HIERARCHICAL CLASSIFICATION OF AUDIO DATA FOR ARCHIVING AND RETRIEVING

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

A hierarchical system for audio classification and retrieval based on audio content analysis is presented in this paper. The system consists of three stages. The first stage is called the coarse-level audio classification and segmentation, where audio recordings are classified and segmented into speech, music, several types of environmental sounds, and silence, based on morphological and statistical analysis of temporal curves of short-time features of audio signals. In the second stage, environmental sounds are further classified into finer classes such as applause, rain, birds' sound, etc. This fine-level classification is based on time-frequency analysis of audio signals and use of the hidden Markov model (HMM) for classification. In the third stage, the query-by-example audio retrieval is implemented where similar sounds can be found according to an input sample audio. It is shown that the proposed system has achieved an accuracy higher than 90% for coarse-level audio classification. Examples of audio fine classification and audio retrieval are also provided.

1. INTRODUCTION

Audio, which includes voice, music, and various kinds of environmental sounds, is an important type of media, and also a significant part of audiovisual data. As there are more and more digital audio databases in place at present, people start to realize the importance of audio database management relying on audio content analysis.

Content-based audio classification and retrieval have a wide range of applications in the entertainment industry, audio archiving management, commercial musical usage, surveillance, etc. For example, it will be very helpful to be able to search sound effects automatically from a very large audio database in film postprocessing, which contains sounds of explosion, windstorm, earthquake, animals, and so on. There are also distributed audio libraries in the World Wide Web, and content-based audio retrieval could be an ideal approach for sound indexing and search. 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 very limited. There are in general three directions. One direction is audio segmentation and classification. One basic problem is speech/music discrimination. Further classification of audio may take other sounds into consideration, as done in [1], where audio was classified into "music", "speech", and "others". The second direction is audio retrieval. One specific technique here is query-by-humming. For generic audio retrieval, two approaches were presented in [2] and [3], respectively. MFCC of audio signals were taken as features, and a tree-structured classifier was built for retrieval in [2]. It turned out that MFCC do not work well in differentiating audio timbres. In [3], statistical values (means, variances, and autocorrelations) of several time- and frequency-domain measurements were used to represent perceptual features such as loudness, brightness, bandwidth, and pitch. This method is only suitable for sounds with a single timbre. The third direction is audio analysis for video indexing. In [4], audio analysis was applied to the distinction of five kinds of video scenes.

Audio classification and retrieval is an important and challenging research topic, while work in this area is still at a 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. There are several distinguishing features of this system. First, we divide the audio classification task into two steps. In the coarse-level step, speech, music, environmental audio, and silence are separated. This classification is generic and model-free. Then, in the fine-level step, more specific classes of natural and synthetic sounds are distinguished within each basic audio type. Second, compared with previous work, we put more emphasis on the environmental audio, which is often ignored in the past. Environmental sounds are an important ingredient in audio recordings, and their analysis is inevitable in many real applications. Third, the audio retrieval is achieved based on audio classification results, thus obtaining semantic meanings and better reliability. Irrelevant or confusing results, as often appearing in image or audio retrieval systems, are avoided by this way. Finally, we investigate physical and perceptual features of different classes of audio, and apply signal processing techniques (including morphological and statistical analysis methods, heuristic method, clustering method, hidden Markov method, etc.) uniquely to the representation and classification of extracted features. The framework of the proposed system is shown in Figure 1.

The paper is organized as follows. In Section 2, audio fea-

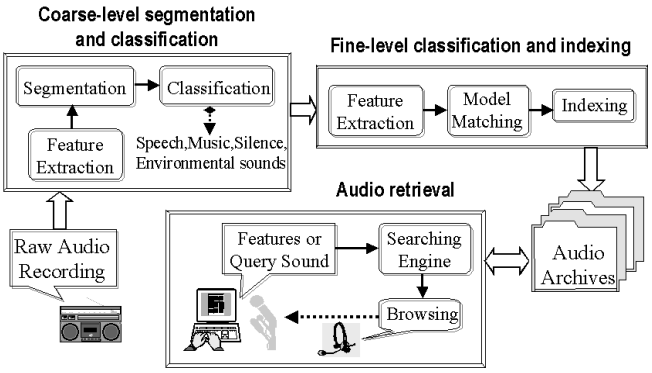

Figure 1: A hierarchical system for content-based audio classification and retrieval.

tures which are important for classification and retrieval are introduced. The procedures for coarse-level audio classification and segmentation, and those for the fine-level audio classification and audio retrieval are described in Sections 3 and 4, respectively. Experimental results are shown in Section 5, and summarizing remarks are given in Section 6.

2. AUDIO FEATURES FOR CLASSIFICATION AND RETRIEVAL

There are two types of audio features: physical features and perceptual features. Physical features refer to mathematical measurements computed directly from the sound wave, such as the energy function, the spectrum, and the fundamental frequency. Perceptual features are subjective terms which are related to the perception of sounds by human beings, including loudness, pitch, timbre, and rhythm. For the purpose of coarse-level classification, we have used temporal curves of three kinds of short-time physical features, i.e., the energy function, the average zero-crossing rate, and the fundamental frequency. For the fine-level classification, one of our most important tasks is to build physical and mathematical models for the perceptual features with which human beings distinguish different classes of sounds. In this work, we consider two kinds of features: timbre and rhythm.

2.1. Physical Features

(1) Short-time Energy Function
The short-time energy of an audio signal is defined as

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

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 of length N. It provides a convenient representation of the amplitude variation over time. For speech signals, it is a basis for distinguishing voiced speech components from unvoiced speech components, as the energy function values for unvoiced components are significantly smaller than those of the voiced components. The energy function can also be used as the measurement to distinguish silence when the SNR is high.

(2) Short-time Average Zero-Crossing Rate (ZCR)
In discrete-time signals, a zero-crossing is said to occur if successive samples have different signs. The short-time average zero-crossing rate, as defined below, gives rough estimates of spectral properties of audio signals.

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

where

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

and $w(m)$ is a rectangle window. It is another measurement to differentiate voiced speech components from unvoiced speech components, as the voiced components have much smaller ZCR values than the unvoiced components. Compared to that of speech, the ZCR curve of music has a remarkablely lower variance and average amplitude. The environmental audio of various origins can be briefly classified according to the differences in ZCR curve properties.

(3) Short-time Fundamental Frequency (FuF)
We define the short-time fundamental frequency to reveal harmonic properties of audio signals.

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

where $w(m)$ is the Hanning window. The operator $fuf\{\cdot\}$ is defined as such that when the sound is harmonic, F_n is equal to the fundamental frequency estimated from the logarithmic spectrum; and when the sound is non-harmonic, F_n is set to zero. Sounds from most musical instruments are harmonic. In speech, 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, or harmonic and non-harmonic mixed.

2.2. Perceptual Features

(1) Timbre
Timbre is generally defined as "the quality which allows one to tell the difference between sounds of the same level and loudness when made by different musical instruments or voices". The problem of building physical models for timbre perception has been investigated for a long time in psychology and music analysis without definite answers. Nevertheless, we may get the conclusion from existing results that the temporal evolution of spectrum of audio signals accounts largely for timbre perception. Here, we extend timbre from a term originally used for harmonic sound (music and voice) to the perception of environmental sound, and analyze it on the time-frequency representation of audio signals. We consider timbre as the most important feature in differentiating different classes of environmental sounds, and to build a model properly for timbre perception based on the spectrogram is one major problem in our research.

(2) Rhythm
Rhythm is a term originally defined for speech and music. It is the quality of happening at regular periods of time. Here, we extend it to environmental sounds to represent the change pattern of timbres in a sound clip. Rhythm is a

significant feature in the perception of sounds like footstep, clock tick, telegraph machine, pager, door knock, etc.

3. COARSE-LEVEL CLASSIFICATION AND SEGMENTATION OF AUDIO

For on-line segmentation and classification of audio recordings, the short-time energy function, short-time average zero-crossing rate, and 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. Each segment is then classified into one of the basic audio types according to a rule-based heuristic procedure. The procedure includes the following steps:

(1) Separating Silence
We define "silence" to be a segment of imperceptible audio, including unnoticeable noise and very short clicks. We use both energy and ZCR measures to detect silence. If the short-time energy function is continuously lower than certain set of thresholds (except for short, sparse clicks) or if most short-time zero-crossing rates are lower than certain set of thresholds, then the segment is indexed as "silence".

(2) Separating Environmental Sounds with Special Features
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". If the fundamental frequency of a sound clip changes over time but only with several values, it is indexed as "harmonic and stable". This step is performed as a screening process for harmonic environmental sounds, so that they will not confuse the differentiation of music. It is also the basis of fine-level classification of harmonic environmental audio.

(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 rates. For each aspect, these is one empirical threshold set and a decision value defined. The four decision values are averaged with certain weights to derive a total probability of the audio segment being music.

(4) Distinguishing Speech
Five aspects are checked to distinguish speech. The first one is the compensative relation between amplitudes of ZCR and energy curves. The second one is the shape of ZCR curve. For speech, the ZCR curve has a stable and low baseline with peaks above it. The third and fourth aspects are the variance and the range of the amplitude of the ZCR curve, respectively. And the fifth aspect is about the property of the short-time fundamental frequency. A decision value, which is a fraction between 0 and 1, is defined for each of the aspects. The weighted average of these decision values gives the possibility of the segment being speech.

(5) Classifying Other Environmental Sounds
The last step is to classify what is left into one type of the non-harmonic environmental sounds: (a) "periodic or quasi-periodic" when either the energy curve or the ZCR curve has approximately periodic peaks; (b) "harmonic and non-harmonic mixed" when the percentage of harmonic sound is within a certain range; (c) "non-harmonic and stable" when the frequency centroid is within a relatively small range compared to the range of the frequency distribution; and (d) "non-harmonic and irregular" when the segment does not satisfy any of the above conditions.

Finally, a post-processing procedure is applied to reduce possible segmentation errors. For more details of these processes, we refer to [5].

4. FINE-LEVEL AUDIO CLASSIFICATION AND AUDIO RETRIEVAL

The core of fine-level audio classification is to build hidden Markov model (HMM) for each class of sounds. Currently, two types of information are contained in HMM: timbre and rhythm. Each kind of timbre is modeled as one state of HMM, and represented with the Gaussian mixture density. The rhythm information is denoted by transition and duration parameters in HMM. Once HMM parameters are set, sound clips can be classified into available classes by matching to models of these classes. The processes are briefly introduced below, while more details can be found in [6].

(1) Feature Extraction
A key point in modeling timbre perception with HMM is to extract feature vector from the short-time spectrum. Up to now, we have used the most direct way to extract features from the frequency distribution, i.e. to use the spectrum coefficients themselves. Taking 128-point FFT of audio signal, we obtain a feature vector of 65 dimensions (i.e., the logarithm of amplitude spectrum at each frequency sample between 0 and π) at each sampled time instant.

(2) Clustering
The feature vectors of one class of sounds are clustered into several sets, with each set denoting one kind of timbre, and modeled later by one state in HMM. We adopted an adaptive sample set construction method for clustering with some modifications. It works well for clustering feature vectors. For example, the sound of dog bark is clustered into three states: bark, intermission, and the transition period in between.

(3) Building Model
The hidden Markov model with continuous observation densities and explicit state duration densities [7] is used to model each class of sound. We denote the complete parameter set of HMM as $\lambda = (A, B, D, \pi)$, with A for the transition probability, B for observation density parameters, D for duration density parameters, and π for initial state distribution. A simplified procedure is taken to train the parameters. First, the parameters of observation density, which takes the form of a Gaussian mixture, are estimated for each state, respectively, through an ML iteration process. Then, the transition probability matrix A and the duration density parameters are calculated statistically according to the state indexes of feature vectors

in the training set. The initial state distribution is set as $\pi_i = 1/N,\ 1 \le i \le N$, where N is the number of states.

(4) Classification
Assume that there are K classes of sounds modeled with parameter sets $\lambda_i,\ 1 \le i \le K$. For a piece of sound to be classified, feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ are extracted. Then, the HMM likelihoods $P_i(\mathbf{X}|\lambda_i),\ 1 \le i \le K$, are computed. Choose the class j which maximizes P_i, i.e. $j = \arg\max\{P_i, 1 \le i \le K\}$, and the sound is classified into this class.

(5) Audio Retrieval
HMM is built for each sound clip in the audio database in the query-by-example audio retrieval. With an input query sound, its feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ are extracted, and the possibilities $P(\mathbf{X}|\lambda_i),\ 1 \le i \le L$ are computed, where L is the number of sound clips in the database and λ_i denotes the HMM parameter set for the ith sound clip. A rank list of audio samples in terms of similarity with the input query are obtained by comparing values of $P(\mathbf{X}|\lambda_i)$. The user may listen to one sound, and be asked "want more like this?". As the database is organized based on audio classification results, if he likes this retrieved sound, the class to which this sound belongs can be retrieved. Sounds within the class will be ordered according to similarity with the query sound, and presented to the user.

5. EXPERIMENTAL RESULTS

5.1. Audio Database and Coarse-level Classification

We have built a generic audio database which includes about 1500 pieces of sound of various types to test the classification and retrieval algorithms. We also collected dozens of audio clips recorded from movies for testing the segmentation performances. The proposed coarse-level classification scheme achieves an accuracy rate of more than 90% with this audio database. Misclassification usually occurs in the hybrid sound which contains more than one basic type of audio. When testing with movie audio recordings, the segmentation and classification together can be achieved in real time. The boundaries are set accurately and each segment is properly classified.

5.2. Example of Fine-level Classification

For a brief test of the fine-level classification algorithm, we built the HMM parameter set for ten classes of sounds, including applause, birds' cry, dog bark, explosion, foot step, laugh, rain, river flow, thunder, and windstorm. Feature vectors extracted from 6-8 sound clips were used for building the model for each class. Then, fifty new sound clips (with five pieces of sound in each class) were used to test the classification accuracy. It turned out that 41 out of the 50 sound clips were correctly classified, achieving an accuracy rate of over 80%. Misclassification happened among classes having perceptually similar sounds, such as applause, rain, river flow, and windstorm.

5.3. Example of Audio Retrieval

In an experiment of audio retrieval, 100 pieces of sound from 15 classes were selected to form a small database, with the HMM parameter set trained for each piece of sound. We chose a sound clip of applause as the query sound, and matched it to each of the 100 HMMs. The resulting top ten sounds in the rank list belonged to the following classes: no.1-5: applause; no.6: rain; no.7-9: applause; no.10: rain. This result is reasonable, as the pouring rain and the applause by a crowd of people sometimes really sound alike. For another example, a sound clip of plane taking off was used as the input query, and the top ten retrieved sounds were: no.1-6: plane; no.7-10: rain. The only 6 pieces of plane sound in the database were ranked at the first 6 places, while the next 4 were taken by sounds of large rain.

6. SUMMARY

A hierarchical system for audio classification and retrieval based on audio content analysis and modeling was presented in this paper. The coarse-level classification is generic and model free, and achieved an accuracy rate of more than 90% tested with our audio database. For fine-level classification and audio retrieval, we focused on modeling environmental sound with the hidden Markov model. Preliminary experiments showed that accuracy rate of over 80% can be achieved with the proposed fine classification method. Results of audio retrieval also proved the HMM-based approach to be promising. Future work will be done to refine the proposed system. First, we will enhance the coarse-level classification by taking hybrid-type sound and noisy sound into consideration. Second, we will look for more effective feature extraction method in the fine-level classification.

7. REFERENCES

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

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

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

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

[5] T. Zhang, C.-C. Kuo: "Content-based Classification and Retrieval of Audio", *SPIE's Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII,* San Diego, July, 1998

[6] T. Zhang, C.-C. Kuo: "Hierarchical System for Content-based Audio Classification and Retrieval", *SPIE's Conference on Multimedia Storage and Archiving Systems III,* Boston, Nov., 1998

[7] L. Rabinar, B. Juang: *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., New Jersey, 1993