Audio content analysis for online audiovisual data segmentation and classification

Article *in* IEEE Transactions on Speech and Audio Processing · June 2001

DOI: 10.1109/89.917689 · Source: IEEE Xplore

C.-C. Jay Kuo
University of Southern California

Audio Content Analysis for Online Audiovisual Data Segmentation and Classification

Tong Zhang, *Member, IEEE*, and C.-C. Jay Kuo, *Fellow, IEEE*

***Abstract*—While current approaches for audiovisual data segmentation and classification are mostly focused on visual cues, audio signals may actually play a more important role in content parsing for many applications. An approach to automatic segmentation and classification of audiovisual data based on audio content analysis is proposed. The audio signal from movies or TV programs is segmented and classified into basic types such as speech, music, song, environmental sound, speech with music background, environmental sound with music background, silence, etc. Simple audio features including the energy function, the average zero-crossing rate, the fundamental frequency, and the spectral peak tracks are extracted to ensure the feasibility of real-time processing. A heuristic rule-based procedure is proposed to segment and classify audio signals and built upon morphological and statistical analysis of the time-varying functions of these audio features. Experimental results show that the proposed scheme achieves an accuracy rate of more than 90% in audio classification.**

***Index Terms*—Audio analysis, audio indexing, audio segmentation, audiovisual content parsing, information filtering and retrieval, multimedia database management.**

I. Introduction

THE task of automatic segmentation, indexing, and retrieval of audiovisual data has important applications in professional media production, audiovisual archive management, education, entertainment, surveillance, and so on. For example, a vast amount of audiovisual material has been archived in television and film databases. If these data can be properly segmented and indexed, it will facilitate the retrieval of desired video segments for the editing of a documentary or an advertisement video clip. To give another example, in audiovisual libraries or family entertainment applications, it will be convenient to users if they are able to retrieve and watch video segments of their interest. As the volume of the available material becomes huge, manual segmentation and indexing is impossible. Automatic segmentation and indexing through computer processing based on multimedia content analysis is clearly the trend.

Current approaches for audiovisual data segmentation and indexing are mostly focused on visual cues such as color histogram differences, motion vectors, and keyframes [1]–[3]. In contrast, the accompanying audio signal receives relatively little attention. There is however a significant amount of information contained in the continuous flow of audio data which may often represent the theme in a simpler fashion than the pictorial part. For instance, all video scenes of gun fight should include the sound of shooting or explosion, while the image content may vary a lot from one video clip to another. In the beginning of the movie "Washington Square," there is a segment which is of several minutes long, showing buildings, streets, and people of a neighborhood. There are many different shots involved, but the continuous accompanying music indicates that they are actually within one audio scene. Moreover, the speech information contained in audio signals is usually critical in identifying the theme of the video segment. By only listening to the dialog in a segment, it is usually enough for us to understand what it is about. However, a viewer can be easily lost by watching pictures only. Thus, it is fair to say that the audio signal may actually play a primary role in content parsing of audiovisual data.

We have been working on the integration of audio and visual information for online video indexing and annotation. The first step is to conduct a segmentation of the video sequence into semantic scenes based on audio content analysis. We call such a segmented unit as "audio scene," and index it as pure speech, pure music, song, speech with music background, environmental sound with music background, silence, etc. based on our audio classification algorithms. Then, further segmentation of audio scenes into visual shots will be done according to visual cues, and keyframes will be extracted from each shot to give the visual index. The combination of audio and visual indexing should provide a great help to users in retrieving and browsing audiovisual segments of their interest from a movie or a TV program. For example, to retrieve "segments of songs performed by Michael Jackson" may be achieved by searching for audio index of "song" and keyframes of "Michael Jackson."

In this paper, we focus on the problem of segmenting and classifying accompanying audio signals in audiovisual data based on audio content analysis. The paper is organized as follows. Existing work on audio content analysis is reviewed in Section II. An overview of the proposed system and major contributions of this research is presented in Section III. The computation and properties of audio features used in this work are analyzed in Section IV. The proposed procedures for the segmentation and indexing of audio stream are described in Section V. Experimental results are shown in Section VI. Finally, concluding remarks and future research plans are given in Section VII.

II. Previous Work on Audio Content Analysis

Existing work on audio content analysis is quite limited and still at a preliminary stage. Current researches can be generally categorized into the following four directions.

Manuscript received August 16, 1999; revised December 5, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wu Chou.

The authors are with the Integrated Media Systems Center and Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, CA 90089-2564 USA (e-mail: cckuo@sipi.usc.edu).

Publisher Item Identifier S 1063-6676(01)02743-2.

1) Audio Segmentation and Classification: One basic problem in audio segmentation and classification is the discrimination between speech and music, since they are the two most important types of audio. The approach presented by Saunders [4] used only the average zero-crossing rate and the energy features, and applied a simple thresholding procedure while Scheirer and Slaney [5] proposed to use thirteen features in the time, frequency, and cepstrum domains, as well as model based classification methods (MAP, GMM, kNN, etc.) to achieve a robust performance. Both approaches reported real-time discrimination of an accuracy rate over 90%. 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. Further classification of audio data may take other sounds, besides speech and music, into consideration. Wyse and Smoliar [6] worked on the classification of audio signals into "music," "speech," and "others." In their work, music was first detected based on the average length of time in which peaks exist in a narrow frequency region. Then, speech was separated out by pitch tracking. This method was developed for the parsing of news stories. An acoustic segmentation approach was also proposed by Kimber and Wilcox [7], where audio recordings were segmented into speech, silence, laughter and nonspeech sounds. They used cepstral coefficients as features and the hidden Markov model (HMM) as the classifier. The method was mainly applied to the segmentation of discussion recordings in meetings. Research by Pfeiffer *et al.* [8] aimed at the analysis of the amplitude, frequency and pitch of audio signals, as well as the simulation of human audio perception so that results may be used to segment audio data streams and to recognize music. These features were also used to detect sounds of shot, cry and explosion which might indicate violence.

2) Content-Based Audio Retrieval: One specific technique in content-based audio retrieval is query-by-humming, through which a song is retrieved by humming the tune of it. A typical system was presented by Ghias *et al.* [9] for this purpose. Foote [10] proposed a music and sound effect retrieval system, where the Mel-frequency cepstral coefficients (MFCC) were taken as features, and a tree-structured classifier was built for retrieval. Since MFCC can not represent the timbre of sounds properly, this method in general fails to distinguish music and environmental sounds with different timbre characters. In the content-based retrieval (CBR) work of Wold *et al.* [11], statistical values (including means, variances, and autocorrelations) of several time- and frequency-domain measurements were used to represent perceptual features like loudness, brightness, bandwidth, and pitch. Since merely statistical values were considered, this method was only suitable for sounds with a single timbre. An audio retrieval method was proposed by Smith *et al.* [12] for searching quickly through broadcast audio to detect and locate sound segments containing a certain reference template based on an active search algorithm and histogram modeling of zero-crossing features. The exact audio segment to be searched should be known *a priori* in this algorithm.

3) Audio Analysis for Video Indexing: In [13] and [14], Liu *et al.* applied audio analysis results to the distinction of five different video scenes: news report, weather report, basketball game, football game and advertisement. The adopted features included the silence ratio, the speech ratio and the subband energy ratio which were extracted from the volume distribution, the pitch contour and the frequency domain, respectively. The multilayer neural network (MNN) and the hidden Markov model (HMM) were used as classifiers. It was shown that, when using MNN, the method worked well in distinguishing among reports, games and advertisements, but had difficulty in classifying the two different types of reports and the two different kinds of games. While using HMM, the overall accuracy rate increased, but there were misclassifications among all the five sorts of scenes. Liu and Huang [15] also applied the same set of audio features in distinguishing news reports from commercials and music in broadcast news programs. A simple hard threshold classifier and a fuzzy classifier were used. Patel and Sethi [16] proposed to perform audio characterization on MPEG compressed data (actually, the subband level data) for the purpose of video indexing. The audio signal was classified into dialog, nondialog and silence intervals. Features were taken from the energy, the pitch, the spectrogram and the pause rate domains, and organized in a thresholding procedure. There were somehow quite a few mistakes occurring in the classification between dialog and nondialog intervals. An approach to video indexing through music and speech detection was proposed by Minami *et al.* [17], where image processing techniques were exploited to analyze the spectrogram of audio signals. Spectral peaks of music were recognized by applying an edge-detection operator, and speech harmonics were detected with a comb filter. They also presented two application systems to demonstrate the indexing method. One system allowed users to access any frame of video randomly while the other created condensations of dramas or movies by excerpting meaningful video segments based on the locations of music and speech.

4) Integration of Audio and Visual Information for Video Segmentation and Indexing: A new trend for video segmentation and indexing is to combine audio and visual information in one framework. This idea was reflected in three recent papers. However, all audio features adopted were quite primitive ones, and no delicate procedure of audio feature extraction for the specific purpose was considered up to now. In the method proposed by Huang *et al.* [18], the same set of audio features as used in [13] were combined with the color and motion information to detect scene and shot breaks. In the approach presented by Naphade *et al.* [19], subband audio data and color histograms of one video segment were integrated to form a "Multiject," and two variations of the hidden Markov model were used to index Multijects. Experimental results of detecting the events of "explosion" and "waterfall" were reported. In the approach of Boreczky and Wilcox [20], color histogram differences, cepstral coefficients of audio data, and motion vectors were used together with a hidden Markov model approach to segment video into regions defined by shots, shot boundaries and camera movement within shots.

Another research field which is quite important for audio content analysis is the *Audio Scene Analysis (ASA)*, which was named after the classic work of Bregman [21]. The goal of this field is to understand the way the auditory system and the

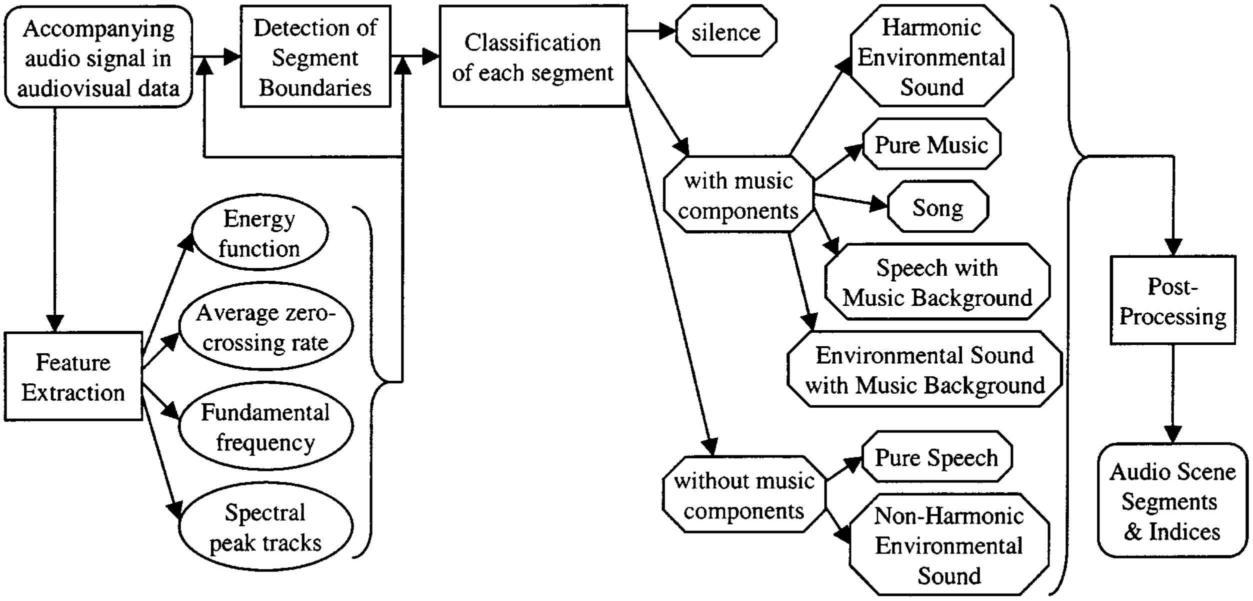

Accompanying audio signal in audiovisual data
Detection of Segment Boundaries
Classification of each segment
silence
Harmonic Environmental Sound
Pure Music
Song
Speech with Music Background
Environmental Sound with Music Background
with music components
without music components
Pure Speech
Non-Harmonic Environmental Sound
Post-Processing
Audio Scene Segments & Indices
Feature Extraction
Energy function
Average zero-crossing rate
Fundamental frequency
Spectral peak tracks

Fig. 1. Automatic segmentation and indexing of audiovisual data based on audio content analysis.

brain of human beings process complex sound environments, where multiple sources that change independently over time are present. Brown and Cooke [22] termed the research area of constructing computer models to perform auditory source segregation as computational audio scene analysis (CASA). One example is the work by Weintraub [23] who used a dynamic programming framework around Licklider's autocorrelation model to separate voices of two speakers whose voices interfere in a single recording. Another example is the system built by Ellis [24], which aimed to analyze the sound and segregate perceptual components from noisy sound mixtures such as a "city-street ambience." The structured audio in MPEG-4 unifies many ideas and efforts in this field and provides semantic and symbolic descriptions of audio (the decoder is standardized while mature techniques for the encoder are still to be developed in the coming years). A summarization of this work was given by Vercoe *et al.* in [25]. This technique is useful for ultra low-bit-rate transmission, flexible synthesis, and perceptually based manipulation and retrieval of sounds.

III. Overview of the Proposed System

In this research, we propose a scheme for the automatic segmentation and annotation of audiovisual data based on audio content analysis. Four kinds of audio features are extracted, namely, the short-time energy function, the short-time average zero-crossing rate, the short-time fundamental frequency and the spectral peak tracks. We perform the morphological and statistical analysis of temporal curves of these features to reveal differences among different types of audio. A rule-based heuristic procedure is then built to segment and classify audio signals with these features. The flowchart of this procedure is illustrated in Fig. 1.

Segment boundaries are first detected by locating abrupt changes in these short-time features. Then, each segment is classified to be one of the basic audio types. Silent segments are distinguished, and nonsilent sounds are separated into two categories, i.e., with or without music components by detecting continuous frequency peaks from the power spectrum of audio signal. Sound segments in the first category are further classified to be harmonic environmental sound, pure music, song, speech with music background, or environmental sound with music background based on the analysis of audio features. Sound segments in the second category are indexed as pure speech or one type of the nonharmonic environmental sound. Finally, a postprocessing step is applied for reducing possible segmentation errors.

Compared with previous work, there are several distinguishing features in the proposed scheme. First, besides commonly studied audio types such as speech and music in existing work, we have taken into account hybrid types of sound which contain more than one kind of audio component. For example, the speech signal with music background and the singing of a person are two types of hybrid sound which have characters of both speech and music. We are able to put these two kinds of sound in different categories with the proposed scheme, and their distinction is important in characterizing audiovisual segments. For example, in documentaries or commercials, there is often a musical background with speech of commentary appearing from time to time. It is also common that clients want to retrieve a segment of video, in which there is singing of one particular song. There are other kinds of hybrid sound included in our system, e.g., speech or music with environmental sounds as the background (where the environmental sounds may be treated as noise), and environmental sounds with music as the background.

Second, we put more emphasis on the distinction of environmental sounds which are often ignored in previous work. Environmental sounds, including sound effects, are an important ingredient in audiovisual recordings, and their analysis is essential in many applications such as the post-processing of films. In our scheme, we separate environmental sounds into six categories

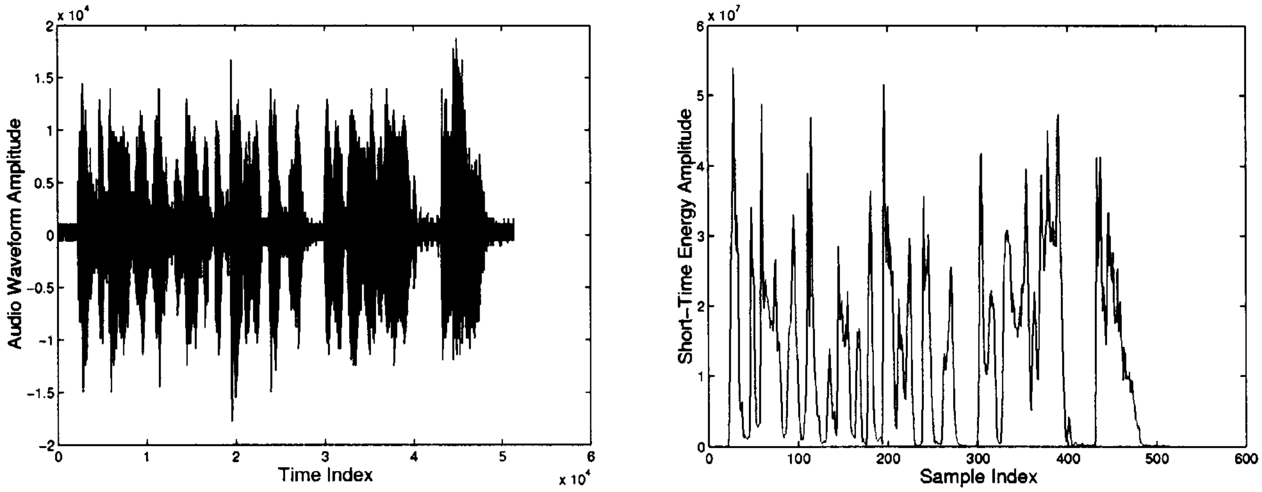


Fig. 2. Audio waveform and the short-time energy function of a speech segment.

according to their harmony, periodicity, or stability properties. There are "harmonic and fixed," "harmonic and stable," "periodic or quasiperiodic," "harmonic and nonharmonic mixed," "nonharmonic and stable," and "nonharmonic and irregular" environmental sounds.

Third, integrated features are exploited for audio classification. For example, short-time features of the energy, the average zero-crossing rate and the fundamental frequency are effectively combined in distinguishing speech, music and silence. We use not only the feature values, but also their change patterns over the time and the relationships among the three kinds of features. We also propose a method to extract spectral peak tracks, and use this feature specifically for the distinction of sound segments of songs and speech with music background. Furthermore, signal processing techniques are applied for the representation and classification of the extracted features, including morphological and statistical analysis, the heuristic method, adaptive search, and fuzzy logic.

Fourth, although the proposed scheme covers a wide range of audio types, the complexity is low since selected audio features are easy to compute and the rule-based indexing procedure is fast. Most audio features used in this system are short-time and one-dimensional, which makes online audiovisual data processing feasible. Among the three short-time features, the fundamental frequency is the most expensive in computation, which only requires one 512-point FFT per 100 input samples. The spectral peak tracking requires a little bit more calculation, but it only has to be computed under certain conditions.

Finally, the proposed audio segmentation and classification approach is based on the observation of different types of audio signals and their physical features, which is generic and model-free. Consequently, it can be easily applied, as the first step processing of digital audiovisual data, to almost any content-based audiovisual material management system. For example, it may be used as the tool for online segmentation and indexing of radio and TV programs. An index table can be generated automatically for each program, and the user is able to choose certain segments (e.g., those of pure music) to browse. Especially, the inclusion of a keyframe for each segment in TV programs will facilitate the retrieval task.

IV. Audio Feature Analysis

A. 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

$x(m)$ discrete time audio signal;

n time index of the short-time energy;

$w(m)$ rectangle window of length N.

It provides a convenient representation of the amplitude variation over time. By assuming that the audio signal changes relatively slowly within a small interval, we calculate E_n once every 100 samples at an input sampling rate of 11 025 samples/s. We set the window duration of $w(m)$ to be 150 samples so that there is an overlap between neighboring frames. The audio waveform of a speech segment and the temporal curve of its short-time energy function are shown in Fig. 2. Note that the sample index of the energy curve is at the ratio of 1 : 100 compared to the corresponding time index of the audio signal.

Major reasons for using the short-time energy feature in our work include

1) for speech signals, it provides a basis for distinguishing voiced speech components from unvoiced speech components because values of E_n for the unvoiced components are in general significantly smaller than those of the voiced components, as can be seen from the peaks and troughs in the energy curve;
2) it can be used as the measurement to distinguish audible sounds from silence when the SNR is high;
3) its change pattern over time may reveal the rhythm and periodicity properties of sound.

B. Short-Time Average Zero-Crossing Rate

For discrete-time signals, a zero-crossing is said to occur if successive samples have different signs. The rate at which zero-

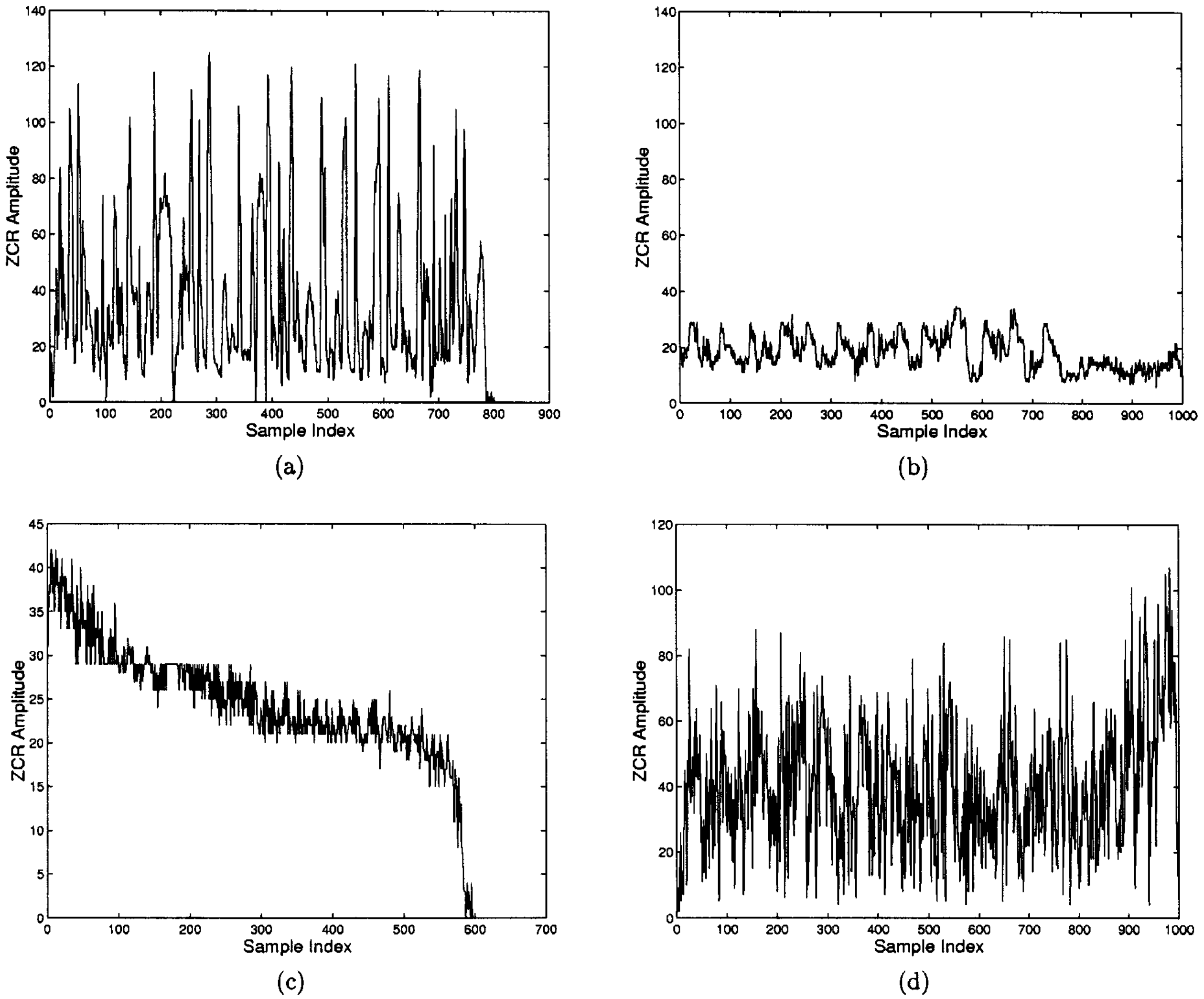


Fig. 3. Short-time average zero-crossing rates of four audio signals: (a) speech, (b) piano, (c) chime, and (d) footstep.

crossings occur is a simple measure of the frequency content of a signal. The short-time average zero-crossing rate is defined as

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

where

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

and $w(n)$ is a rectangle window. Temporal curves of the short-time average zero-crossing rate (ZCR) for several audio samples are shown in Fig. 3. Similar to the computation of the short-time energy function, we also choose to compute the ZCR value at every 100 input samples, and set the window width to 150 samples.

The average zero-crossing rate can be used as another measure to distinguish between voiced and unvoiced speech signals, because unvoiced speech components normally have much higher ZCR values than voiced ones [26]. As shown in Fig. 3(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 amplitude for the ZCR curve. Note also that the ZCR curve has a relatively low and stable baseline with high peaks above it. Comparatively, the ZCR curve of music plotted in Fig. 3(b) has a much lower variance and average amplitude, suggesting that the zero-crossing rate of music is normally much more stable during a certain period of time. ZCR curves of music generally have irregular waveforms with a changing baseline and a relatively small range of amplitude. Since environmental audio consists of sounds of various origins, their ZCR curves can have very different properties. For example, the zero-crossing rate of the sound of chime as shown in Fig. 3(c) reveals a continuous drop of the frequency centroid over time while that of the footstep sound in Fig. 3(d) is rather irregular. We may briefly classify environmental sounds according to their ZCR curve properties such as regularity, periodicity, stability and range of amplitude.

C. 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 nonharmonic sounds. The spectra of sounds generated by trumpet and applause are illustrated in Fig. 4. It is clear that the former one is harmonic while the latter one is nonharmonic.

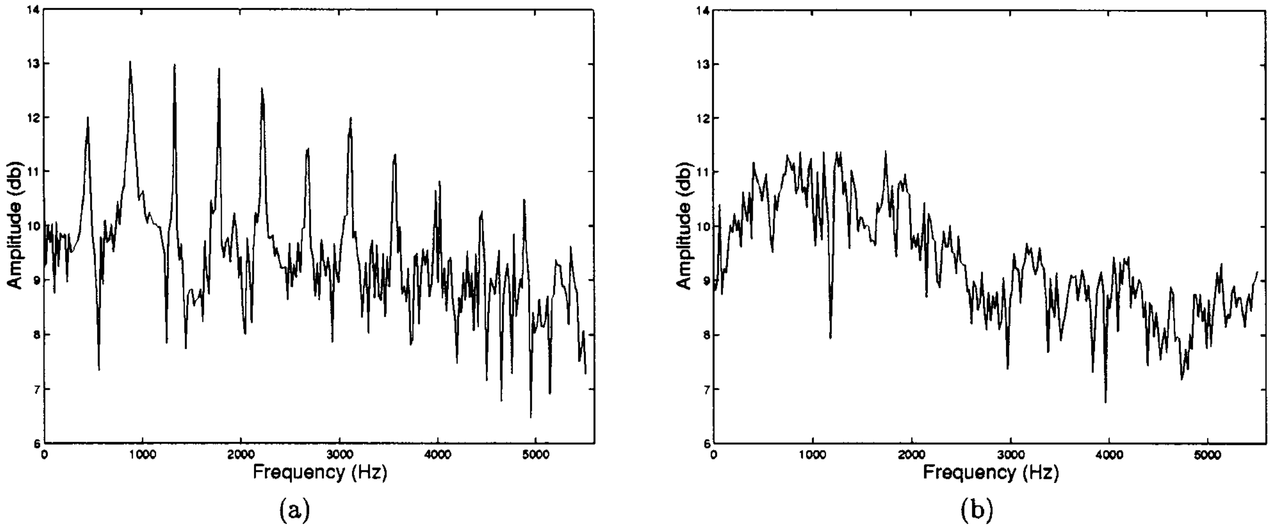


Fig. 4. Spectra of harmonic and nonharmonic sound computed directly with FFT: (a) trumpet and (b) applause.

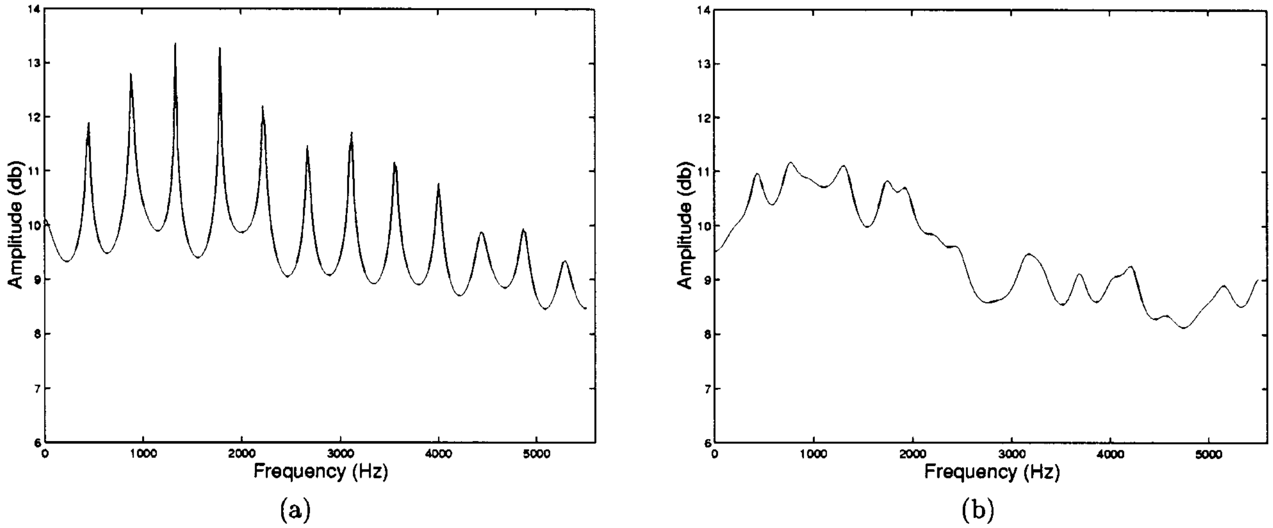


Fig. 5. Spectra of harmonic and nonharmonic sound generated with the AR model: (a) trumpet and (b) applause.

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 nonharmonic mixed sound, since voiced components are harmonic while unvoiced components are nonharmonic. Most environmental sounds are nonharmonic, such as the sounds of applause, footstep, and explosion. However, there are also examples of sound effect which are harmonic and stable, such as the sounds of doorbell and touch-tone; and those which are harmonic and nonharmonic mixed like laughter and dog bark.

In order to measure the harmony feature of sound, we define the short-time fundamental frequency (SFuF) as such: when the sound is harmonic, the SFuF value is equal to the fundamental frequency estimated at that instant; and when the sound is nonharmonic, the SFuF value is set to zero. Although there are many schemes proposed for fundamental frequency estimation or pitch detection in speech and music analysis [26]–[29] (it is worthwhile to point out that the fundamental frequency is a physical measurement while the pitch is rather a perceptual term [30]), none of them is perfectly satisfactory for a wide range of audio signals. As our primary purpose of estimating the fundamental frequency is to determine the harmonic property for all kinds of audio signals, we tend to develop a method which is efficient and robust, but not necessarily perfectly precise. In this work, the short-time fundamental frequency is calculated based on peak detection from the spectrum of 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. Comparing the spectra shown in Fig. 5, which were generated with the AR model with those computed directly from the FFT of audio signals as shown in Fig. 4, we can see that detecting peaks associated with the harmonic frequencies is much easier in the AR generated spectrum than in the directly computed one. We choose the order of the AR model to be 40. With this order, harmonic peaks are remarkable while there are also nonharmonic peaks appearing. However, compared with harmonic peaks, nonharmonic ones not only lack a precise harmonic re-

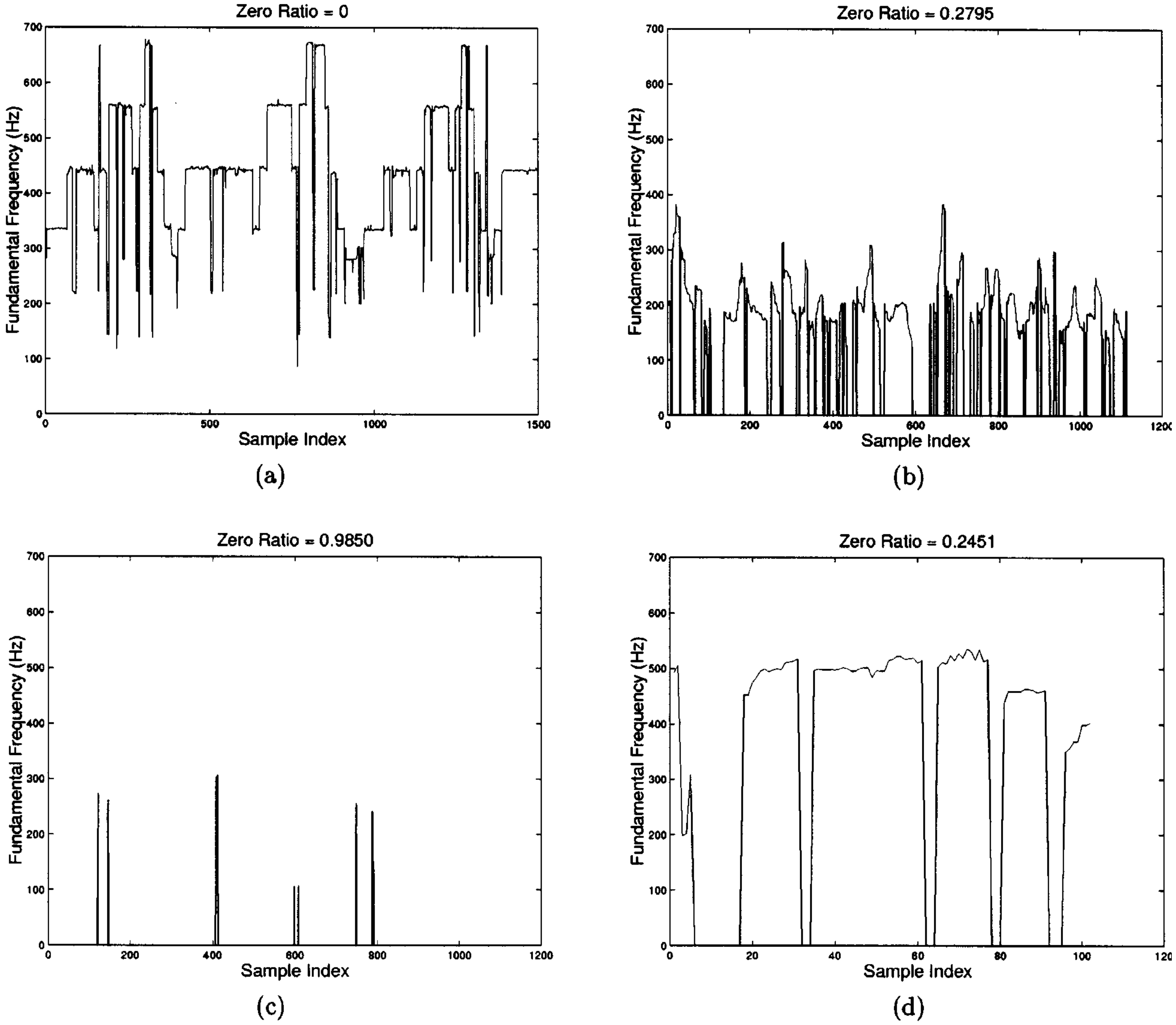


Fig. 6. Short-time fundamental frequency of audio signals: (a) trumpet, (b) speech, (c) rain, and (d) laugh.

lation among them, but also appear to be less sharp at the maximum and of smaller amplitude (i.e., the maximum to minimum distance of the peak) which is clearly observed in Fig. 5. Thus, for a sound to be regarded as harmonic, there should be the greatest-common-divider relation among peaks, and some of the peaks should be sharp and high enough.

All maxima in the spectrum are detected as potential harmonic peaks, and the amplitude, the width and the sharpness of each peak are calculated using morphological analysis. It is checked among locations of these peaks whether a certain amount of them have a common divider and at least some of them have sharpness, amplitude and width values satisfying certain criteria. If all conditions are met, the SFuF value is estimated as the frequency corresponding to the greatest common divider of locations of harmonic peaks; otherwise, SFuF is set to zero. SFuF is computed once every 100 input samples. After the temporal curve of SFuF is obtained for a segment of a certain length, there is a postprocessing step in which singular points in the temporal curve of SFuF are removed to improve the accuracy of the SFuF estimation.

Illustrated in Fig. 6 are examples of SFuF curves of sounds. Shown on the top of each picture is the "zero ratio" of the SFuF curve for that sound segment, which is defined as the ratio between the number of samples with a zero SFuF value (i.e., nonharmonic sound) and the total number of samples in the curve. One can see that music is in general continuously harmonic. Also, the fundamental frequency usually changes more slowly than that of other kinds of sounds, and the SFuF value tends to concentrate on certain frequency for a short period of time. Harmonic and nonharmonic components appear alternately in the SFuF curve of the speech signal, since voiced components are harmonic and unvoiced components are nonharmonic. The fundamental frequency of voiced components is normally in the range of 100–300 Hz. Most environmental sounds are nonharmonic with zero ratios over 0.9 such as the sound of rain. An instance of harmonic and nonharmonic mixed sound effects is the sound of laughing, in which voiced segments are harmonic, while intermissions in between as well as transitional parts are nonharmonic. It has a zero ratio of 0.25 which is similar to that of the speech segment.

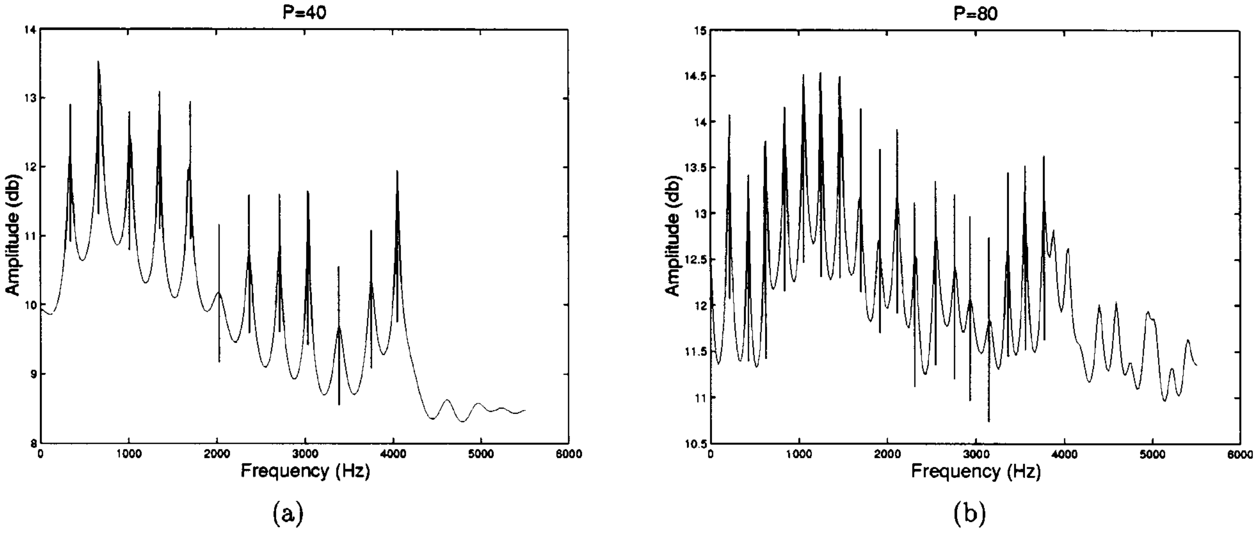


Fig. 7. Detecting harmonic peaks from power spectrum generated with the AR model parameters for song and speech segments: (a) female song with $P = 40$ and (b) female speech with $P = 80$. P is order of the AR model.

D. Spectral Peak Track

Peak tracks in the spectrogram of an audio signal often reveal characteristics of the type of 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 which align tidily in the shape of a comb. Spectral peak tracks in song segments may exist in a broad range of frequency bands, and the fundamental frequency ranges from 87 Hz to 784 Hz. There are relatively long tracks in songs which are stable because the voice may stay 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 segments normally lie in lower frequency bands, and are more close to each other due to the fundamental frequency range of 100–300 Hz. They also tend to be of shorter length because there are intermissions between voiced syllables, and may fluctuate slowly because the pitch may change during the pronunciation of certain syllables.

In this work, we extract spectral peak tracks for the purpose of characterizing sounds of song and speech. Basically, it is done by detecting peaks in the power spectrum generated by the AR model parameters and checking harmonic relations among the peaks. The range of fundamental frequency of harmonic peaks under consideration is set to 80 Hz–800 Hz due to the property of song and speech. With a 512-point FFT, the frequency resolution should be enough to detect harmonic peaks for such a range if the order of the AR model is chosen properly. For example, when $P = 40$, harmonic peaks with a fundamental frequency higher than 250 Hz can be easily detected, which fits for most song segments. However, this resolution is not enough for most male and female speech segments. By experiments, we found that $P = 80$ was normally suitable for female speech signals (with a pitch at about 150–250 Hz), and male speech signals might require an order of $P = 100$ when the pitch is between 100–150 Hz. Nevertheless, with these higher values of P, artifact peaks will appear in the estimated spectra of sounds having higher fundamental frequencies, and may severely impair the quality of peak detection in these sounds.

We currently fix the order of AR model at three levels: 40, 80 and 100. The idea is that it should be able to detect harmonic peaks with one of these orders for sounds of concern. The procedure to determine the proper order is stated below. If, in the previous frame of an audio signal, harmonic peaks were detected from the power spectrum generated with the AR model of order P_1 (P_1 may be 40, 80, or 100), we begin to detect harmonic peaks for the current frame with the spectrum of order P_1. If harmonic peaks are found in this spectrum, we go on to the next frame. Otherwise, we try the spectra generated with the other two order levels. If no harmonic peaks were detected in the previous frame, we try the three order levels one by one for the current frame until harmonic peaks are found or the conclusion of no harmonic peaks existing is obtained. Harmonic peaks should have harmonic relations among them and satisfy some sharpness, amplitude and width conditions. Since there are many spurious peaks in the spectrum generated with $P = 80$ or 100, we add the restriction in such cases that harmonic peaks should align consecutively in the lower-to-mid frequency bands and the fundamental frequency should be below 250 Hz based on the features of speech signals. Also, we apply a confidence level to the detection result when $P = 40$, which is set to 1 if the detected harmonic peaks satisfy certain criteria; and set to 0 otherwise. If the confidence level is 1, we proceed to the next signal frame; Otherwise, we attempt to detect harmonic peaks with a higher resolution (i.e., $P = 80$ and 100). If no harmonic peaks are detected in these spectra, we come back to take the result of $P = 40$. Otherwise, we adopt harmonic peaks detected in a spectrum with a higher order. Harmonic peaks detected through the above procedure for two frames of song and speech signals are shown in Fig. 7, where each detected peak is marked with a vertical line.

Harmonic peaks are detected once every 100 input samples, and each signal frame contains 512 samples. The locations of detected peaks are aligned in the temporal order to form the spectral peak tracks. In order to correct detection errors, two post-processing steps are applied to the obtained tracks. The first step is called "linking," in which missing points in the tracks are added to make these tracks complete. This is done

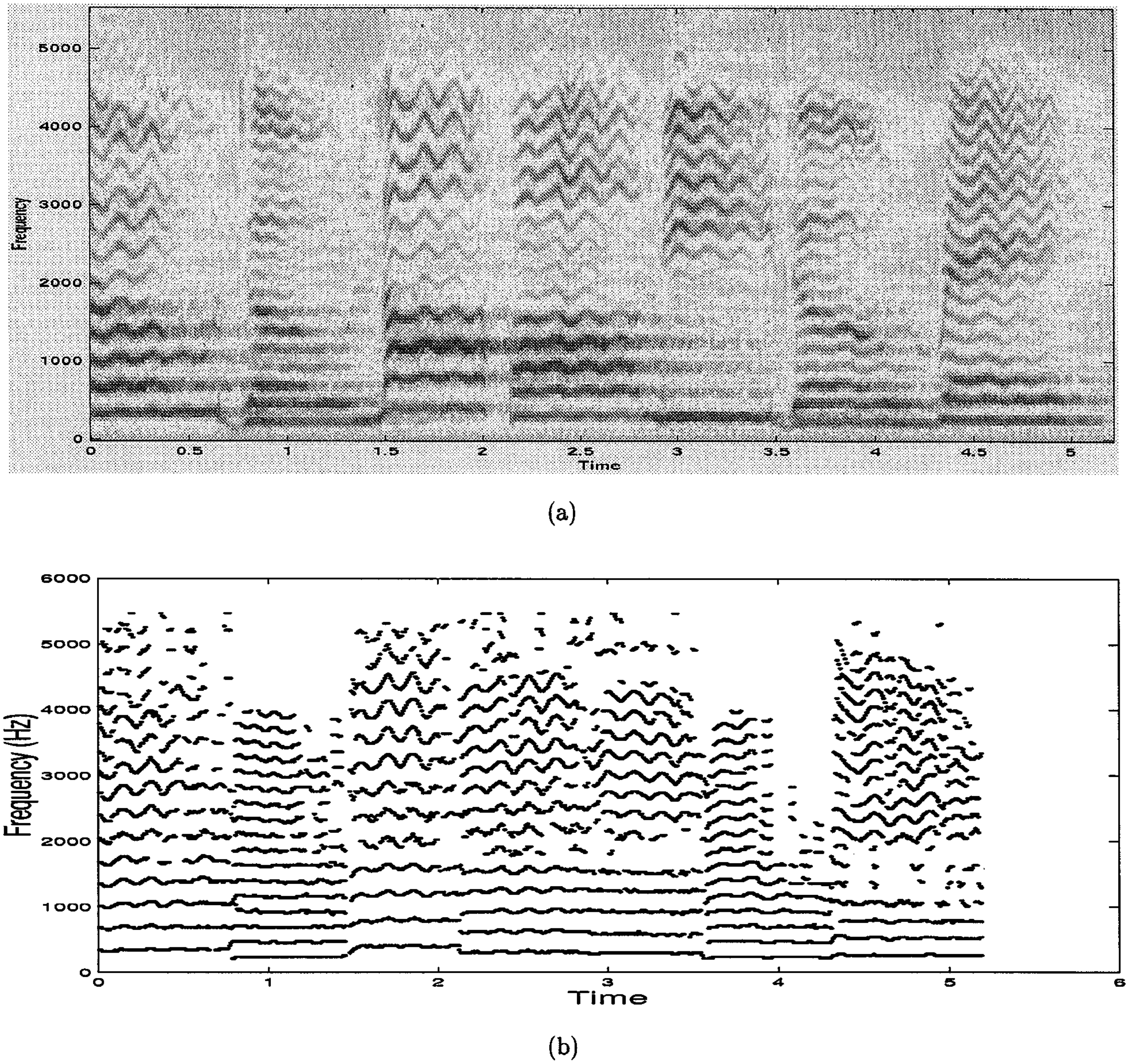


Fig. 8. Spectrogram and spectral peak tracks of female vocal solo.

by searching for holes (one to three samples wide) in the tracks. These missing points may result from weak or overlapped harmonic peaks which are difficult to detect. The second step is called “cleaning,” which is to remove isolated points that are out of the line of any track. Spectrograms and spectral peak tracks estimated with the proposed method for two segments of song and speech signals are illustrated in Figs. 8 and 9, respectively. The first segment is female vocal solo which contains seven notes sung as “5-1-6-4-3-1-2.” We can see that the pitch and the duration of each note are clearly reflected in the detected peak tracks. Each note lasts for about 0.7–0.8 s. Harmonic tracks range from the fundamental frequency at about 225–400 Hz up to 5000 Hz, and are in a ripple-like shape. The second segment is female speech having music and other noise in the background. However, the speech signal is dominant in the spectrogram, and spectral peak tracks are nicely detected despite the interference. The harmonic peak tracks are shorter than those in the song segment with a pitch level of 150–250 Hz.

V. Segmentation and Indexing of Audio Stream

A. Detection of Segment Boundaries

For online segmentation of audiovisual data, short-time values of the energy function, the average zero-crossing rate and the fundamental frequency are computed on the fly with incoming audio data. Whenever there is an abrupt change detected in any of these three features, a segment boundary is set. For the temporal curve of each feature, there are two adjoining sliding windows installed with the average feature value computed within each window as illustrated in Fig. 10. The sliding windows proceed together with each newly computed feature value, and the corresponding average values $Ave(w1)$ and $Ave(w2)$ are updated. These two values are compared. Whenever there is a big difference between them, an abrupt change is claimed to be detected at the common edge of the two windows (i.e., the point E). We choose the length of each window to be 100 feature samples, which corresponds to about 1 s in time with a sampling rate of 11 025 Hz.

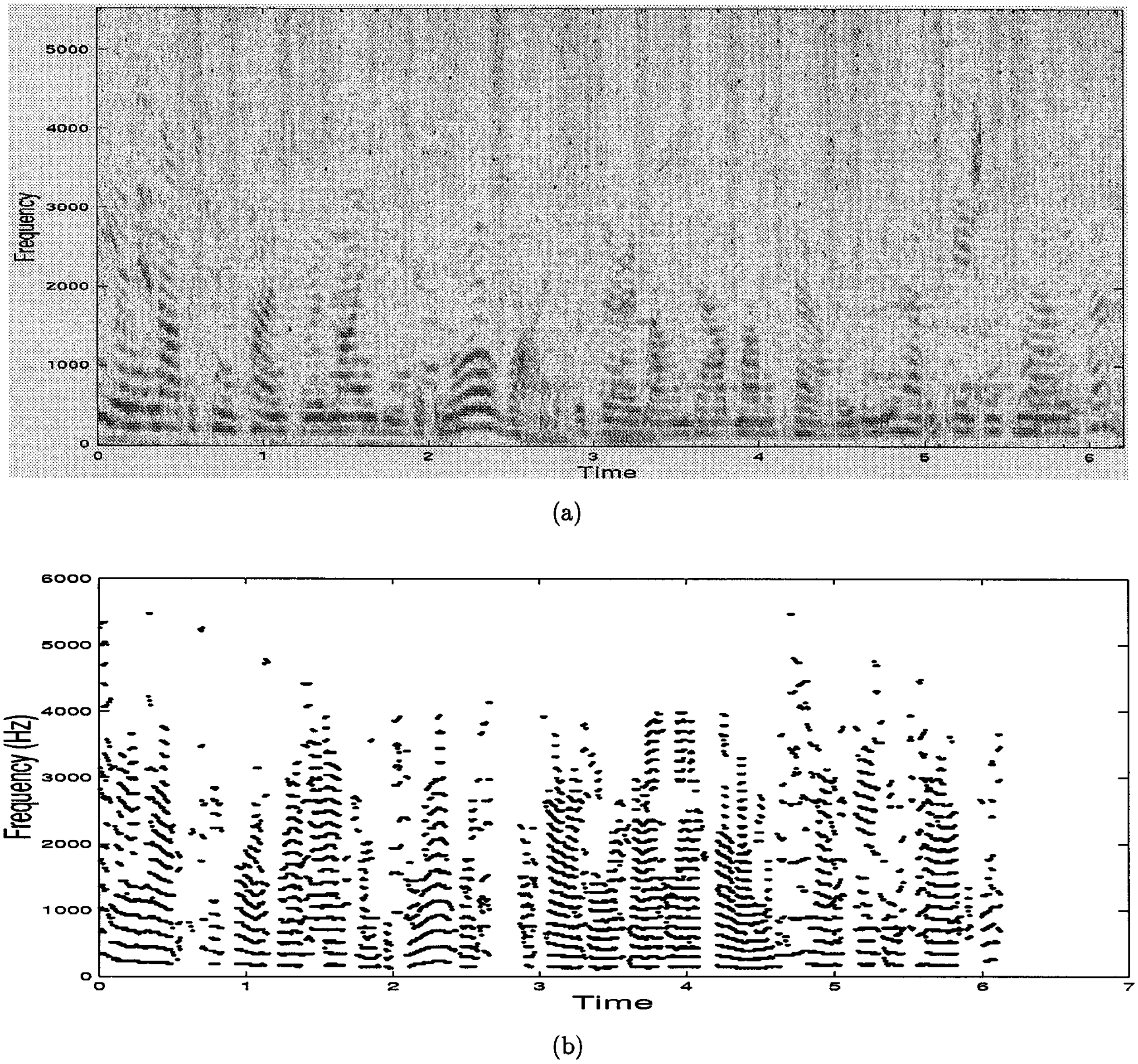


Fig. 9. Spectrogram and spectral peak tracks of female speech with music and noise in the background.

Examples of boundary detection within the temporal curves of the short-time energy function and the short-time fundamental frequency are shown in Fig. 11. Since the temporal evolution pattern and the range of amplitudes of these short- time features are different for speech, music, environmental sound and so on, dramatic changes can be detected at the boundaries of different audio types by applying statistical analysis to these features.

B. Classification of Each Segment

After segment boundaries are detected, each segment of sound is classified into one of the basic audio types according to the procedure as illustrated in Fig. 1. Details about each step in the classification process are described in the following.

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, it is found that the energy level of some noise pieces is not lower than that of some music pieces. The reason that we can hear the music while may not notice the noise is that the frequency-level of the 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 (there may be durations in which the energy is higher than the threshold, but the durations should be short enough and far apart from each other), or if most short-time average zero-crossing rates in the segment are lower than certain set of thresholds, then the segment is indexed as "silence."

2) Separating Sounds with/without 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, the nonsilence audio segments are first separated into two categories: with or without music components, by detecting continuous and stable frequency peaks from the power spectrum.

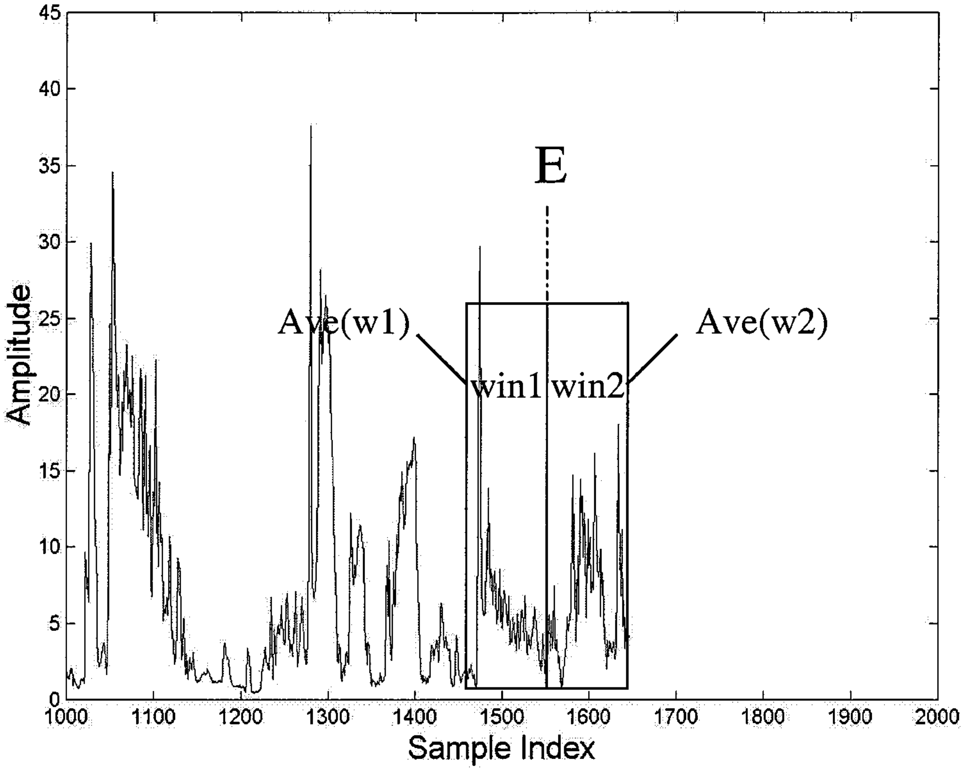


Fig. 10. Setting sliding windows in the temporal curve of audio feature for boundary detection.

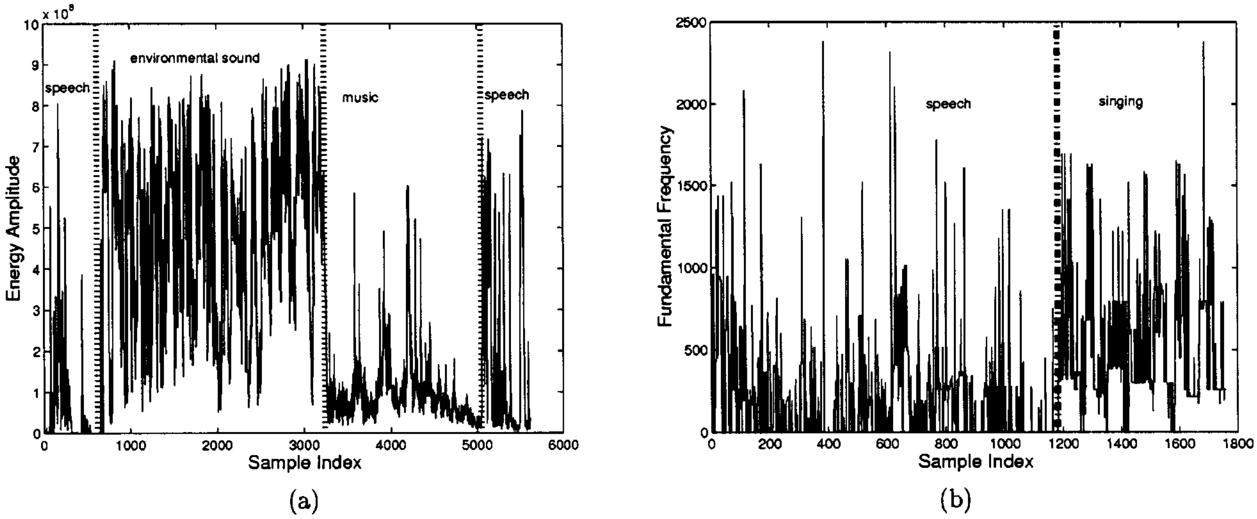


Fig. 11. Boundary detection in the temporal curves of (a) short-time energy function and (b) short-time fundamental frequency.

The power spectrum is generated from AR model parameters of order 40, and is calculated once every 400 input samples. Each signal frame for computing the spectrum contains 512 samples. 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. To avoid the influence from speech harmonic peaks or low frequency noise, only spectral peaks above 500 Hz are considered since most music components are in this range. Signal frames below a certain energy level are also ignored. 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 an 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 (which is called “zero ratio”). The higher the ratio is, the less music components are contained in the sound. Shown in Fig. 12 are index sequences of several sound segments.

By examining zero ratios of different types of audio, we have the following observations.

1) *Speech:* Although the speech signal contains harmonic components, the frequency peaks change faster and last

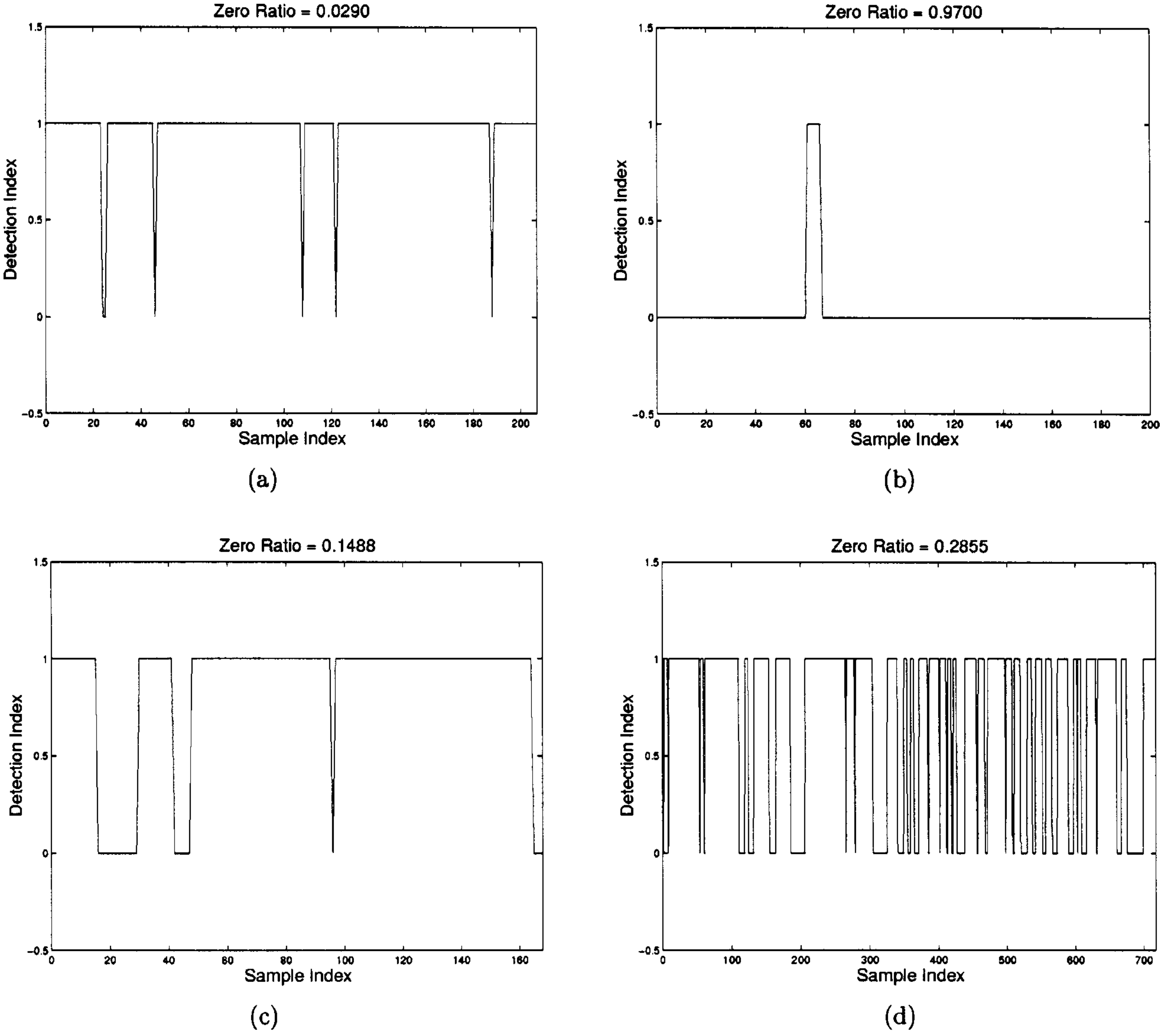


Fig. 12. Index sequences of music components detection in sound segments: (a) pure music, (b) pure speech, (c) speech with music background, and (d) song.

for a shorter time than those in music. Zero ratios for speech segments are normally above 0.95.

2) *Environmental Sound:* Harmonic and stable environmental sounds are all indexed as having music components, while nonharmonic sounds are all indexed as not having music components. However, there are some exceptional cases in between such as certain harmonic and nonharmonic mixed sounds, for which we have to add rules in the program to properly place them.

3) *Pure Music:* Zero ratios for all pure music segments are below 0.3. Indexing errors normally come from short notes, low volume or low frequency parts, nonharmonic components, and the intermissions between notes.

4) *Song:* Most song segments have zero ratios below 0.5. Those parts not detected as having music components result from: peak tracks that shape like ripples instead of lines when the note is long, the intermissions between notes, low volume or low frequency sounds. When the ripple- shaped peak tracks are detected and indexed as music components, zero ratios for songs are significantly reduced.

5) *Speech with Music Background:* When the speech signal is strong, the background music is normally hidden and can not be detected. However, music components can be detected in the intermission periods of speech or when music signal becomes stronger. We make the distinction of the following two cases. In the first case, music is stronger or there are many intermissions in speech so that music is a prominent part of the sound, the zero ratios are below 0.6. In the second case, music is weak while speech is strong and continuous, so that speech is the major component and music may be ignored. Zero ratios are higher than 0.8 in such a case.

Thus, 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. In the second category, there are pure speech and nonharmonic environmental sound. Further classification is done within each category.

3) Detecting Harmonic Environmental Sounds: Within the first category, environmental sounds which are harmonic and stable are separated out first. 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 the sounds of doorbell and pager. This classification step is performed here as a screening process for harmonic environmental sounds, so that they will not interfere with the differentiation of music.

4) Distinguishing Pure Music: Pure music is distinguished based on statistical analyzes of the ZCR and SFuF curves. Four aspects are checked: the degree of being harmonic, the degree of the fundamental frequency's concentration on certain values during a period of time, the variance of the average zero-crossing rate, and the range of amplitude of the average zero-crossing rate. For each aspect, there is one empirical threshold set and a decision value defined. If the threshold is reached, 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 segment's being pure music. For an audio segment to be indexed as "pure music," this probability should be above a certain threshold and at least three of the decision values should be above 0.5.

5) Distinguishing Songs: Up to this step, what are left in the first category include the sound segments of song, speech with music background and environmental sound with music background. We extract the spectral peak tracks of these segments, and differentiate the three audio types based on morphological analyzes of these tracks. The song segments are characterized by one of the three features: ripple-shaped harmonic peak tracks (due to the vibration of vocal chords), tracks which are of longer durations compared to those in speech, and tracks which have a fundamental frequency higher than 300 Hz. The groups of tracks are checked whether any of these three features are matched. The segment will be indexed as "song" if either the sum of durations in which the harmonic peak tracks satisfy one of the features gets to a certain amount, or its comparison to the total length of the segment reaches a certain ratio. The ripple-shaped tracks are detected by taking the first-order difference of the track and checking the pattern of the resulted sequence. One thing to point out is that while some musical instruments such as violin may also generate ripple-shaped peak tracks; they are, however, normally at higher frequency bands.

6) Separating Speech/Environmental Sound with Music Background: According to [21], "when sounds with peaked spectra are mixed, energy from one or other source generally dominates each channel." Therefore, even though there is music in the background, as long as the speech is strong, 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 whether they concentrate in the lower-to-mid frequency bands (with fundamental frequencies 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." Finally, what left in the first category are the segments which have music components but do not meet the criteria for any of the above audio types. They are indexed as "environmental sound with music background."

7) Distinguishing Pure Speech: Within the second category, pure speech is first distinguished and five aspects of conditions are checked. The first aspect is the relation between the temporal curves of ZCR and 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 to zero for speech segments because the peaks appear at different times in the two curves, while it is much larger for other types of audio.

The second aspect is the shape of ZCR curve. For speech, the ZCR curve has a stable and low baseline with peaks above it. We define the baseline to be the linking line of lowest points of troughs in the ZCR curve. The mean and variance of the baseline are calculated. The parameters (amplitude, width, and sharpness) and the appearance frequency of the peaks are also considered. The third and fourth aspects are the variance and the range of amplitude of the ZCR curve, respectively. Contrary to music segments where the variance and the range of amplitude are normally lower than certain thresholds, a typical speech segment has a variance and a range of amplitude that are higher than certain thresholds. The fifth aspect is the fundamental frequency property. As speech is harmonic and nonharmonic mixed, it has a harmony percentage within a certain range. There is also relation between the SFuF curve and the energy curve, i.e., the harmonic parts in SFuF correspond to peaks in the energy curve while the zero parts in SFuF 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 aspects. The weighted average of these decision values represent the possibility of the segment's being speech. When the possibility is above a certain threshold and at least three of the decision values are above 0.5, the segment is indexed as "pure speech."

8) Classifying Nonharmonic Environmental Sounds: The last step is to classify what is left in the second category into one type of nonharmonic environmental sound as the following.

1) If either the energy function curve or the ZCR curve has peaks which have approximately equal intervals between neighboring peaks, the segment is indexed as *"periodic or quasiperiodic"*. Examples of this type include the sounds of clock tick and footstep.
2) If the percentage of harmonic parts in the SFuF curve is within a certain range (lower than the threshold for music, but higher than the threshold for nonharmonic sound), the segment is indexed as *"harmonic and nonharmonic mixed"*. For example, the sound of train horn, which is harmonic, appears with a nonharmonic background. Also, the sound of cough consists of both harmonic and nonharmonic components.

3) If the ZCR values are within a relatively small range compared to the absolute range of the frequency distribution, the segment is indexed as *"nonharmonic and stable"*. One example is the sound of birds' cry, which is nonharmonic but its ZCR curve is concentrated in the range of 80–120 with an absolute range of 150.
4) Finally, if the segment does not satisfy any of the above conditions, it is indexed as *"nonharmonic and irregular"*. Most environmental sounds belong to this type, such as the sounds of thunder, earthquake, and fire.

C. Postprocessing

The postprocessing step is to reduce possible segmentation and classification 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. 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. Through post-processing, these segments are to be combined to other segments and are reindexed based on their contextual relations.

Here are some examples of heuristic rules used in the postprocessing step: if a "silence" segment is shorter than 2 s and the two segments prior and next to it have the same index, the three segments are merged into one and indexed as the same as the first segment; if a "harmonic and fixed" or "harmonic and stable" environmental sound segment is shorter than 5 s and is next to a "pure music" or "song" segment, it is merged into that segment; if a "harmonic and nonharmonic mixed environmental sound" is shorter than 2 s and is between two segments of "speech with music background" or "environmental sound with music background," the three segments are merged into one and indexed according to the first segment.

VI. Experimental Results and Discussion

A. Audio Database

We have built a generic audio database to be used as the testbed of the proposed algorithms, which consists of the following contents: 1000 clips of environmental audio including the sounds of applause, animal, footstep, raining, explosion, knocking, vehicles and so on; 100 pieces of classical music played with ten kinds of instruments, 100 other music pieces of different styles (classic, jazz, blues, light music, Chinese and Indian folk music, etc.); 50 clips of songs sung by male, female, or children, with or without musical instrument accompaniment; 200 speech pieces in different languages (English, German, French, Spanish, Japanese, Chinese, etc.) and with different levels of noise; 50 clips of speech with the music background; 40 clips of environmental sound with the music background; and 20 samples of silence segment with different types of low-volume noise (clicks, brown noise, pink noise and white noise). These short pieces of sound clips (with duration from several seconds to more than 1 min) are used to test the audio classification performances. We also collected dozens of longer audio clips recorded from movies or video programs. These pieces last from several minutes to half an hour, and contain various types of audio. They are used to test the performances for audiovisual data segmentation and indexing.

TABLE I
Classification Results for Audio Categories

Audio Category	Test Samples	Correct Samples		False Alarms	
	Number	Number	Sensitivity	Number	Recall
Silence	20	20	100%	0	100%
With music components	380	362	95.3%	0	100%
Without music components	800	800	100%	18	97.8%

TABLE II
Classification Results for Basic Audio Types

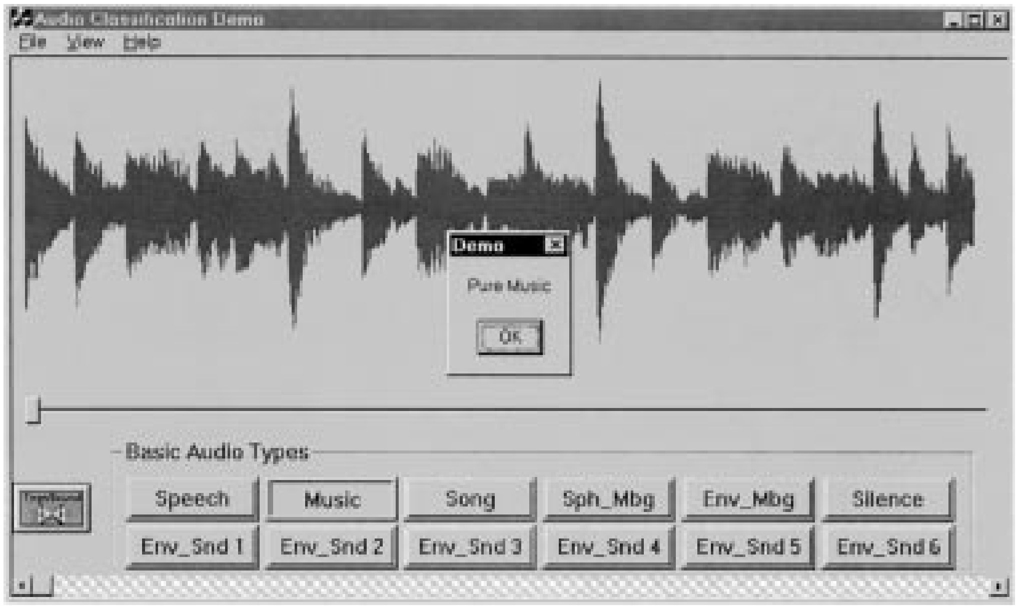
Audio Type	Test Samples	Correct Samples		False Alarms	
	Number	Number	Sensitivity	Number	Recall
Pure speech	200	182	91%	16	91.9%
Pure music	200	189	94.5%	4	97.9%
Song	50	42	84%	2	95.5%
Speech with MBG	50	43	86%	5	89.6%
Sound effect with MBG	40	35	87.5%	6	85.4%
Harmonic sound effect	40	36	90%	0	100%
Non-harmonic sound effect	600	591	98.5%	29	95.3%

B. Generic Audio Data 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. Listed in Tables I and II are results at the two classification hierarchies, respectively, where sensitivity rate is defined as the ratio between the number of correctly classified samples and the actual number of samples in one category, and recall rate is the ratio between the number of correctly indexed samples and the total number of samples as indexed in one audio type (including false alarms). "MBG" is the abbreviation for "music background." In order to obtain threshold values which are used in the heuristic procedures, 10–50% of samples in each audio type were randomly selected to form a training set (i.e., there were 20–60 samples from each type). The threshold values were determined step by step according to the segmentation and indexing process as outlined in Section V. And in each step, an iterative procedure of modifying and testing the threshold values was conducted until an optimal result was achieved. Then, the whole data set was used to testify the classification performances.

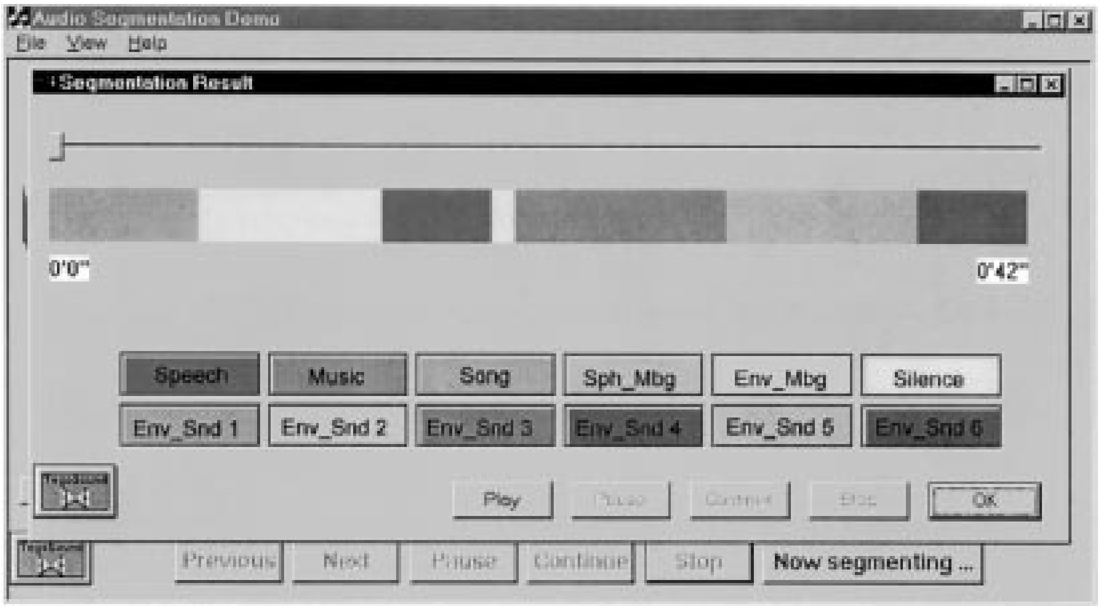
From Table I, we can see that sounds without music components are correctly categorized since they all have a rather high zero ratio. For sounds in the first category, there are classification errors with some song segments (especially those without instrument accompaniment) and some speech with music background segments in which the music components are weak. However, with hybrid-type sounds and sound effects involved here, the overall accuracy for categorizing music/nonmusic sounds is still comparable to previous results for pure speech/music discrimination tasks.

Within the first category, several harmonic sound effects are indexed as "pure music," and there are music segments misclassified as song, speech with MBG or sound effect with MBG. A couple of song segments which lack the ripple-shaped spectral



Pure Music

Fig. 13. Demonstration of generic audio data classification.



Now segmenting ...

Fig. 14. Demonstration of audiovisual data segmentation.

peak tracks are taken as sound effect with MBG. For the second category, apart from false alarms from the first category, there are also 27 misclassifications between pure speech and nonharmonic environmental sound segments. It should be noted that most mistakes result from the very noisy background in some speech, music and song segments. While our approach is normally robust in distinguishing speech and music with a rather high level of noise, the algorithm needs to be further improved so that speech and music components are correctly detected as long as their contents can be recognized by human perception. On the whole, the ratio of the total number of correctly indexed samples in the eight audio types (i.e., audio types listed in Table II plus silence) to the size of the data set reaches 94.8%. We also calculated the accuracy rate of samples not included in the training set, which is 90.7%. A demonstration program was made for the online audio classification, which shows the waveform, the audio features and the classification result for a given sound, as illustrated in Fig. 13.

C. Audiovisual Data Segmentation and Indexing Results

We tested the segmentation procedure with audio clips recorded from movies and video programs. With Pentium333 PC/Windows NT, segmentation and classification tasks can be achieved together with less than one eighth of the time required to play the audio clip. We made a demonstration program for online audiovisual data segmentation and indexing as shown in Fig. 14, where different types of audio data are represented by different colors. Displayed on this figure is the segmentation

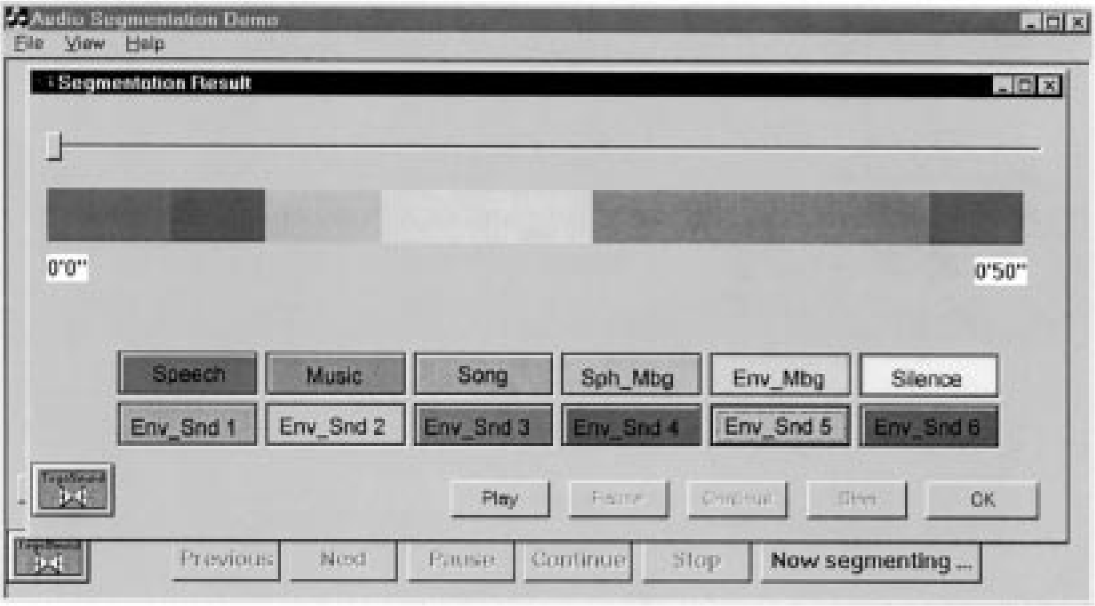


Fig. 15. Segmentation of a movie audio clip.

and indexing result for a 42-s long audio clip recorded from a Spanish cartoon video called "Don Quijote de la Mancha." The first segment in this audio clip is song performed by children and with musical instrument accompaniment, which is indexed as "song." Then, after a period of silence which is indexed as "silence," there is a segment of female speech, and it is indexed as "pure speech." Afterwards, there is a short pause indexed as "silence" and followed by a segment of music which is indexed as "pure music." Next, with the music as background, comes the speech of an old male, and the segment is indexed as "speech with the music background." Finally, the music stops, and there is speech of a boy which is indexed as "pure speech."

For another example, an audio clip recorded from the movie "Washington Square" was segmented as illustrated in Fig. 15. In this 50-s 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 "nonharmonic and irregular environmental sound"), followed by a period of unrecognizable conversation of multiple people simultaneously mixed with baby cry (indexed as the mix of harmonic and nonharmonic 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 as background (indexed as "pure music"). And finally, there is a short conversation between a male and a female (indexed as "pure speech").

Besides the above two examples, we also performed experiments on twenty or so other audio clips. In general, boundaries between segments of different audio types are set quite precisely with a precision within 1 s as compared to human perception. Using human judgement as the ground truth, our algorithm is sensitive enough to detect more than 95% of audio type changes. As to the indexing accuracy, the result is similar to that of the audio classification experiment described in last section, i.e., over 90% of the segments are correctly indexed.

VII. Conclusion and Future Work

A scheme for the automatic segmentation and indexing of audiovisual data based on audio content analysis was presented in this paper. Previous work on video segmentation and annotation has been mostly focused on the visual information. The common framework is to detect video shot changes using histogram difference and motion vectors, and extract keyframes to represent each video shot. However, this visual-based processing often leads to a far too fine segmentation of the audiovisual sequence with respect to the semantic meaning of data. For example, in the video sequence of a song performance, there may be shots appearing in turn of the singer, of the band, of the audience, and of some other designed views. According to the visual information, these shots will be indexed separately. But according to the audio information, we know that they are actually within one performance of a song. Therefore, in our approach for video content parsing, the first step is to conduct a segmentation of the video sequence into semantic scenes based on audio cues, and index each scene with the proposed audio classification algorithms.

While current approaches for audio content analysis are normally developed for specific scenarios, a generic scheme was investigated in this work to cover all sorts of audio signals, including hybrid-type sounds and environmental sounds which are important in many applications, but seldom considered in previous work. Four kinds of audio features including the energy function, the average zero-crossing rate, the fundamental frequency and the spectral peak track are analyzed to reveal differences among different types of audio data. Methods are also proposed for estimating the fundamental frequency and extracting spectral peak tracks from the AR model generated spectrum. Based on audio feature analysis, a procedure for online segmentation and classification of the accompanying audio signal in audiovisual data into twelve basic audio types was accomplished. An accurate classification rate higher than 90% was achieved. In the segmentation and indexing of audio data recorded from movies and video programs, segment boundaries were precisely set, and each segment was properly annotated.

There are several related tasks to be conducted in the future. We will first improve the audio classification procedure so as to make it more robust to all kinds of situations. For example, spectral peak tracks in the segments of chorus may be different from those of solo songs, and may lack the ripple-shaped feature. We will also work on extracting audio features in the compression domain (e.g., the MPEG bitstreams), since most digital audiovisual data available these days are in the compressed format. Then, analysis results of the audio information will be integrated into those of the visual information and the caption in video programs so that a fully functional system for video content parsing can be achieved.

References

[1] S. W. Smoliar and H. Zhang, "Content-based video indexing and retrieval," *IEEE Multimedia*, pp. 62–72, Summer 1994.

[2] M. Flickner, H. Sawhney, and W. Niblack *et al.*, "Query by image and video content: The QBIC system," *Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[3] S.-F. Chang, W. Chen, and H. J. Meng *et al.*, "A fully automated content based video search engine supporting spatio-temporal queries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 602–615, Sept. 1998.

[4] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'96*, vol. 2, Atlanta, GA, May 1996, pp. 993–996.

[5] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'97*, Munich, Germany, Apr. 1997.

[6] L. Wyse and S. Smoliar, "Toward content-based audio indexing and retrieval and a new speaker discrimination technique," Inst. Syst. Sci., Nat. Univ. Singapore, http://www.iss.nus.sg/People/lwyse/lwyse.html, Dec. 1995.

[7] D. Kimber and L. Wilcox, "Acoustic segmentation for audio browsers," in *Proc. Interface Conf.*, Sydney, Australia, July 1996.

[8] S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic audio content analysis," Praktische Informatik IV, Univ. Mannheim, Mannheim, Germany, http://www.informatik.uni-mannheim.de/pfeiffer/publications/, Apr. 1996.

[9] A. Ghias, J. Logan, and D. Chamberlin, "Query by humming-musical information retrieval in an audio database," in *Proc. ACM Multimedia Conf.*, 1995, pp. 231–235.

[10] J. Foote, "Content-based retrieval of music and audio," *Proc. SPIE*, 1997.

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

[12] G. Smith, H. Murase, and K. Kashino, "Quick audio retrieval using active search," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'98*, Seattle, WA, May 1998, pp. 3777–3780.

[13] Z. Liu, J. Huang, and Y. Wang *et al.*, "Audio feature extraction and analysis for scene classification," in *Proc. IEEE 1st Multimedia Workshop*, 1997.

[14] Z. Liu, J. Huang, and Y. Wang, "Classification of TV programs based on audio information using hidden Markov model," in *Proc. IEEE 2nd Workshop Multimedia Signal Processing*, Redondo Beach, CA, Dec. 1998, pp. 27–32.

[15] Z. Liu and Q. Huang, "Classification of audio events in broadcast news," in *Proc. IEEE 2nd Workshop Multimedia Signal Processing*, Dec. 1998, pp. 364–369.

[16] N. Patel and I. Sethi, "Audio characterization for video indexing," in *Proc. SPIE Conf. Storage Retrieval Still Image Video Databases*, vol. 2670, San Jose, CA, 1996, pp. 373–384.

[17] K. Minami, A. Akutsu, and H. Hamada *et al.*, "Video handling with music and speech detection," *IEEE Multimedia*, pp. 17–25, Fall 1998.

[18] J. Huang, Z. Liu, and Y. Wang, "Integration of audio and visual information for content-based video segmentation," in *Proc. IEEE Conf. Image Processing*, Oct. 1998.

[19] M. R. Naphade, T. Kristjansson, and B. Frey *et al.*, "Probabilistic multimedia objects (MULTI-JECTS): A novel approach to video indexing and retrieval in multimedia systems," in *Proc. IEEE Conf. Image Processing*, Chicago, IL, Oct. 1998.

[20] J. S. Boreczky and L. D. Wilcox, "A hidden Markov model framework for video segmentation using audio and image features," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'98*, May 1998, pp. 3741–3744.

[21] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.

[22] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 2, pp. 297–336, 1994.

[23] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ. , Stanford, CA, 1985.

[24] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Mass. Inst. Technol., Cambridge, MA, 1996.

[25] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer, "Structured audio: Creation, transmission, and rendering of parametric sound representations," *Proc. IEEE*, vol. 86, pp. 922–939, May 1998.

[26] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[27] A. Choi, "Real-time fundamental frequency estimation by least-square fitting," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 201–205, Mar. 1997.

[28] B. Doval and X. Rodet, "Estimation of fundamental frequency of music sound signals," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'91*, vol. 5, Toronto, ON, Canada, Apr. 1991, pp. 3657–3660.

[29] W. B. Kuhn, "A real-time pitch recognition algorithm for music applications," *Comput. Music J.*, vol. 14, no. 3, pp. 60–71, Fall 1990.

[30] F. Everest, *The Master Handbook of Acoustics*. New York: McGraw-Hill, 1994.

Tong Zhang (M'98) received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees of biomedical engineering from Tsinghua University, Beijing, China, in 1992, 1994, and 1996, respectively, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, in 1998 and 1999, respectively.

Since February 2000, she has been a Technical Contributor at the Hewlett-Packard Laboratories, Palo Alto, CA. Her research interests are in the areas of digital signal and image processing, audio and video content analysis, and multimedia database management.

Dr. Zhang is a member of SPIE and ACM.

C.-C. Jay Kuo (S'83–M'86–SM'92–F'99) received the B.S. degree from the National Taiwan University, Taipei, Taiwan, R.O.C., in 1980 and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively, all in electrical engineering.

He was Computational and Applied Mathematics (CAM) Research Assistant Professor with the Department of Mathematics, University of California, Los Angeles, from October 1987 to December 1988. Since January 1989, he has been with the Department of Electrical Engineering—Systems and the Signal and Image Processing Institute, University of Southern California, Los Angeles, where he currently has a joint appointment as Professor of electrical engineering and mathematics. His research interests are in the areas of digital signal and image processing, audio and video coding, media communication technologies and delivery protocols, and network computing. He has authored more than 400 technical publications in international conferences and journals and graduated around 30 Ph.D. students. He is the Editor-in-Chief for the *Journal of Visual Communications and Image Representation* and Editor for the *Journal of Information Science and Engineering*.

Dr. Kuo is a member of SIAM, ACM, and SPIE. He is an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING and was Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING from 1995 to 1998 and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 1995 to 1997. He received the National Science Foundation Young Investigator Award (NYI) and Presidential Faculty Fellow (PFF) Award in 1992 and 1993, respectively.