Content-based classification and retrieval of audio

Zhang, Tong, Kuo, C.-C. Jay

Tong Zhang, C.-C. Jay Kuo, "Content-based classification and retrieval of audio," Proc. SPIE 3461, Advanced Signal Processing Algorithms, Architectures, and Implementations VIII, (2 October 1998); doi: 10.1117/12.325703

Event: SPIE's International Symposium on Optical Science, Engineering, and Instrumentation, 1998, San Diego, CA, United States

Content-Based Classification and Retrieval of Audio

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
Email:{tzhang,cckuo}@sipi.usc.edu

ABSTRACT

An online audio classification and segmentation system is presented in this research, where audio recordings are classified and segmented into speech, music, several types of environmental sounds and silence based on audio content analysis. This is the first step of our continuing work towards a general content-based audio classification and retrieval system. The extracted audio features include temporal curves of the energy function, the average zero-crossing rate, the fundamental frequency of audio signals, as well as statistical and morphological features of these curves. The classification result is achieved through a threshold-based heuristic procedure. The audio database that we have built, details of feature extraction, classification and segmentation procedures, and experimental results are described. It is shown that, with the proposed new system, audio recordings can be automatically segmented and classified into basic types in real time with an accuracy of over 90%. Outlines of further classification of audio into finer types and a query-by-example audio retrieval system on top of the coarse classification are also introduced.

Keywords: audio content analysis, audio segmentation and classification, audio database, average zero-crossing rate, fundamental frequency

1 INTRODUCTION

Audio, including voice, music and various kinds of environmental sounds, is an increasingly important type of media, and plays a significant role for audiovisual data. While there are quite a few systems for content-based image and video retrieval at present, very little work has been done on the audio portion of a multimedia stream. However, since there are many digital audio databases in place, research on effective management of audio databases is expected to gain more attention these days.

Audio content analysis, classification, and retrieval have a wide range of applications in the entertainment industry, audio archive management, commercial musical usage, surveillance, etc. Let us consider several examples below. It will be very helpful to be able to search sound effects automatically from a very large audio database in the film postprocessing, which contains sounds of explosion, windstorm, earthquake, animals, and so on. In Karaoke or music/video stores, the ability to retrieve songs or musical products by humming and/or playing only a segment of melody would be very convenient to customers. There are also distributed audio libraries in the World Wide Web for management. While the use of keywords for sound browsing and retrieving provides a possible solution, it is however time- and labor-consuming in indexing. Moreover, an objective and consistent description of these sounds is lacking, since features of sounds are very difficult to describe. Consequently, content-based audio retrieval would be the ideal approach for sound indexing and searching. Furthermore, content analysis of audio is useful in audio-assisted video analysis. Possible applications include video scene classification, automatic segmentation and indexing of raw audiovisual recordings, and audiovisual database browsing.

Existing research on content-based audio data management is quite primitive. It can be generally put into three categories: (1)audio segmentation and classification; (2)audio retrieval; and (3)audio analysis for video indexing.

Part of the SPIE Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII • San Diego, California • July 1998
SPIE Vol. 3461 • 0277-786X/98/$10.00

One basic problem in audio segmentation and classification is speech/music discrimination. The approach presented in [1] used only the average zero-crossing rate and energy features, and applied a simple thresholding procedure. While in [2], 13 features in time, frequency, and cepstrum domains, as well as more complicated classification methods were used to achieve a robust performance. Since speech and music have different spectral distribution and temporal changing patterns, it is not very difficult to reach a relatively high level of discrimination accuracy. A further classification of audio may take other sounds, besides speech and music, into consideration. In [3], audio was classified into "music", "speech" and "others". Music was first detected based on the average length of time that 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 in [4], where audio recordings were segmented into speech, silence, laughter and non-speech 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.

One specific technique in content-based audio retrieval is query-by-humming. The approach in [5] defined the sequence of relative differences in the pitch to represent the melody contour and adopted the string matching method to search similar songs. It was reported that, with 10-12 pitch transitions, 90% of the 183 songs contained in a database could be discriminated. A music and audio retrieval system was proposed in [7], where the Mel-frequency cepstral coefficients (MFCC) were taken as features, and a tree-structured classifier was built for retrieval. Since MFCC do not represent the timbre of sounds well, this method in general failed to distinguish music and environmental sounds with different timbre characters. In the content-based retrieval (CBR) work of the Musclefish Company [6], they took statistical values (including means, variances, and autocorrelations) of several time- and frequency-domain measurements to represent perceptual features like loudness, brightness, bandwidth, and pitch. As merely statistical values are used, this method is only suitable for sounds with a single timbre.

In [8], audio analysis was applied 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 was adopted as the classifier. It was shown that 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. In [9], audio characterization was performed on MPEG data (actually, the sub-band level data) for the purpose of video indexing. Audio was classified into dialog, non-dialog and silence intervals. Features were taken from the energy, pitch, spectrogram, pause rate domains, and organized in a threshold procedure. There were somehow quite a few mistakes occurring in the classification between dialog and non-dialog intervals.

Audio classification and retrieval is an important and challenging research topic. The classification can be done in different ways through different depths. The retrieval can be emphasized on different types of audio according to various application needs. As described above, work in this area is still at a very preliminary stage. Our objective in this research is to build a hierarchical system which consists of coarse-level and fine-level audio classification and audio retrieval. In coarse classification, speech, music, environmental audio, and silence are separated. In fine classification, more specified classes of natural and man-made sounds are discriminated. And in audio retrieval, desirable sounds may be searched by an example query or a set of features.

Compared with previous work, we put more emphases on the environmental audio, which has often been ignored in the past. Environmental sounds are an important ingredient in audio recordings, and their analysis is inevitable in many real applications. We also investigate physical and perceptual features of different classes of audio and apply signal processing techniques to the representation and classification of extracted features.

The paper is organized as follows. An overview of the proposed content-based audio classification and retrieval system is presented in Section 2. Details about the building blocks such as audio feature extraction, coarse-level classification and on-line segmentation are described in Sections 3-5, respectively. Experimental results are shown in Section 6, and concluding remarks are given in Section 7.

2 OVERVIEW OF PROPOSED AUDIO CLASSIFICATION AND RETRIEVAL SYSTEM

We are currently working on a hierarchical system for audio content analysis and classification. With such a system, audio data can be archived appropriately for the ease of retrieval at the query stage. To build such a system, we divide its implementation into three stages. In the first stage, audio signals are classified into basic types, including speech, music, several types of environmental sounds and silence. It is called the coarse-level classification. For this level, we use relatively simple features such as the energy function, the average zero-crossing rate, and the fundamental frequency to ensure the feasibility of real-time processing. We have worked on morphological and statistical analysis of these features to reveal differences among different types of audio. A rule-based heuristic procedure is built to classify audio signals based on these features. An on-line segmentation and indexing of audio/video recordings is achieved based on the coarse-level classification. For example, in arranging the raw recording of meetings or performances, segments of silence or irrelevant environmental sounds (including noise) may be discarded, while speech, music and other environmental sounds can be classified into corresponding archives. Techniques and demonstrations of this stage will be presented in later sections.

In the second stage, further classification is conducted within each basic type. For speech, we can differentiate it into voices of man, woman, child as well as speech with a music background. For music, we classify it according to playing instruments and types (for example, classics, blues, jazz, rock and roll, music with singing and the plain song). For environmental sounds, we divide them into finer classes such as applause, bell ring, footstep, windstorm, laughter, birds' cry, and so on. This is known as the fine-level classification. Based on this result, a finer segmentation and indexing result of audio material can be achieved. Due to differences in the origination of the three basic types of audio, i.e. speech, music and environmental sounds, very different approaches can be taken in their fine classification. In this paper, we focus primarily on the fine classification of environmental audio. Features are extracted from the time-frequency representation of audio signals to reveal subtle differences of timbre, pitch, and change pattern among different classes of sounds. The hidden Markov model (HMM) is used as the classifier, because it can properly represent the evolution of features over time which is important for audio data. One HMM is built for each class of sound. The fine classification of audio is well suited for automatic indexing and browsing of audio/video databases and libraries.

In the third stage, an audio retrieval system is built based on the archiving scheme described above. There are two retrieval approaches. One is query-by-example, where the input is an example sound, and the output is a rank list of sounds in the database which shows the similarity of retrieved sounds to the input query. Similar to image retrieval systems where the search of images may be done according to color, texture, or shape features, audio clips may also be retrieved with distinct features such as timbre, pitch, and rhythm. The user may choose one feature or a combination of features with respect to the sample audio clip. The other one is query-by-keywords (or features), where various aspects of audio features are defined in a special keyword list. The keywords include both conceptual definitions (such as violin, applause, or cough) and perceptual descriptions (such as fastness, brightness, and pitch) of sounds. In an interactive retrieval process, users may choose from a given menu a set of features, listen to the retrieved samples, and modify the input feature set to get a better matched result. Application examples of this system may include searching sound effects in producing films, audio editing in making TV or radio programs, selecting and browsing materials in audio libraries, and so on.

The procedure of the proposed audio classification and retrieval approach in an audio archive management system is illustrated in Figure 1. Raw audio recordings are analyzed and segmented based on abrupt changes of features. Then, audio segments are classified and indexed. They are stored in corresponding archives. The audio archives are organized in a hierarchical way for the ease of the storage and retrieval of audio clips. When a user wants to browse the audio samples in the archives, he may put a set of features or a query sound into the computer. The search engine will then find the best matched sounds and present them to the user. The user may also refine the query to get more audio material relevant to his interest.

In the following three sections, we will introduce in detail the features and procedures for the coarse-level content-based audio classification and segmentation.

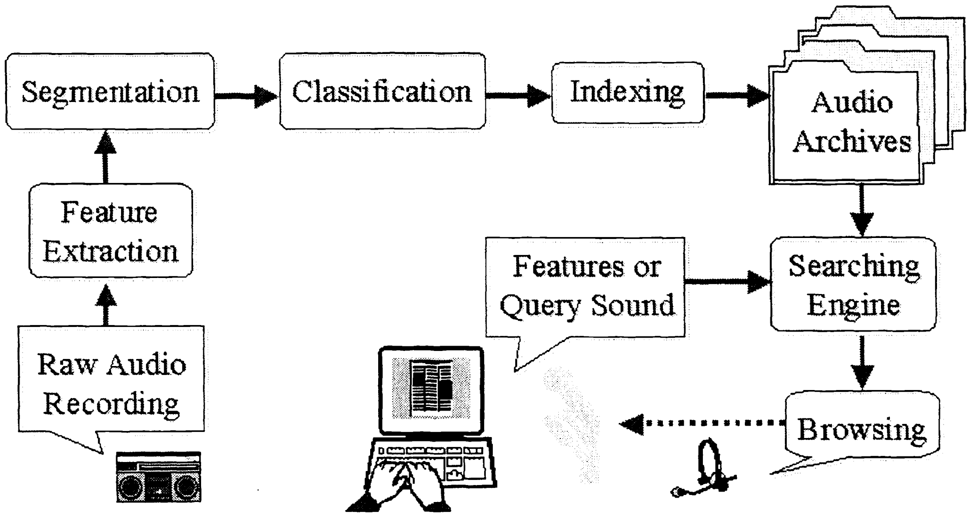


Figure 1: Application of content-based audio classification and retrieval to audio archive management.

3 AUDIO FEATURE EXTRACTION

Three kinds of features are used in our work, namely, the short-time energy function, the average zero-crossing rate, and the fundamental frequency. They are detailed below.

3.1 Short-Time Energy Function

The short-time energy 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.

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

It provides a convenient representation of the amplitude variation over the 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 11025 samples per second. We set the window duration of $w(n)$ to be 150 samples so that there is an overlap between neighboring frames. The audio waveform of a typical speech segment and its short-time energy curve are shown in Figure 2. Note that the sample index of the energy curve is at the ratio of 1:100 compared to the corresponding time index of audio signal.

For speech signals, one major significance of the energy function is that 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 significantly smaller than those for the voiced components, as can be seen from the peaks and troughs in the energy curve. In many applications, the energy function can also be used as the measurement to distinguish silence.

3.2 Average Zero-Crossing Rate (ZCR)

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. This is

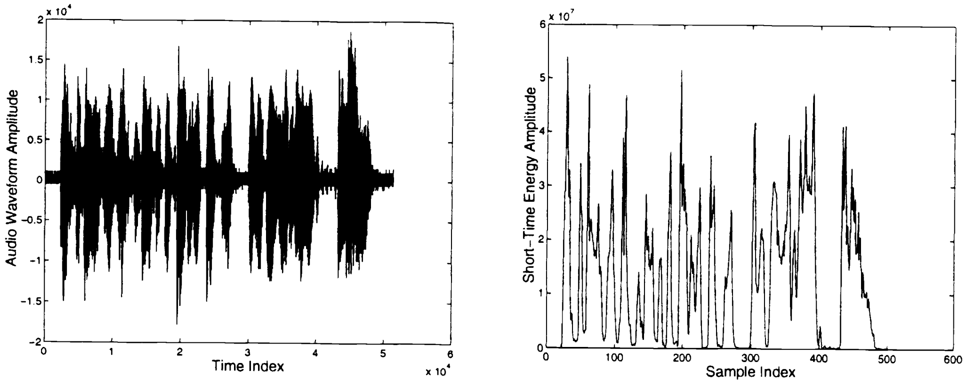


Figure 2: The audio waveform and the short-time energy of a speech segment

particularly true of narrowband signals. Since audio signals may include both narrowband and broadband signals, the interpretation of the average zero-crossing rate is less precise. However, rough estimates of spectral properties can still be obtained using a representation based on the short-time average zero-crossing rate, as defined below:

$$Z_n = \sum_m |sgn[x(m)] - sgn[x(m-1)]| w(n-m), \qquad (2)$$

where

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

and

$$w(n) = \begin{cases} 1/2, & 0 \leq n \leq N-1, \\ 0, & \text{otherwise.} \end{cases}$$

The short-time average zero-crossing rate (ZCR) curves of several audio samples are shown in Figure 3. Similar to the computation of short-time energy function, we also choose to compute ZCR every 100 input samples, and set the window width to 150 samples.

The speech production model suggests that the energy of voiced speech signals is concentrated below 3 kHz because of the spectral fall-off introduced by the glottal wave whereas most of the energy is found at higher frequencies for unvoiced speech signals [10]. Since high (or low) frequencies imply high (or low) zero-crossing rates, a reasonable rule is that if the zero-crossing rate is high, the speech signal is unvoiced while if the zero-crossing rate is low, the speech signal is voiced. Hence, the zero-crossing rate can be used for making distinction between voiced and unvoiced speech signals. As shown in Figure 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 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 plotted in Figure 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 an irregular waveform with a changing baseline and a relatively small range of the 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 reveals a continuous drop of the frequency centroid over the time while that of the footstep sound is rather irregular. We may briefly classify environmental sounds according to the properties of their ZCR curves such as regularity, periodicity, stability, and the range of amplitude, for both coarse-level separation and fine-leval classification.

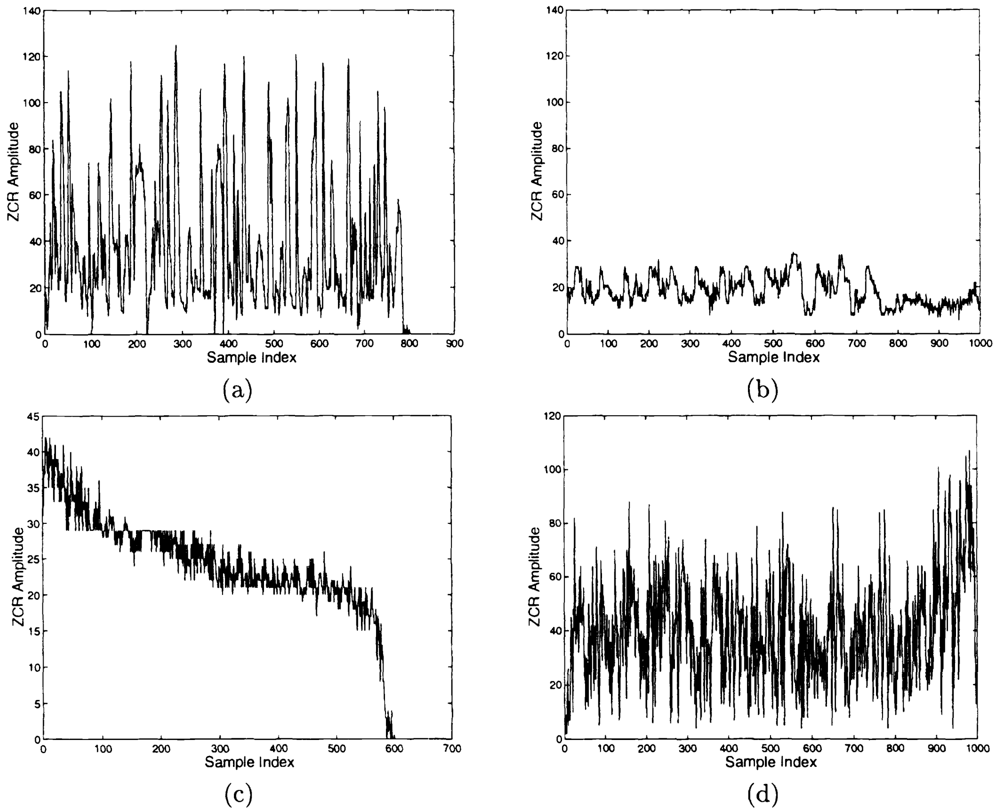


Figure 3: Average zero-crossing rates of four audio signals: (a)speech, (b)piano, (c)chime and (d)footstep

3.3 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 sound generated by violin and applause are illustrated in Figure 4, respectively. It is clear that the former one is harmonic while the latter one is non-harmonic.

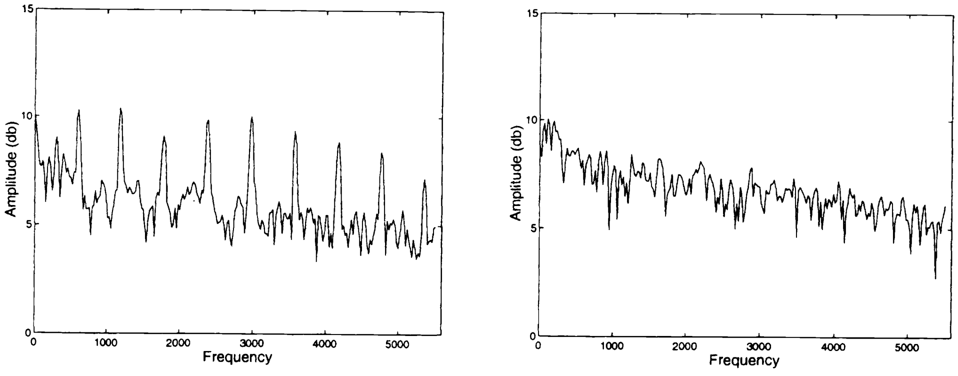
Whether an audio segment is harmonic or not depends on its source. Sounds from most musical instruments are harmonic. The speech signal is a mixed harmonic and non-harmonic sound, since voiced components are harmonic while unvoiced components are non-harmonic. Most environmental sounds are non-harmonic except that there are some examples which are harmonic and stable (such as the sound of doorbell), or mixed harmonic and non-harmonic (such as clock tick). With our experience, the harmonic feature of a sound often plays an important role in the coarse-level classification. To measure the harmonic feature, let us define the short-time fundamental frequency as follows:

$$F_n = fuf\{\log|FFT(x(m)w(n-m))|\}, \tag{3}$$

where

$$w(n) = \begin{cases} 0.5(1-\cos(2\pi\frac{n}{N-1})), & 0 \le n \le N-1, \\ 0, & \text{otherwise.} \end{cases}$$

is the Hanning window, which is chosen for its relatively small side-lobes and fast attenuation which makes it easier for frequency peak detection. In our actual implementation, the audio signal is first multiplied with $w(n)$



(a) (b)

Figure 4: Spectra of harmonic and non-harmonic sound: (a) violin and (b) applause.

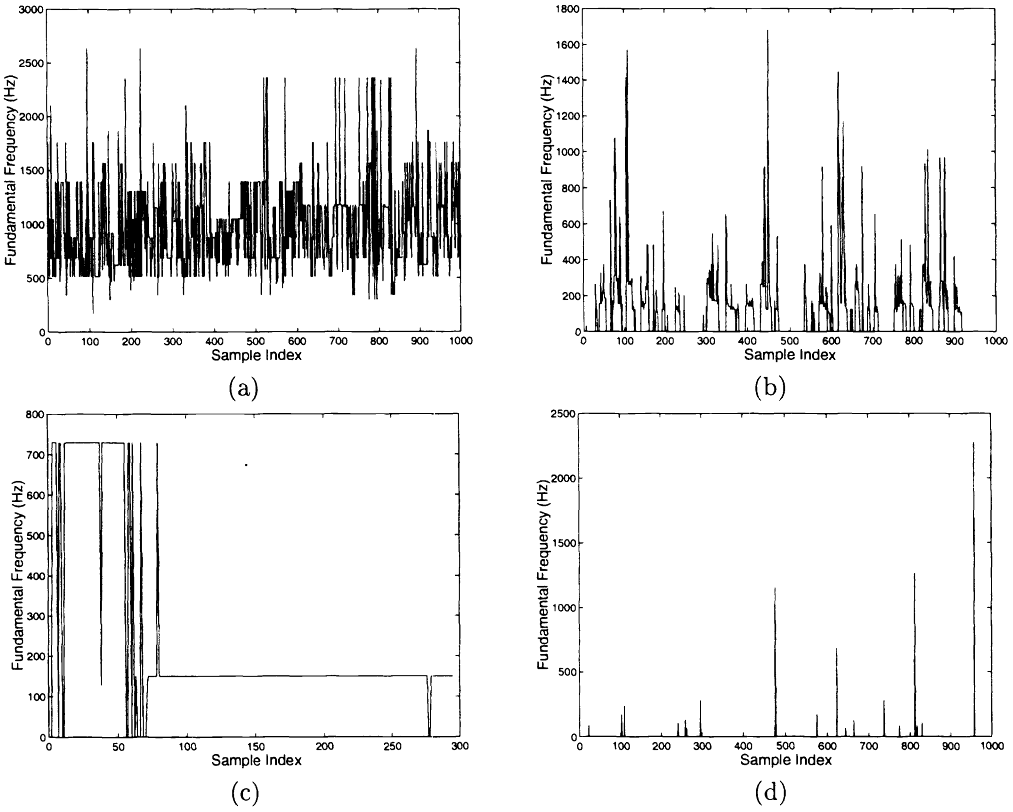
of 512-sample wide (i.e., $N = 512$). Then, the amplitude spectrum is calculated, and the logarithm is taken. The remaining key task is the estimation of the fundamental frequency from the short-time spectrum, which is denoted by the operator $fuf\{\cdot\}$. This operation is detailed below.

Fundamental frequency estimation, or equivalently pitch detection, has been one of the most important problems in speech/music analysis. (It is however worthwhile to point out that the fundamental frequency is a physical measurement while the pitch is rather a perceptual term which is analogous to the frequency but not exactly the same, as stated in [11].) There are many schemes proposed to solve this problem, but none of them is perfectly satisfactory for a wide range of audio signals. Our primary purpose of estimating the fundamental frequency is to detect the harmonic property for all kinds of audio signals. Thus, we desire to choose a method which is simple, robust, but not necessarily perfectly precise. The chosen approach primarily consists of two steps. First, peaks in the spectrum which might represent the harmonics are detected. These peaks should be well above the average amplitude of the frequency response as illustrated in Figure 4(a). Adaptive thresholding based on the moving average of the spectrum amplitude is applied. Other thresholds of amplitudes and widths are also performed to further confine peak locations. Second, it is checked whether there are harmonic relations among detected peaks, to be more precisely, whether the peaks (or some of them) are integer multiples of a common frequency which corresponds to the fundamental frequency. If so, the fundamental frequency is estimated from these peaks. Otherwise, the spectrum does not contain harmonic components so that there is no fundamental frequency. For such cases, we set the value of the fundamental frequency to zero.

We plot the short-time fundamental frequency curves of five sample audio signals in Figure 5. Again, the fundamental frequency is calculated once for every 100 input samples. We can see that the music clip played with the organ is continuously harmonic with the fundamental frequency concentrated in 500-1500Hz most of the time. The speech signal is a mixed harmonic and non-harmonic type. The voiced speech is harmonic with a fundamental frequency normally below 600Hz. The unvoiced speech is non-harmonic as denoted by zeros in the curve. Most environmental sounds are non-harmonic like the example shown in Figure 5(d) with more than 90% of the curve being zero. But there are exceptions such as the sound of doorbell which is harmonic and the values of fundamental frequency represent the two phases of the sound (i.e. the first interval has a higher pitch while the second interval has a lower pitch).

4 AUDIO CLASSIFICATION

With a certain segment of audio, the temporal curves of the three short-time features as described above are computed. Then, through a rule-based heuristic procedure, the segment is classified into one of the basic audio types.



(a) (b) (c) (d)

Figure 5: Short-time fundamental frequency of audio signals: (a) organ, (b) speech, (c) doorbell and (d)ping-pong.

4.1 Separating 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 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 zero-crossing rates are lower than certain set of thresholds, then the segment is indexed as "silence".

4.2 Separating Environmental Sounds with Special Features

The second step is to separate out environmental sounds which are harmonic and stable. The short-time fundamental frequency curve is checked. If most parts of the temporal curve are harmonic, and the fundamental frequency is fixed at one particular value, then the segment is indexed as "harmonic and unchanged". A typical example of this type is the sound of touchtone. 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 the doorbell and the pager. This classification step is performed here as a screening process for harmonic environmental sounds, so that they will not confuse the classification of music. It is also the basis of further fine classification of harmonic environmental audio.

4.3 Distinguishing Music

Music is distinguished based on the zero-crossing rate and the fundamental frequency properties. Four aspects are checked, i.e. the degree of being harmonic, the degree of the fundamental frequency concentration on certain values during a period of time, the variance of zero-crossing rates, and the range of the amplitude of the zero-crossing rate. For each aspect, these 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 certain weights to derive a total probability of the audio segment being music. For a segment to be indexed as "music", this probability should be above a certain threshold and at least three of the decision values should be above 0.5.

4.4 Distinguishing Speech

When distinguishing speech, five aspects of conditions are checked. The first aspect is the relation between ZCR and energy curves. For speech, 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. One example is shown in Figure 6. We cut both ZCR and energy curves at 1/3 of the maximum amplitude and remove the lower part, so that only peaks of the two curves remain. Then, the inner product of the two residual curves is calculated. The product is normally near to zero for speech segments, but 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, where the baseline is defined to be the linking line of lowest points of troughs. We check the mean and the variance of the baseline. The shape and the frequency of peaks are also considered. The third and fourth aspects are the variance and the range of the amplitude of the ZCR curve, respectively. Contrary to music where the variance and the range of the amplitude are normally lower than certain thresholds, a typical speech segment has a variance and a range of the amplitude that are higher than certain thresholds. The fifth aspect is about 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 correspond to peaks in the energy curve while the zero parts 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 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 "speech".

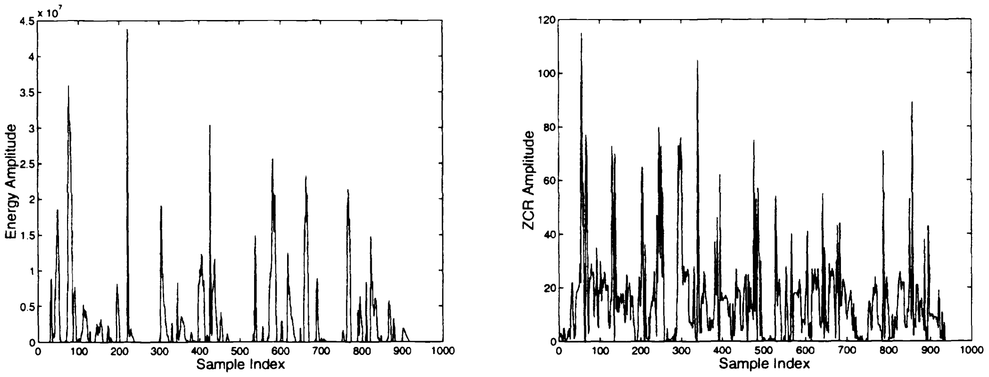


Figure 6: Energy and ZCR curves of a piece of speech.

4.5 Classifying Other Environmental Sounds

The last step is to classify what is left into one type of non-harmonic environmental sounds. If either the energy curve or the ZCR curve has peaks which have approximately equal intervals between neighboring peaks, the segment is indexed as "periodic or quasi-periodic". Examples include sounds of the clock tick and the regular footstep. This is a beginning of rhythm analysis. More complicated rhythm analysis will be done in the fine-level classification. If the percentage of harmony 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. Also, the sound of cough consists of both harmonic and non-harmonic components. If the frequency centroid 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 non-harmonic but its ZCR curve is concentrated within the range of 80-120. Finally, 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 sounds of thunder, earthquake and fire.

5 AUDIO SEGMENTATION

The classification procedure described in the previous section classifies one audio segment into one of the basic types. For on-line segmentation of audio recordings, there are three steps involved, i.e. detection of segment boundaries, classification of each segmented interval, and post-processing to refine segmented results.

5.1 Detection of Segment Boundaries

The short-time energy function, the short-time average zero-crossing rate, and the short-time 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 each feature curve, there is a sliding window to compute the average amplitude within the window. The sliding window proceeds, and we compare the average amplitude of the current window with that of the window right next to it. Whenever a big difference is observed, we claim that an abrupt change is detected. Detected boundaries in the energy and fundamental frequency curves are illustrated in Figure 7.

5.2 Classification of Each Segment

After segment boundaries are detected, each segment is classified into one of the basic audio types by using the classification procedure described in Section 4.

5.3 Post-Processing

The post-processing procedure is to reduce possible segmentation errors. We have adjusted our 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 be even misclassified as "harmonic and stable environmental sound" because of the unchanged tune in the segment. Through post-processing, these segments can be combined together according to their contextual relation.

6 EXPERIMENTAL RESULTS

6.1 Audio Database

We have built a generic audio database as a testbed for various audio classification and segmentation algorithms. It includes the following contents: 1000 environmental sound clips, 100 pieces of music played with 10 kinds of instruments, other music pieces of different styles, songs sung by male and female, speech in different languages and with different levels of noise, speech with the music background. These short pieces of sound clips (with duration from several seconds to more than one 1 minute) are used to test the classification performance. We have also collected dozens of longer audio clips from movies. These pieces last from several minutes to half an hour, and

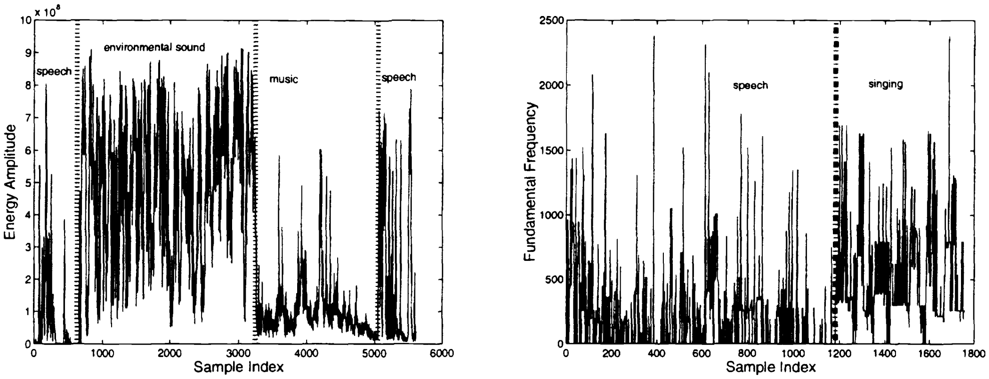


Figure 7: Boundary detection in the energy and fundamental frequency curves.

contain various types of audio. They are used to test the segmentation performance.

6.2 Classification Result

The proposed coarse-level classification scheme achieved an accuracy rate of more than 90% by using the audio test database described above. Misclassification usually occurs in the hybrid sound which contains more than one basic type of audio. For example, the speech signal with the music background and the singing of a person are two types of hybrid sounds which have characters of both speech and music. In the future, we will put these two kinds of sounds as separate categories. Also, the speech with the environmental sound background, where the environmental sound may be treated as noise, is sometimes misclassified as the harmonic and non-harmonic mixed environmental sound. We will continue to improve the classifier so that it has a more robust performance for such a case. It is desirable that the speech signal can be detected when its SNR is not too low (in other words, when the contents of speech can be easily identified by human being).

6.3 Segmentation Result

We tested the segmentation procedures with audio clips recorded from movies. With Pentium166 PC/Windows NT, we can achieve the segmentation and classification together with less than one half of the time required to play the audio clip. It is expected that much less time may be needed when using a more advanced CPU available today. One segmentation example is shown in Figure 8. Nine types of audio (speech, music, silence and six classes of environmental sounds) are represented by different colors. For this 50-second long audio clip, there is first a segment of speech spoken by a female (classified as speech), then a segment of screams by a group of people (classified as non-harmonic and irregular), followed by a period of unrecognizable conversation of multi-people simultaneously mixed with baby cry (classified as the mix of harmonic and non-harmonic sounds), by a segment of music (classied as music) and, finally, by a short conversation between a male and a female (classified as speech). The boundaries are set accurately and each segment is accurately classified.

7 CONCLUSION AND FUTURE WORK

An online audio classification and segmentation system was presented in this paper, where audio recordings are classified and segmented into speech, music, several types of environmental sounds and silence based on audio content analysis. This is the first step of our continuing work towards a general content-based audio classification and retrieval system. We focused on features and procedures for coarse-level segmentation and classification scheme based on morphological and statistical properties of the temporal curves of three short-time features. It is generic and model-free. Tested with an audio database containing about 1500 pieces of sound, it is shown that more than 90% of the audio clips can be correctly classified into one of the basic types, i.e. speech, music, environmental

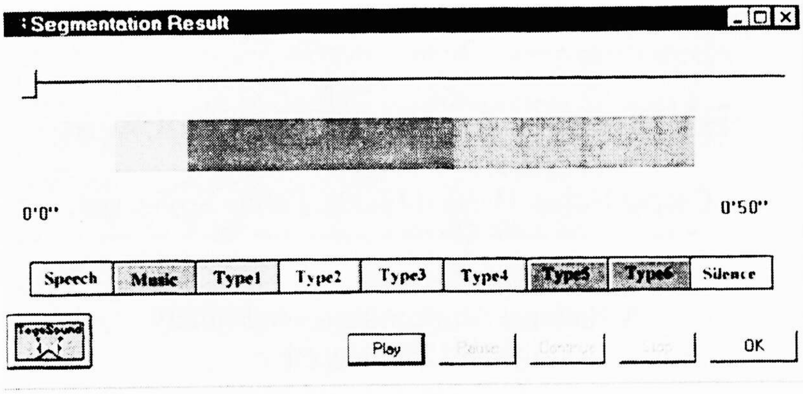


Figure 8: Segmentation of a movie audio clip.

sounds that are further broken down into six classes and silence. For long audio clips consisting of mixed types of sounds, segment boundaries can be accurately found and each segmented result can be properly classified.

In this paper, we mainly described the approach and results of our work on the coarse-level audio classification and segmentation. For the next step, we will work on the improvement of the robustness of both coarse- and fine-level classifications, and build an interface for interactive audio retrieval. The developed audio classification and retrieval techniques will be integrated into a complete system for professional media production, audio/video archive management or surveillance.

8 REFERENCES

[1] J. Saunders: "Real-Time Discrimination of Broadcast Speech/Music", *Proc. ICASSP'96*, vol.II, pp.993-996, Atlanta, May, 1996

[2] E. Scheirer, M. Slaney: "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", *Proc. ICASSP'97*, Munich, Germany, April, 1997

[3] L. Wyse, S. Smoliar: "Toward Content-based Audio Indexing and Retrieval and a New Speaker Discrimination Technique", downloaded from *http://www.iss.nus.sg/People/lwyse/lwyse.html*, Institute of Systems Science, National Univ. of Singapore, Dec., 1995

[4] D. Kimber, L. Wilcox: "Acoustic Segmentation for Audio Browsers", *Proc. Interface Conference*, Sydney, Australia, July, 1996

[5] A. Ghias, J. Logan, D. Chamberlin: "Query By Humming - Musical Information Retrieval in An Audio Database", *Proc. ACM Multimedia Conference*, pp.231-235, Anaheim, CA, 1995

[6] E. Wold, T. Blum, D. Keislar, *et al.*: "Content-Based Classification, Search, and Retrieval of Audio", *IEEE Multimedia*, pp.27-36, Fall, 1996

[7] J. Foote: "Content-Based Retrieval of Music and Audio", *Proc. SPIE'97*, Dallas, 1997

[8] Z. Liu, J. Huang, Y. Wang, *et al.*: "Audio Feature Extraction and Analysis for Scene Classification", *Proc. of IEEE 1st Multimedia Workshop*, 1997

[9] N. Patel, I. Sethi: "Audio Characterization for Video Indexing", *Proc. SPIE on Storage and Retrieval for Still Image and Video Databases*, Vol.2670, pp.373-384, San Jose, 1996

[10] L. Rabinar, R. Schafer: *Digital Processing of Speech Signals*, Prentice-Hall, Inc., New Jersey, 1978

[11] F. Everest: *The Master Handbook of Acoustics*, McGraw-Hill, Inc., 1994