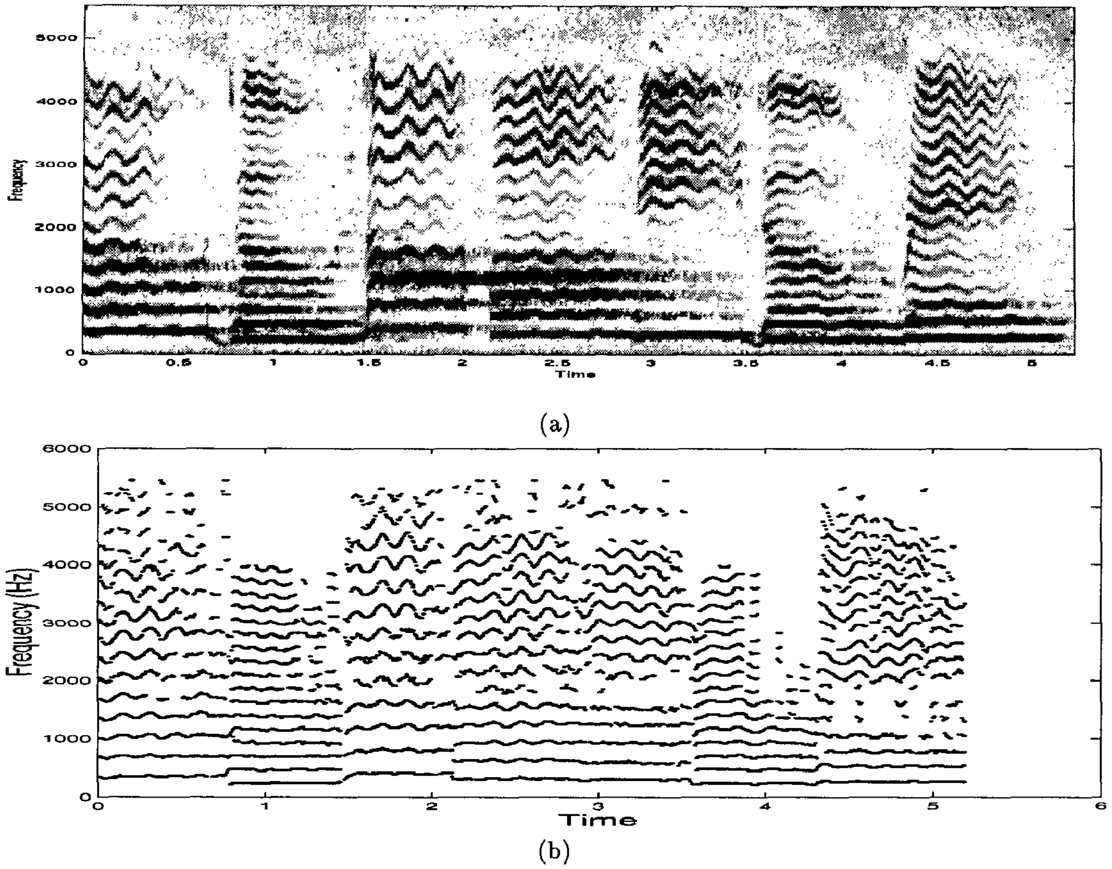
Figure 5: Detecting harmonic peaks from the power spectrum generated by the AR model parameters for song and speech segments: (a) female song with $P = 40$ and (b)female speech with $P = 80$ where P is the order of the AR model.

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.

3.2. Classification of Each Segment

After segmenting boundaries are detected, each segment is classified into one of the basic audio types through the following steps.
(1) Detecting Silence
The first step is to check whether the audio segment is silence or not. We define "silence" to be a segment of imperceptible audio, including unnoticeable noise and very short clicks. The normal way to detect silence is by energy thresholding. However, we have found that the energy level of some noise pieces is not lower than that of some music pieces. The reason that we can hear music while may not notice noise is that the frequency-level of noise is much lower. Thus, we use both energy and ZCR measures to detect silence. If the short-time energy function is continuously lower than a certain set of thresholds, or if most short-time average zero-crossing rates in the segment are lower than a certain set of thresholds, then the segment is indexed as "silence".

(2) Separating Sounds with Music Components
As observed from movies and video programs, music is an important type of audio component frequently appearing, either alone or as the background of speech or environmental sounds. Therefore, we first separate the audio segments into two categories, i.e. with or without music components, mainly by detecting continuous frequency peaks from the power spectrum.



(a)

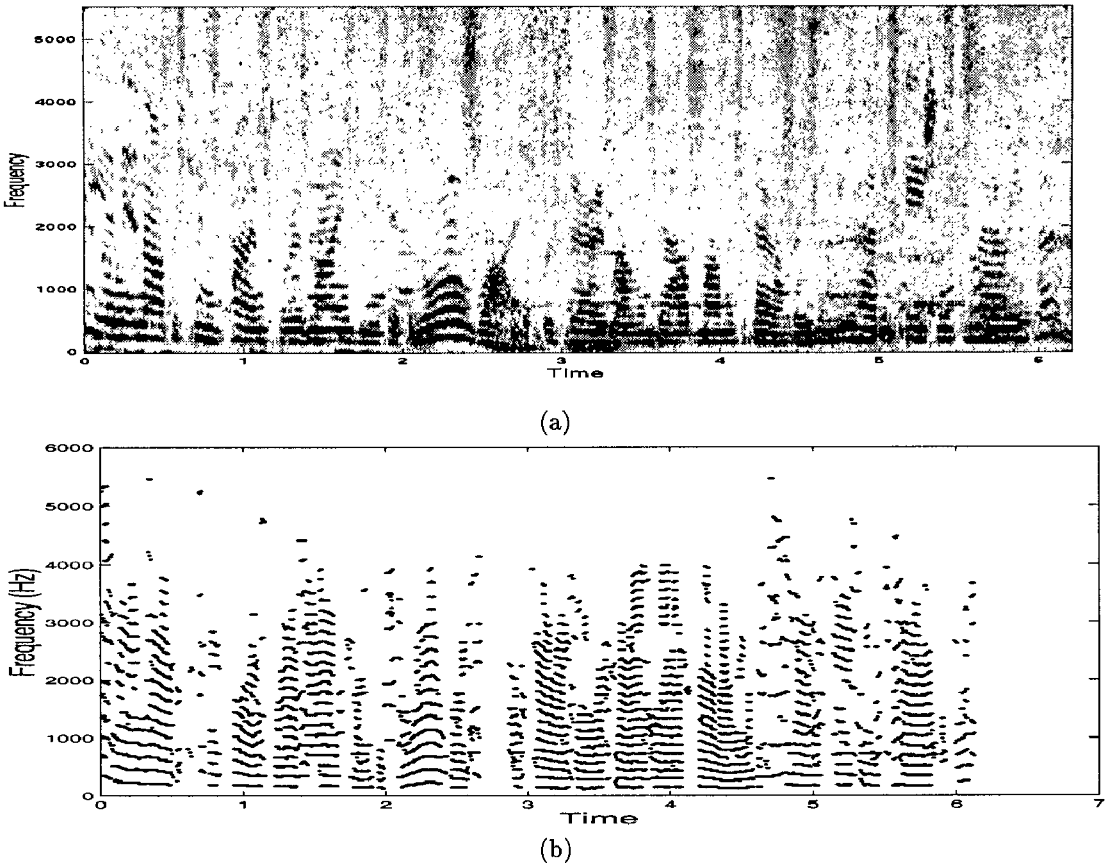
(b)

Figure 6: The spectrogram and spectral peak tracks of female vocal solo.

The power spectrum is generated by an AR model. If there are peaks detected in consecutive power spectra which stay at about the same frequency level for a certain period of time, this period of time is indexed as having music components. An index sequence is generated for each segment of sound, i.e. the index value is set to 1 if the sound is detected as having music components at that instant and to 0, otherwise. The ratio between the number of zeros in the index sequence and the total number of indices in the sequence can thus be a measurement of the sound segment as having music components or not (we call it the "zero ratio"). The higher the ratio is, the less music components are contained in the sound.

We examine the zero ratio of different types of sounds, and summarize our observation below. (1) Speech. Although the speech signal contains many harmonic components, the frequency peaks change faster and last for a shorter time than those of music. The zero ratio for speech segments is normally above 0.95. (2) Environmental Sound. Harmonic and stable environmental sounds are all indexed as having music components, while non-harmonic sounds are indexed as not having music components. (3) Pure Music. The zero ratio for all pure music segments is below 0.3. Indexing errors normally come from short notes, low volume or low frequency parts, non-harmonic components, and the intermissions between two notes. (4) Song. Most song segments have a zero ratio below 0.5. Those parts not detected as having music components result from peak tracks shaped like ripples (instead of lines) when the note is long, intermissions between notes, low volume and/or low frequency sounds. When the ripple-shaped peak tracks are detected and indexed as music components, the corresponding zero ratio for songs are significantly reduced. (5) Speech with Music Background. When the speech is strong, the background music is normally hidden and cannot be detected. However, music components can be detected in intermission periods in speech or when music becomes stronger. We make a distinction of the following two cases. For the first case, when music is stronger or there are many intermissions in speech so that music is a prominent part of the sound, the zero ratio is below 0.6. For the second case, when music is weak while speech is strong and continuous, speech is the major component and music may be ignored. The zero ratio is higher than 0.8 in such a case.

Therefore, based on a threshold for the zero ratio at about 0.7 together with some other rules, audio segments can be separated into two categories as desired. The first category contains harmonic and stable environmental sound, pure music, song, speech with the music background, and environmental sound with the music background. For the second category, there are pure speech and non-harmonic environmental sounds. Further classification will be done



(a)

(b)

Figure 7: The spectrogram and spectral peak tracks of female speech with the background of music and noise.

within each category.

(3) Detecting Harmonic Environmental Sounds
The next step is to separate out environmental sounds which are harmonic and stable. The temporal curve of the short-time fundamental frequency is checked. If most parts of the curve are harmonic, and the fundamental frequency is fixed at one particular value, the segment is indexed as "harmonic and unchanged". A typical example of this type is the sound of touch-tone. If the fundamental frequency of a sound clip changes over time but only with several values, it is indexed as "harmonic and stable". Examples of this type include sounds of the doorbell and the pager.

(4) Distinguishing Pure Music
Pure music is distinguished based on properties of the averaged zero-crossing rate and the fundamental frequency. Four aspects are checked. They are the degree of being harmonic, the degree of the fundamental frequency's concentration on certain values during a period of time, the variance of zero-crossing rates, and the range of amplitudes of zero-crossing rates. For each aspect, there is one empirical threshold set and a decision value defined. If the threshold is satisfied, the decision value is set to 1; otherwise, it is set to a fraction between 0 and 1 according to the distance to the threshold. The four decision values are averaged with predetermined weights to derive a total probability of the audio segment to be pure music. For a segment to be indexed as "pure music", this probability should be above a certain threshold, and at least three of the four decision values should be above 0.5.

(5) Distinguishing Songs
Up to now, what left in the first category are the sound segments of song, speech with the music background and environmental sound with the music background. We extract spectral peak tracks for these segments, and differentiate the three audio types based on the analysis of these tracks. Songs may be characterized by one of the three features: ripple-shaped harmonic peak tracks (due to the vibration of vocal chords), tracks which are of a longer durations compared to those in speech, and tracks which have a fundamental frequency higher than 300Hz. Tracks are checked to see whether any of these three features is matched. The segment will be indexed as "song" if either the sum of durations where harmonic peak tracks satisfy one of the features is above a certain amount, or its comparison to the total length of the segment reaches a certain ratio.

(6) Separating Speech with Music Background and Environmental Sound with Music Background
In speech with the music background, as long as the speech is strong (i.e. the pronunciations are clear and loud enough for human perception), the harmonic peak tracks of the speech signal can be detected in spite of the existence of music components. We check the groups of tracks to see

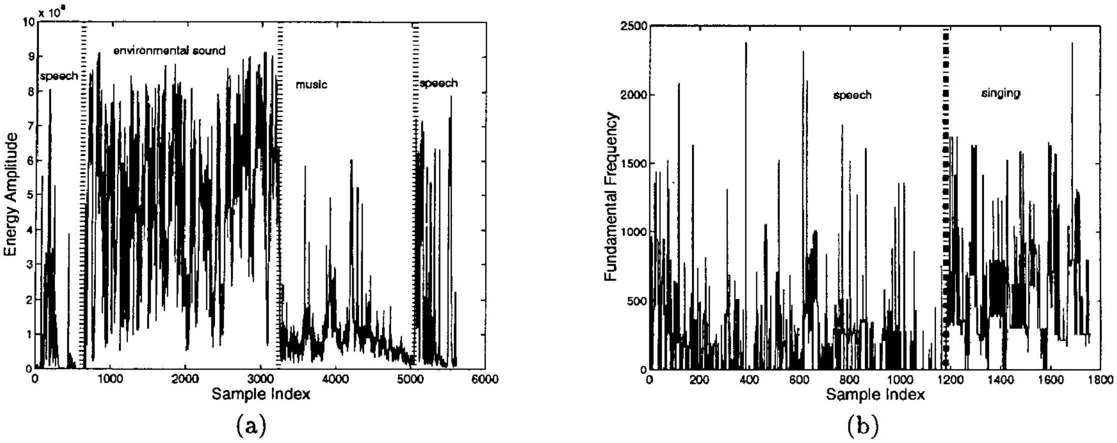


Figure 8: Boundary detection in the temporal curves of (a)energy function and (b)fundamental frequency.

whether they concentrate in the lower to middle frequency bands (with the fundamental frequency between 100 to 300 Hz) and have lengths within a certain range. If there are durations in which the spectral peak tracks satisfy these criteria, the segment is indexed as "speech with music background". An example is shown in Figure 9. Then, what left in the first category will be indexed as "environmental sound with the music background".

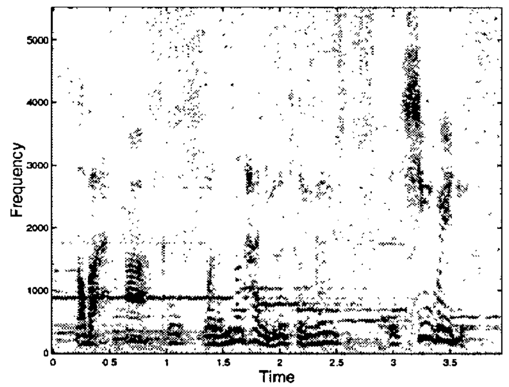


Figure 9: The spectrogram for a segment of speech with the music background.

(7) Distinguishing Pure Speech

When distinguishing pure speech, five conditions are checked. The first one is the relation between temporal curves of the zero-crossing rate and the energy function. In speech segments, the ZCR curve has peaks for unvoiced components and troughs for voiced components, while the energy curve has peaks for voiced components and troughs for unvoiced components. Thus, there is a compensative relation between them. We clip both ZCR and energy curves at one third of the maximum amplitude and remove the lower parts, so that only peaks of the two curves will remain. Then, the inner product of the two residual curves is calculated. This product is normally near zero for speech segments because peaks appear at different times in the two curves, while the product value is much larger for other types of audio. The second aspect is the shape of the ZCR curve. For speech, the ZCR curve has a stable and low baseline with peaks above it. The baseline is defined as the linking line of lowest points of troughs in the curve. The mean and the variance of the baseline are calculated. The parameters and the frequency of appearance of peaks are also considered. The third and fourth aspects are the variance and the range of amplitudes of the ZCR curve, respectively. Contrary to music segments where the variance and the range of amplitudes are normally lower than certain thresholds, a typical speech segment has a variance and a range of amplitudes that are higher than certain thresholds. The fifth aspect is related to the property of the short-time fundamental frequency. As voiced components are harmonic and unvoiced components are non-harmonic, speech has a percentage of harmony within a certain range. There is also a relation between the fundamental frequency curve and the energy curve. That is, the harmonic parts in the FuF curve correspond to peaks in the energy curve while the zero parts in the FuF curve correspond to troughs in the energy curve.

A decision value, which is a fraction between 0 and 1, is defined for each of the five conditions. The weighted average of these decision values represent the possibility of the segment's being speech.

(8) Classifying Non-harmonic Environmental Sounds

The last step is to classify what left in the second category into one type of non-harmonic environmental sounds as the following. We apply the following four rules. (1) If either the energy function curve or the average zero-crossing rate curve has peaks which have approximately equal intervals between neighboring peaks, the segment is indexed as "periodic or quasi-periodic". Examples for this type include sounds of clock tick and the regular footstep. (2) If the percentage of harmonic parts in the fundamental frequency curve is within a certain range (lower than the threshold for music, but higher than the threshold for non-harmonic sound), the segment is indexed as "harmonic and non-harmonic mixed". For example, the sound of train horn, which is harmonic, appears with a non-harmonic background. (3) If the frequency centroid (denoted by the average zero-crossing rate value) is within a relatively small range compared to the absolute range of the frequency distribution, the segment is indexed as "non-harmonic and stable". One example is the sound of birds' cry, which is nonharmonic while its ZCR curve is concentrated within the range of 80-120. (4) If the segment does not satisfy any of the above conditions, it is indexed as "non-harmonic and irregular". Many environmental sounds belong to this type such as the sounds of thunder, earthquake and fire.

3.3. Post-Processing

The post-processing step is to reduce possible segmentation errors. We have adjusted the segmentation algorithm to be sensitive enough to detect all abrupt changes. Thus, it is possible that one continuous scene is broken into several segments. In the post-processing step, small pieces of segments are merged with neighboring segments according to certain rules. For example, one music piece may be broken into several segments due to abrupt changes in the energy curve, and some small segments may even be misclassified as "harmonic and stable environmental sound" because of the unchanged tune in the segment. With post-processing, these segments can be combined into one segment reindexed based on its contextual relation.

4. EXPERIMENTAL RESULTS

We have built a generic audio database as the testbed of the proposed algorithms. It includes around 1500 audio clips of various types. The short pieces of sound clips (with duration from several seconds to one minute) are used to test the classification accuracy. We have also collected dozens of longer audio clips recorded from movies. These pieces last from several minutes to half an hour, and contain different types of audio. They are used to test the segmentation and indexing performances.

4.1. Classification Results

The proposed classification approach for generic audio data achieved an accuracy rate of more than 90% by using a set of 1200 audio pieces including all types of sound selected from the audio database described above. A demonstration program was made for on-line classification, which shows the waveform, the audio features, and the classification result for a given sound, as illustrated in Figure 10.

Misclassifications used to occur in hybrid sounds which contain more than one basic type of audio. After these types of sounds (e.g. song, the speech with a music background and the environmental sound with a music background) are separated out, such errors have been significantly reduced. Now, major mistakes result from the very noisy background in some speech and music segments. Actually, our approach is normally robust in distinguishing speech and music signals with rather low SNR. We will further improve our algorithm so that the speech or the music segment can be detected as long as its content can be recognized by human being. When SNR is too low, and environmental sounds are actually dominant, our algorithm will classify the segment into a proper type of environmental sound.

4.2. Segmentation and Indexing Results

We tested the segmentation procedure with audio clips recorded from movies and TV programs. With Pentium333 PC/Windows NT, segmentation and classification tasks can be completed together with less than one eighth of the time required to play the audio clip. We made a demonstration program for on-line audiovisual data segmentation and indexing as shown in Figure 11, where different types of audio are represented by different colors. Displayed in this figure is the segmentation and indexing result for an audio clip recorded from the movie "Washington Square". In this 50-second long audio clip, there is first a segment of speech spoken by a female (indexed as "pure speech"), then a segment of screams by a group of people (indexed as "non-harmonic and irregular" environmental sound), followed by a period of unrecognizable conversation of multi-people simultaneously mixed with baby cry (indexed as the mix of harmonic and non-harmonic sounds). Then, a low volume music appears in the background (indexed as "environmental sound with music background"). Afterwards, there is a segment of music with very low level environmental sounds in the background (indexed as "pure music"). Finally, there is a short conversation between a male and a female (indexed as "pure speech").

In the above example as well as many other experiments made, boundaries between segments of different audio types were set very precisely and each segment was accurately classified.

5. CONCLUSION AND FUTURE WORK

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 boudaries were set accurately, and a correct classification rate higher than 90% was achieved.

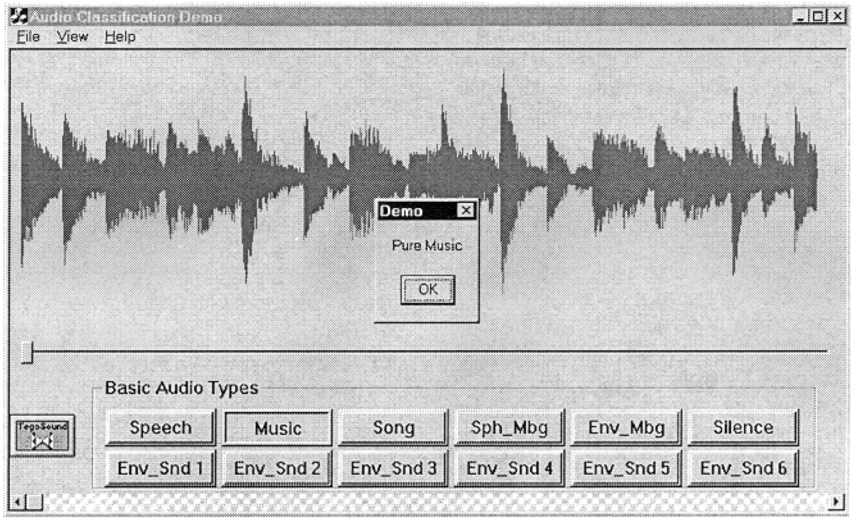
Further research can be done in two areas. One is audio feature extraction in the compressed domain such as MPEG bitstreams. The other is the integration of audio features with visual and textual information to achieve superior video segmentation and indexing performances.

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.

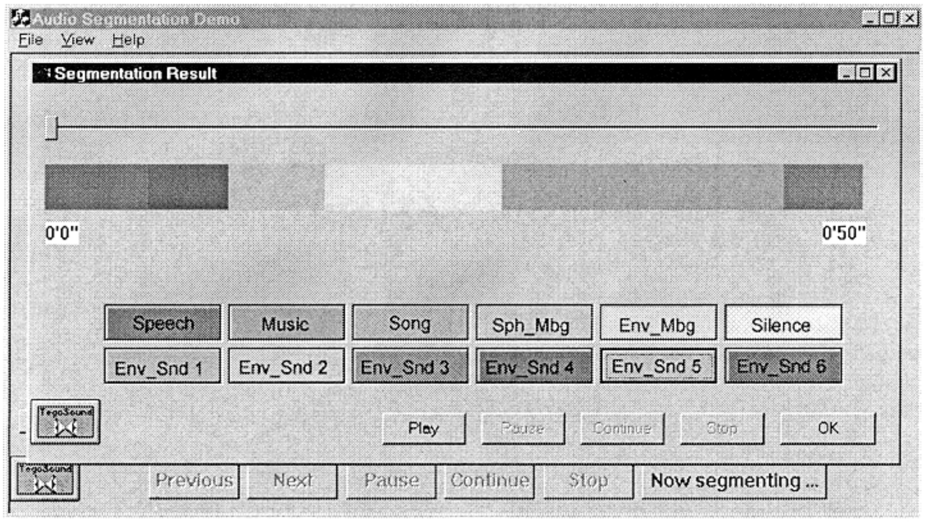
[2] Foote, J.: Content-based retrieval of music and audio, in Proceedings of SPIE'97, Dallas, 1997.

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



Pure Music

Figure 10: Demonstration of generic audio data classification.



Now segmenting ...

Figure 11: Demonstration of audiovisual data segmentation.

[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 on Image Processing, Chicago, Oct. 1998.

[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, vol.2670, pp.373-384, San Jose, 1996.

[8] Saunders, J.: Real-time discrimination of broadcast speech/music, in Proceedings of ICASSP'96, vol.II, pp.993-996, May 1996.

[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.

[10] Wold, E., Blum, T. and Keislar, D. *et al.*: Content-based classification, search, and retrieval of audio, IEEE Multimedia, pp.27-36, Fall, 1996.

[11] 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. 1995.