International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 19, No. 1 (2005) 63–78

World Scientific
www.worldscientific.com

A NEW APPROACH FOR CLASSIFICATION OF GENERIC AUDIO DATA*

RUEI-SHIANG LIN and LING-HWEI CHEN†

Department of Computer and Information Science
National Chiao Tung University, 1001 Ta Hsueh Rd.
Hsinchu, Taiwan 30050, R.O.C.
†*lhchen@cc.nctu.edu.tw*

The existing audio retrieval systems fall into one of two categories: single-domain systems that can accept data of only a single type (e.g. speech) or multiple-domain systems that offer content-based retrieval for multiple types of audio data. Since a single-domain system has limited applications, a multiple-domain system will be more useful. However, different types of audio data will have different properties, this will make a multiple-domain system harder to be developed. If we can classify audio information in advance, the above problems can be solved. In this paper, we will propose a real-time classification method to classify audio signals into several basic audio types such as pure speech, music, song, speech with music background, and speech with environmental noise background.

In order to make the proposed method robust for a variety of audio sources, we use Bayesian decision function for multivariable Gaussian distribution instead of manually adjusting a threshold for each discriminator. The proposed approach can be applied to content-based audio/video retrieval. In the experiment, the efficiency and effectiveness of this method are shown by an accuracy rate of more than 96% for general audio data classification.

Keywords: Audio classification; spectrogram; Bayesian decision function; multivariable Gaussian distribution.

1. Introduction

Audio classification[1,3–7,9,11–17] has many applications in professional media production, audio archive management, commercial music usage, content-based audio/video retrieval, and so on. Several audio classification schemes have been

*This research was supported in part by the Department of Industrial Technology of R.O.C. under contract 93-EC-17-A-02-S1-032 and the National Science Council of R.O.C. under contract NSC-92-2213-E-009-101.
†Author for correspondence.

proposed. These methods tend to roughly divide audio signals into two major distinct categories: speech and music. Scherier and Slaney[13] provided such a discriminator. Based on thirteen features including cepstral coefficients, four multidimensional classification frameworks are compared to achieve better performance. The approach presented by Saunders[12] takes a simple feature space and is performed by exploiting the distribution of zero-crossing rate. In general, speech and music have quite different properties in both time and frequency domains. Thus, it is not hard to reach a relatively high level of discrimination accuracy. However, two-type classification for audio data is not enough in many applications, such as content-based video retrieval.[1] Recently, video retrieval has become an important research topic. To raise the retrieval speed and precision, a video is usually segmented into several scenes.[1,17] In general, neighboring scenes will have different types of audio data. Thus, if we can develop a method to classify audio data, the classified results can be used to assist scene segmentation. Different kinds of videos will contain different types of audio data. For example, in documentaries, commercials or news report, we can usually find the following audio types: speech, music, speech with musical or environmental noise background and song.

Wyse and Smoliar[15] presented a method to classify audio signals into "music", "speech", and "others". The method was developed for the parsing of news stories. In Ref. 5, audio signals are classified into speech, silence, laughter, and non-speech sounds for the purpose of segmenting discussion recordings in meetings. The abovementioned approaches are developed for specific scenarios, only some special audio types are considered. The research in Refs. 6, 16 and 17 has taken more general types of audio data into account. In Ref. 6, 143 features are first studied for their discrimination capability. Then, the cepstral-based features such as Mel-frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC), etc. are selected to classify audio signals. The authors concluded that in many cases, the selection of features is actually more critical to the classification performance. More than 90% accuracy rate is reported. Zhang and Kuo[17] first extracted some audio features including the short-time fundamental frequency and the spectral tracks by detecting the peaks from the spectrum. The spectrum is generated by autoregressive model (AR model) coefficients, which are estimated from the autocorrelation of audio signals. Then, the rule-based procedure, which uses many threshold values, is applied to classify audio signals into speech, music, song, speech with music background, etc. More than 90% accuracy rate is reported. The method is time-consuming due to the computation of autocorrelation function. Besides, many thresholds used in this approach are empirical, they are improper when the source of audio signals is changed. To avoid these disadvantages, in this paper, we will provide a method with only few thresholds used to classify audio data into five general categories: pure speech, music, song, speech with music background and speech with environmental noise background. These categories are the basic sets needed in the content analysis of audiovisual data.

The proposed method consists of three stages: feature extraction, the coarse-level classification and the fine-level classification. Based on statistical analysis, four effective audio features are first extracted to ensure the feasibility of real-time processing. They are the energy distribution model, variance and the third moment associated with the horizontal profile of the spectrogram and the variance of the differences of temporal intervals. Then, the coarse-level audio classification based on the first feature is conducted to divide audio signals into two categories: single-type and hybrid-type, i.e. with or without background components. Finally, each category is further divided into finer subclass through Bayesian decision function.[2] The single-type sounds are classified into speech and music; the hybrid-type sounds are classified into speech with environmental noise background, speech with music background and song. The system diagram is shown in Fig. 1. Experimental results show that the proposed method achieves an accuracy rate of more than 96% in audio classification.

The paper is organized as follows. In Sec. 2, the proposed method will be described. Experimental results and discussion will be presented in Sec. 3. Finally, the conclusions will be given in Sec. 4.

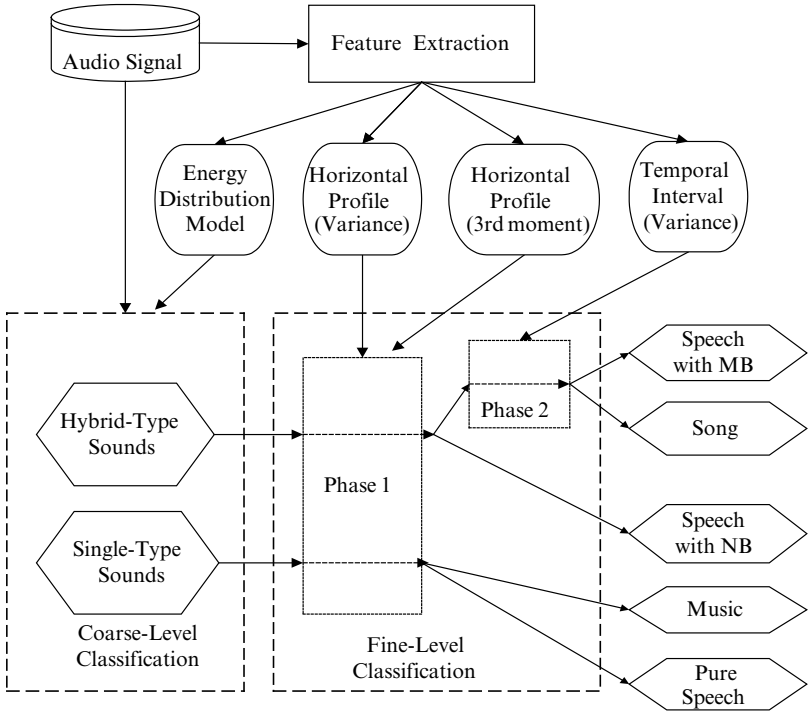


Fig. 1. Block diagram of the proposed system, where "MB" and "NB" are the abbrevations for "music background" and "noise background", respectively.

2. The Proposed System

In this paper, an input audio clip is transformed to a spectrogram that is the time-varying spectrum of a signal and in a three-dimensional (time, frequency, intensity) space known as a time-frequency distribution.[10] To construct a spectrogram, the Short Time Fourier Transform (STFT) is applied. As for STFT, the input audio signal is first divided into several frames. Each frame contains consecutive n audio signal samples, and two neighboring frames will overlap 50%. Then, the Fourier transform is applied to each frame tapered with a window function in succession. Let $s(t)$ denote the audio signal and $STFT(\tau, \omega)$ be the result of STFT, that is,

$$STFT(\tau, \omega) = \sum_{t=0}^{n-1} s\left(t + \frac{\tau}{2}n\right) r^*(t) e^{-j\omega t}, \tag{1}$$

where $r^*(t)$ is the window function, τ stands for the frame number, n is set to 512 and ω is the frequency parameter. Then, the spectrogram, $S(\tau, \omega)$, is the energy distribution associated with the Short Time Fourier Transform, that is,

$$S(\tau, \omega) = 10 \log_{10} \left(\frac{|STFT(\tau, \omega)|}{M} \right)^2,$$

where

$$M = \max_{\tau, \omega} |STFT(\tau, \omega)|. \tag{2}$$

Traditionally, a spectrogram is displayed with gray levels, where the darkness of a given point is proportional to its energy. The vertical axis in a spectrogram represents frequency and the horizontal axis represents time (or frame). Figures 2(a)–2(e) show five examples of the spectrograms of music, speech with music background, song, pure speech and speech with environmental noise background, respectively.

The proposed audio classification method is based on the spectrogram and consists of three phases: feature extraction, the coarse-level classification and the fine-level classification. First, four effective audio features are extracted. Then, based on the first feature, the coarse-level audio classification is conducted to classify audio signals into two categories: single-type and hybrid-type. Finally, based on the remaining features, each category is further divided into finer subclasses. The single-type sounds are classified into pure speech and music. The hybrid-type sounds are classified into song, speech with environmental noise background and speech with music background.

In the following, the proposed method will be described in details.

2.1. *Feature extraction phase*

Four kinds of audio features are used in the proposed method, they are energy distribution model, variance and the third moment associated with the horizontal profile of the spectrogram, and variance of the differences of temporal intervals (which will be defined later). To get these features, the audio spectrogram for an

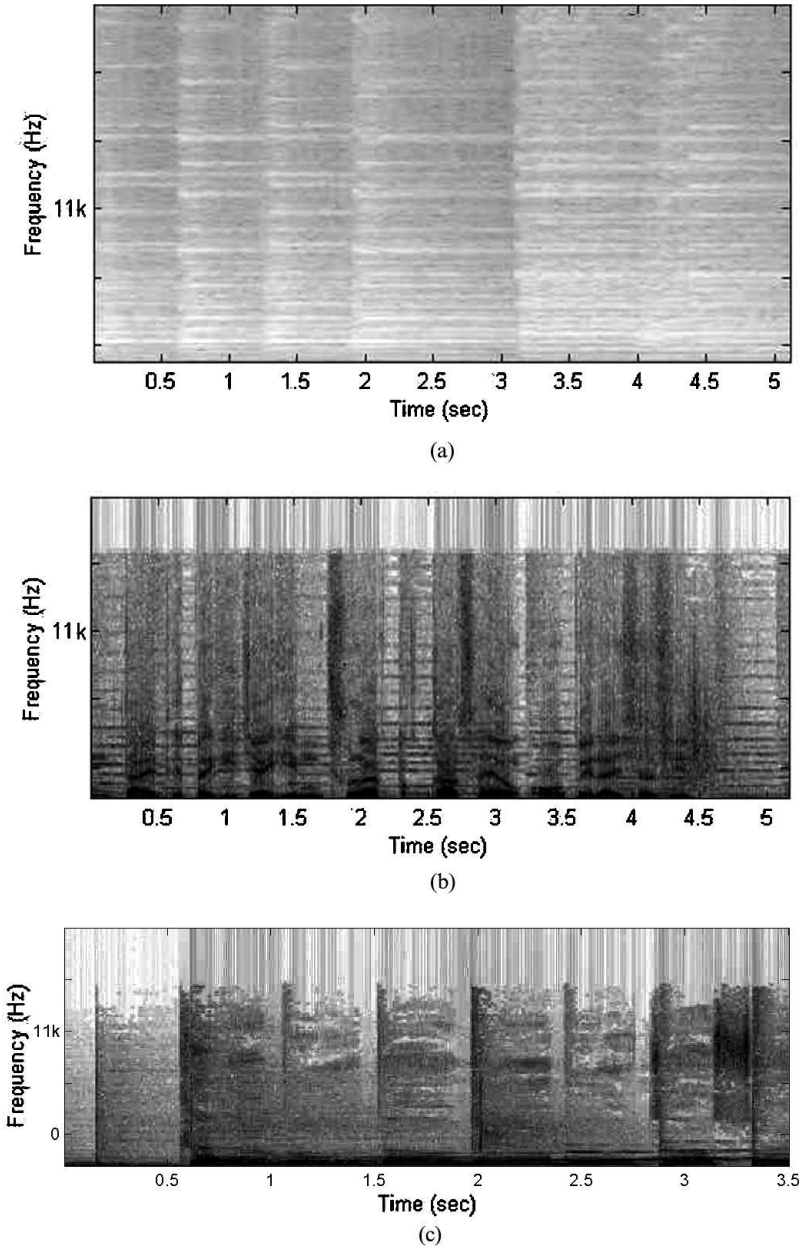


Fig. 2. Five spectrogram examples. (a) Music. (b) Speech with music background. (c) Song. (d) Speech. (e) Speech with environmental noise background.

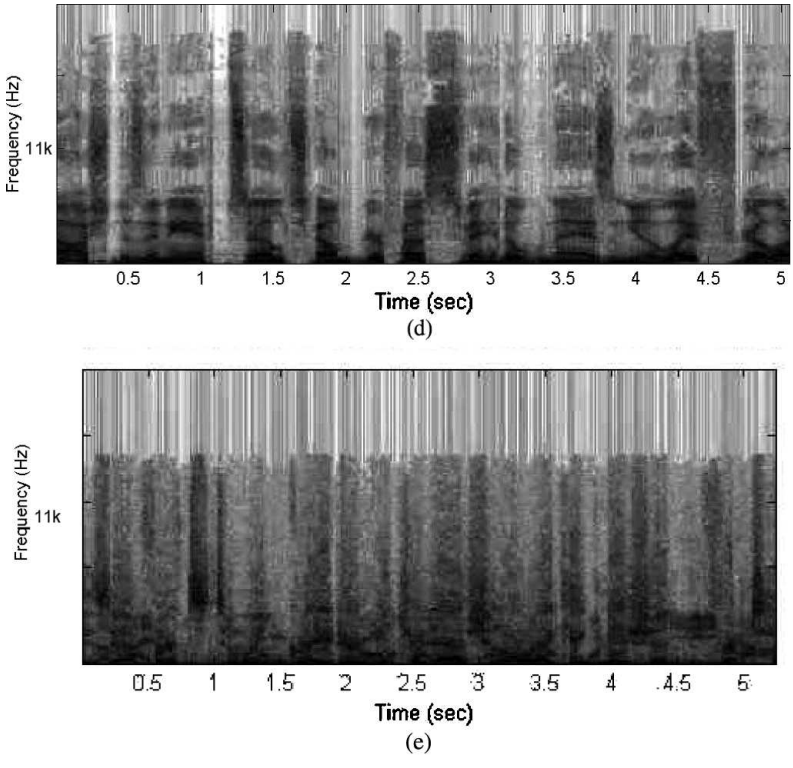


Fig. 2. (*Continued*)

audio signal is constructed first. Based on the spectrogram, these four features are extracted and described as follows.

2.1.1. *The energy distribution model*

For the purpose of characterizing single-type and hybrid-type sounds, i.e. with or without background components, the energy distribution model is proposed. The histogram of a spectrogram is also called the energy distribution of the corresponding audio signal. In our experiments, we found that there are two kinds of energy distribution models: unimodel and bimodel [see Figs. 3(a) and 3(b)], in audio signals. In Fig. 3, the horizontal axis represents the spectrogram energy.

For a hybrid-type sound, its energy distribution model is bimodel; otherwise, it is unimodel. Thus, to discriminate single-type sounds from hybrid-type sounds, we only need to detect the type of the corresponding energy distribution model. To reach this, for an audio signal, the histogram of its corresponding spectrogram, $h(i)$, is established first. Then, the mean μ and the variance σ^2 of $h(i)$ are calculated. In general, if μ approaches to the position of the highest peak in h, $h(i)$ will be a unimodel [see Fig. 3(a)]. On the other hand, for a bimodel, dividing $h(i)$ into two parts from μ, each part will be unimodel [see Fig. 3(b)]. Thus, if we find a local

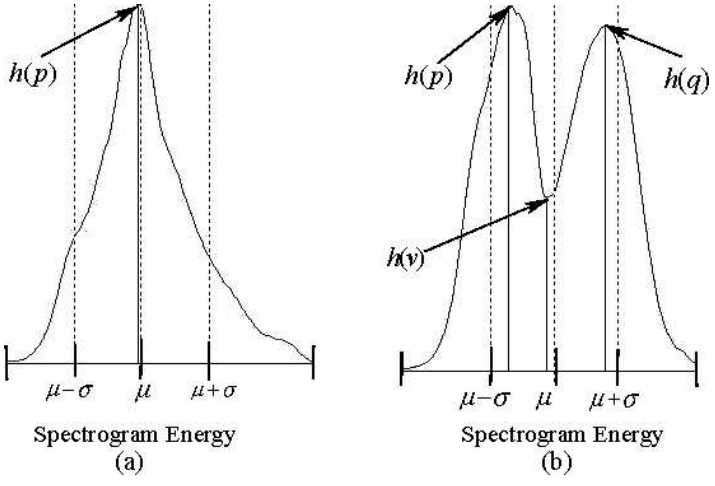


Fig. 3. Two examples of the energy distribution models. (a) Unimodel [the histogram of the energy distribution of Fig. 2(a)]. (b) Bimodel [the histogram of the energy distribution of Fig. 2(c)].

peak in each part, these two peaks will not be close. Based on these phenomena, a model decision algorithm is provided and described as follows.

Algorithm 1. Model decision Algorithm

Input: The spectrogram $S(\tau, \omega)$ of an audio signal.
Output: The model type, T, and two parameters *T1*, *T2*.
Step 1. Establish the histogram, $h(i)$, $i = 0, \ldots, 255$, of $S(\tau, \omega)$.
Step 2. Compute the mean μ and the variance σ^2 of $h(i)$.
Step 3. Find the position p of the highest peak in $h(i)$.
Step 4. If $|p - \mu| \leq 5$, $T =$ unimodel, go to Step 9.
Else
Use μ to set the search range $\Re_p$ as follows:

$$\Re_p = \begin{cases} (\mu, \mu + \sigma], & \text{if } p < \mu \\ [\mu - \sigma, \mu), & \text{if } p > \mu \end{cases}.$$

End if.
Step 5. Find the position q of the highest peak $h(q)$within $\Re_p$.
Step 6. Find the position v of the lowest valley $h(v)$ in the range between p and q.
Step 7. Set $dst = |p - q|$.
Step 8. Set $T =$ bimodel if the following two conditions are satisfied
Condition 1: $dst \geq \frac{\sigma}{2}$.
Condition 2: $h(q) \geq \frac{1}{2}h(p)$ and $h(q) \geq \frac{6}{5}h(v)$.
Else $T =$ unimodel.
Step 9. Output T and assign μ to *T1*, $\mu + \sigma$ to *T2*.

End of *Algorithm 1.*

Through the model decision algorithm described above, the model type for an audio signal can be determined. Note that in the algorithm, except the model type extracted, two parameters, *T1* and *T2*, which will be used later, will be also obtained.

2.1.2. *The horizontal profile analysis*

In this section, we will base on two facts to discriminate an audio clip with or without music components. One fact is that if an audio clip contains musical components, we can find many horizontal long-line like tracks [see Figs. 2(a)–2(c)] in its spectrogram. The other fact is that if an audio clip does not contain musical components, most energy in the spectrogram of each frame will concentrate on a certain frequency interval [see Figs. 2(d)–2(e)]. Based on these two facts, two novel features will be derived and used to distinguish music from speech.

To obtain these features, the horizontal profile of the audio spectrogram is constructed first. Note that the horizontal profile [see Figs. 4(a)–4(e)] is defined as the projection of the spectrogram of the audio clip on the vertical axis. Based on the first fact, we can find that for an audio clip with musical components, there will be many peaks in its horizontal profile [see Figs. 4(a)–4(c)], and the location difference between two adjacent peaks is small and near constant. On the other hand, based on the second fact, we can see that for an audio clip without musical components, only few peaks can be found in its horizontal profile [see Figs. 4(d)–4(e)], and the location difference between any two successive peaks is larger and variant. Based on the above description, for an audio clip, all peaks, P_i, in its horizontal profile are first extracted; and the location difference, dP_i, between any two successive peaks is evaluated. Note that in order to avoid the influence of noise in high frequency, the frequency components above $Fs/4$ are discarded, where Fs is the sampling rate.

Then the variance, v_{dP_i}, and the third moment, m_{dP_i}, of $dP_i s$ are taken as the second and third features and used to discriminate audio clips with or without music components. Note that variance and the third moment stand for the spread and skewness of the location differences of all two successive peaks in the horizontal profile, respectively. For an audio clip with musical components, variance and the third moment will be small; however, for an audio clip without musical component, these two features will be larger.

2.1.3. *The temporal intervals*

Up to now, we have provided three features. By processing the audio signals through these features, all audio signals can be classified successfully except the simultaneous speech and music category, which contains two kinds of signals: speech with music background and song. To discriminate these, a new feature is provided. One important characteristic to distinguish them is the duration of the music-voice.

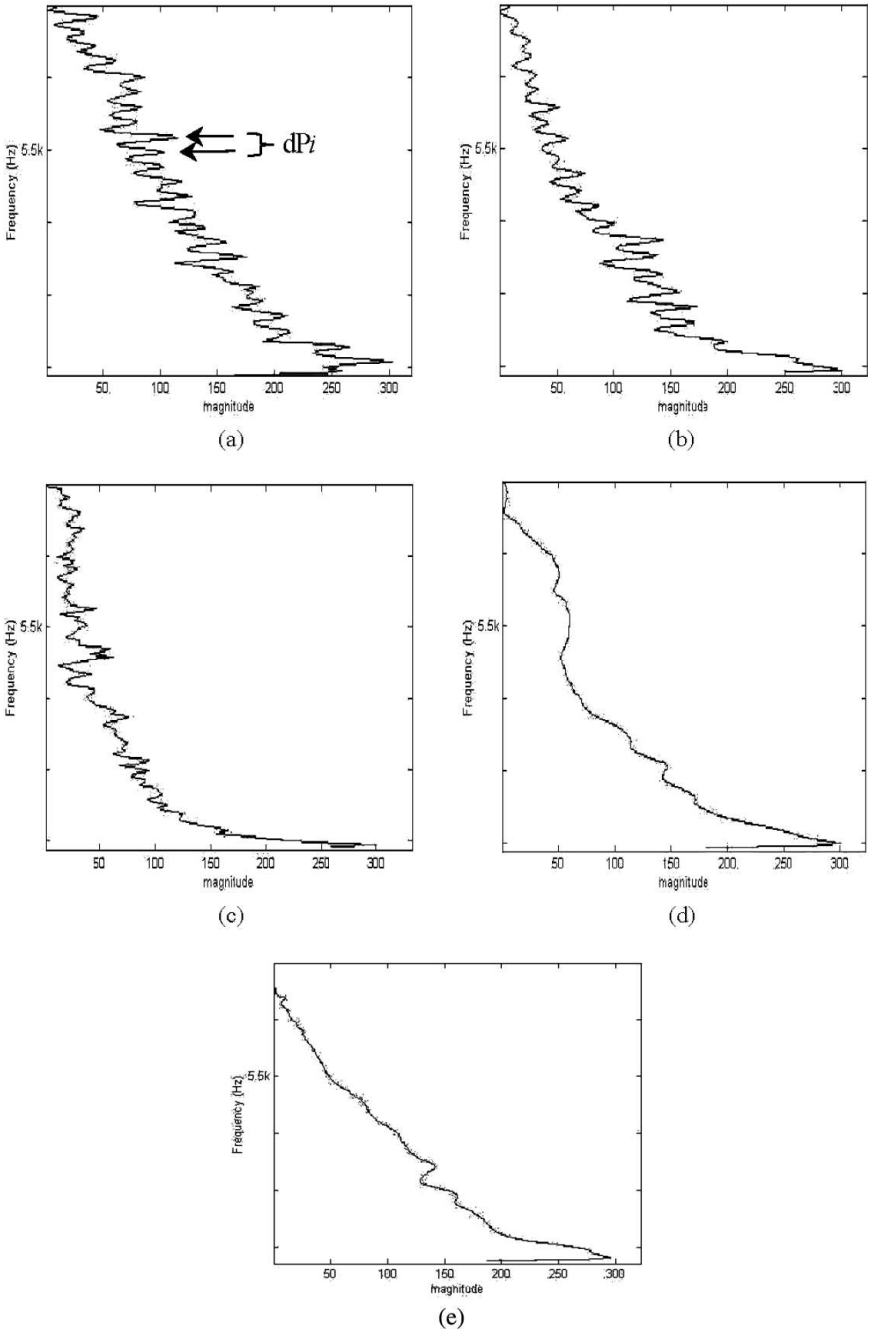


Fig. 4. Five examples of the horizontal profiles. (a–e) are the horizontal profiles of Figs. 2(a)–2(e), respectively.

The duration of music-voice is defined as the duration of music appearing with human voice simultaneously. That is, two successive durations of music-voice are separated by the duration of a pure music component. For speech with music background, in order to emphasize the message of the talker, the signal energy contribution of voice is greater than the contribution of the music. In general, it is strongly speech-like, the difference between any two adjacent duration of music-voice is variable [see Fig. 5(c)]. Conversely, song is usually melodic and rhythmic, the difference between any two adjacent durations of music-voice in song is small and near constant [see Fig. 5(a)].

By observing the spectrogram in different frequency bands, we can see that music-voice (i.e. speech and music appear simultaneously) has more energy in the neighboring middle frequency bands, while music without voice will possess more energy in the lower frequency band. These phenomena are shown in Fig. 5.

Based on these phenomena, the property of the duration of each continuous part of the simultaneous speech and music in a sound is used to discriminate the speech with music background from song. First, a novel feature associated with the temporal interval is derived. This interval is defined as the duration of a continuous part of music-voice of a sound. Note that the signal between two adjacent temporal intervals will be music without human voice. Based on the phenomenon of the energy distribution in different frequency bands described previously, an algorithm will be proposed to determine the continuous music-voice parts in a sound. Note that some frequency noises usually exist in an audio clip, i.e. these noises will contribute to those frequencies with lower energy in spectrogram. In order to avoid the influence of frequency noise, a filtering procedure is applied in advance to get rid of those with lower energy. The proposed filtering procedure is provided and described as follows.

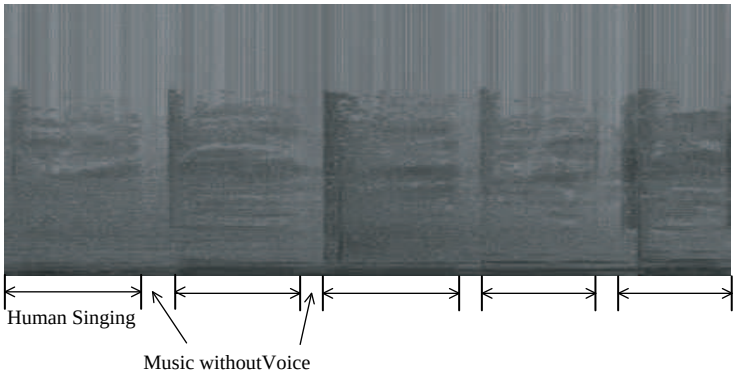
Filtering Procedure:

(1) *Filter out the higher frequency components with lower energy*:
For the spectrogram of each frame τ, $S(\tau, \omega)$, find the highest frequency ω_h with $S(\tau, \omega_h) > T2$. Set $\hat{S}(\tau, \omega) = 0, \forall\, \omega > \omega_h$.

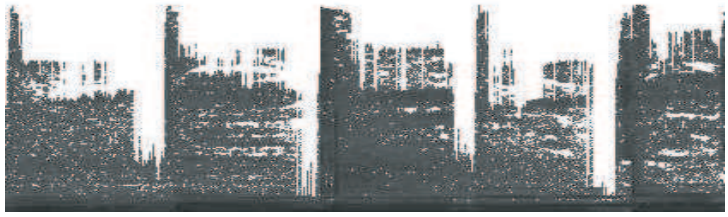
(2) *Filter other components*:
For $\omega < \omega_h$, $\hat{S}(\tau, \omega) = \begin{cases} 0, & \text{if } S(\tau, \omega) < T1 \\ S(\tau, \omega), & \text{otherwise.} \end{cases}$

Figures 5(b) and 5(d) show the filtered spectrograms of Figs. 5(a) and 5(c), respectively. In what follows, we will be interested in how to determine the temporal intervals.

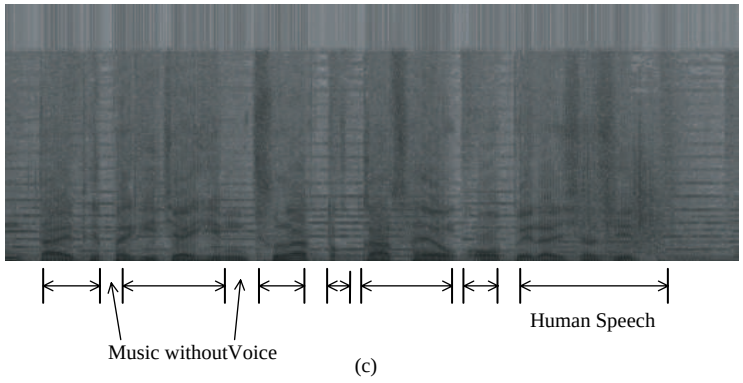
Note that an audio clip of the simultaneous speech and music category contains several temporal intervals and some short periods of background music, each will separate two temporal intervals [see Fig. 5(a)]. To extract temporal intervals, the entire frequency band $[0, Fs/2]$ is first divided into two subbands of unequal width:



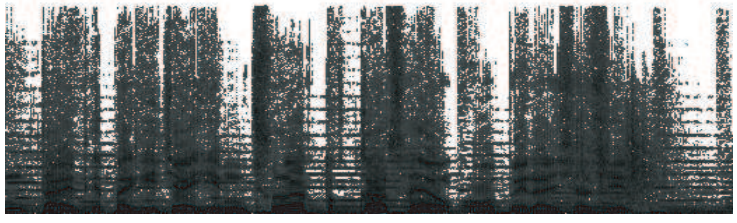
(a)



(b)



(c)



(d)

Fig. 5. Two examples of the filtered spectrogram. (a) The spectrogram of song. (b) The filtered spectrogram of (a). (c) The spectrogram of speech with music background. (d) The filtered spectrogram of (c).

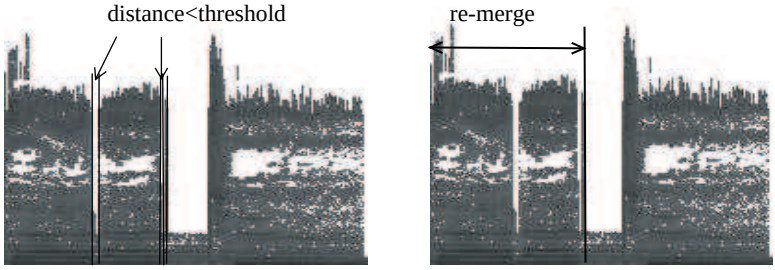


Fig. 6. An example of the remerged process. (a) Initial temporal intervals. (b) Result after remerged process.

$[0, Fs/8]$ and $[Fs/8, Fs/2]$. Next, for each frame, evaluate the ratio of the nonzero part in each subband to the total nonzero part. If the ratio is larger than 10%, mark the subband. Based on the marked subbands, we can extract the temporal intervals. First, those neighboring frames with the same marked subbands are merged to form a group. If the higher subband (i.e. $[Fs/8, Fs/2]$) in a group is marked, the group will be regarded as a part of music-voice (also called raw temporal interval). That is, a temporal interval is a sequence of frames with higher energy in higher subband.

Since the results obtained after filtering procedure are usually sensitive to unvoiced speech and slight breathing, a remerged process is then applied to the raw temporal intervals. During the remerged process, two neighboring intervals are merged if the distance between them is less than a threshold. Figure 6 shows an example of the remerge process. Once we complete this step, we will obtain a set of temporal intervals and the duration difference between any two successive intervals is evaluated. Finally, the variance of these differences, v_{dt}, is taken as the last feature.

2.2. *Audio classification*

Since there are some similar properties among most of the five classes considered, it is hard to find distinguishable features for all of these five classes. To treat this problem, a hierarchical system is proposed. It will do coarse-level classification first, then the fine-level classification is performed. To meet the aim of online classification, features described above are computed on the fly with incoming audio data.

2.2.1. *The coarse-level classification*

The aim of coarse-level audio classification is to separate the five classes into two categories such that we can find some distinguishable features in each category. Based on the energy distribution model, audio signals can be first classified into

two categories: single-type and hybrid-type, i.e. with or without background components. Single-type sounds contain pure speech and music. And hybrid-type sounds contain song, speech with environmental noise background and speech with music background.

2.2.2. *The fine-level classification*

The coarse-level classification stage yields a rough classification for audio data. To get the finer classification result, the fine-level classifier is used. Based on the extracted feature vector X, the classifier is designed using a Bayesian approach under the assumption that the distribution of the feature vectors in each class w_k is a multidimensional Gaussian distribution $N_k(m_k, C_k)$. The Bayesian decision function[2] for class w_k, $d_k(X)$ has the form:

$$d_k(X) = \ln P(w_k) - \frac{1}{2}\ln|C_k| - \frac{1}{2}(X - m_k)^T C_k^{-1}(X - m_k), \tag{3}$$

where m_k and C_k are the mean vector and covariance matrix of X, and $P(w_k)$ is *a priori* probability of class w_k. For a piece of sound, if its feature vector X satisfies $d_i(X) > d_j(X)$ for all $j \neq i$, it is assigned to class w_i.

The fine-level classifier consists of two phases. During the first phase, we take (v_{dP_i}, m_{dP_i}) as the feature vector X and apply Bayesian decision function to each of the two coarse-level classes separately. For each audio signal of the single-type class, we can successfully classify it as music or pure speech. And the classification is well done without needing any further processing. For those of the hybrid-type sounds, which may be speech with environmental noise background, speech with music background or song, the same procedure is applied. Speech with environmental noise background is distinguished and what is left in the first phase is the subclass including speech with music background and song. An additional process is needed to do further classification for the subclass. For this, the Bayesian decision function with the feature v_{dt} is applied. And we can successfully classify each signal in this subclass as speech with music background or song.

3. Experimental Results

3.1. *Audio database*

In order to compare, we have collected a set of 700 generic audio pieces of different types of sound according to the collection rule described in Ref. 17 as the testing database. Care was taken to obtain a wide variation in each category, and most clips are taken from MPEG-7 content set.[8,17] For single-type sounds, there are 100 pieces of classical music played with varied instruments, 100 other music pieces of different styles (jazz, blues, light music, etc.), and 200 clips of pure speech in different languages (English, Chinese, Japanese, etc.). For hybrid-type sounds, there are 200 pieces of song sung by male, female, or children, 50 clips of speech with background music (e.g. commercials, documentaries, etc.), and 50 clips of speech

with environmental noise (e.g. sport broadcast, news interview, etc.). These audio clips (with duration from several seconds to no more than half minute) are stored as 16-bit per sample with 44.1 kHz sampling rate in the WAV file format.

3.2. *Classification results*

Tables 1 and 2 show the results of the coarse-level classification and the final classification results, respectively. From Table 2, it can be seen that the proposed classification approach for generic audio data can achieve an accuracy rate of more than 96% by using the testing database. The training is done using 50% of randomly selected samples in each audio type, and the test is operated on the remaining 50%. By changing training set several times and evaluating the classification rates, we find that the performance of the system is stable and independent on the particular test and training sets. Note that the experiments are carried out on a Pentium II 400 PC/Windows 2000, and one twentieth of the time is required to play the audio clip for processing an audio clip. The only computationally expensive part is the spectrogram, and the other processing is simple by comparison (e.g. variances, peak finding, etc.). In order to compare, we also like to cite the efficiency of the existing system described in Ref. 17, which also includes the five audio classes considered in our method and uses similar database to ours. The authors of Ref. 17 report that less than one eighth of the time required to play the audio clip are needed to process an audio clip. They also report that their accuracy rates are more than 90%.

Table 1. Coarse-level classification results.

Audio Type		Number	Correct Rates (%)
Single-type sounds	Pure speech	200	100
	Pure music	200	100
Hybrid-type sounds	Song	200	100
	Speech with MB	50	100
	Speech with NB	50	100

Table 2. Final classification results.

Audio Type		Number	Correct Rates (%)
Single-type sounds	Pure speech	200	100
	Pure music	200	97.6
Hybrid-type sounds	Song	200	98.53
	Speech with MB	50	96.5
	Speech with NB	50	100

4. Conclusion

In this paper, we have presented a new method for the automatic classification of generic audio data. An accurate classification rate higher than 96% was achieved. Two important and distinguishing features compared with previous work in the proposed scheme are the complexity and running time. Although the proposed scheme covers a wide range of audio types, the complexity is low due to the easy computing of audio features, and this makes online processing possible.

Besides the general audio types such as music and speech tested in existing work, we have taken hybrid-type sounds (speech with music background, speech with environmental noise background and song) into account. While current existing approaches for audio content analysis are normally developed for specific scenarios, the proposed method is generic and model free. Thus, our method can be widely applied to many applications.

References

1. 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, Seattle, pp. 3741–3744.
2. S.-T. Bow, *Pattern Recognition and Image Preprocessing* (Marcel Dekker, 1992).
3. J. Foote, An overview of audio information retrieval, *ACM Multimed. Syst.* **7**(1) (1999) 2–11.
4. I. Fujinaga, Machine recognition of timbre using steady-state tone of acoustic instruments, in *Proc. ICMC 98*, 1998, Ann Arbor, Michigan, pp. 207–210.
5. D. Kimber and L. Wilcox, Acoustic segmentation for audio browsers, in *Proc. Interface Conf.*, July 1996, Sydney, Australia.
6. D. Li, I. K. Sethi, N. Dimitrova and T. McGee, Classification of general audio data for content-based retrieval, *Patt. Recogn. Lett.* **22**(5) (2001) 533–544.
7. G. Lu and T. Hankinson, A technique towards automatic audio classification and retrieval, in *Proc. Int. Conf. Signal Processing'98*, Vol. 2, 1998, pp. 1142–1145.
8. MPEG Requirements Group, Description of MPEG-7 content set, Doc. ISO/MPEG N2467, MPEG Atlantic City Meeting, October 1998.
9. S. Pfeiffer, S. Fischer and W. Effelsberg, Automatic audio content analysis, in *Proc. ACM Multimedia'96*, April 1996, Boston, MA, pp. 21–30.
10. S. Qian and D. Chen, *Joint Time-Frequency Analysis Methods and Applications* (Prentice-Hall, Upper Saddle River, NJ, 1996).
11. S. Rossignol, X. Rodet and J. Soumagne *et al.*, Feature extraction and temporal segmentation of acoustic signals, in *Proc. ICMC 98*, 1998, Ann Arbor, Michigan, pp. 199–202.
12. J. Saunders, Real-time discrimination of broadcast speech/music, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'96*, May 1996, Vol. 2, Atlanta, GA, pp. 993–996.
13. E. Scherier and M. Slaney, Construction and evaluation of a robust multifeature speech/music discriminator, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'97*, April 1997, Munich, Germany, pp. 1331–1334.
14. E. Wold, T. Blum, D. Keislar and J. Wheaton, Content-based classification, search, and retrieval of audio, *IEEE Multimed. Mag.* **3**(3) (1996) 27–36.

15. L. Wyse and S. Smoliar, Toward content-based audio indexing and retrieval and a new speaker discrimination technique, Institute of Systems Sciences, Nat. Univ. Singapore, http://www.iss.nus.sg/People/lwyse/lwyse.html, December 1995.
16. T. Zhang and C.-C. J. Kuo, Hierarchical classification of audio data for archiving and retrieving, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'99*, Vol. 6, 1999, pp. 3001–3004.
17. T. Zhang and C.-C. J. Kuo, Audio content analysis for online audiovisual data segmentation and classification, *IEEE Trans. Speech Audio Process.* **9**(4) (2001) 441–457.

Ruei-Shiang Lin received the B.S. and M.S. degrees in electrical engineering from Tamkang University, Taiwan, in 1996 and Tatung University, Taiwan, in 1998, respectively. He is currently a Ph. D. student in the Department of Computer and Information Science at National Chiao Tung University, Taiwan.

His research interests include image processing, pattern recognition and audio analysis.

Ling-Hwei Chen received the B.S. degree in mathematics and the M.S. degree in applied mathematics from the National Tsing Hua University, Hsinchu, Taiwan in 1975 and 1977, respectively, and the Ph.D. in computer engineering from National Chiao Tung University, Hsinchu, Taiwan in 1987.

From August 1977 to April 1979, she worked as a research assistant in the Chung-Shan Institute of Science and Technology, Taoyan, Taiwan, after which she worked as a research associate in the Electronic Research and Service Organization, Industry Technology Research Institute, Hsinchu, Taiwan. From March 1981 to August 1983, she worked as an engineer in the Institute of Information Industry, Taipei, Taiwan and is now a Professor at the Department of Computer and Information Science at the National Chiao Tung University.

Her current research interests include image processing, pattern recognition, document processing, image compression and image cryptography.