

國立交通大學

資訊科學系

博士論文

一個關於一般音訊資料之音訊分類，音訊分段及音訊檢索之研究

A Study on Classification, Segmentation and
Retrieval for Generic Audio Data

研究生：林瑞祥

指導教授：陳玲慧 博士

中華民國九十三年九月

一個關於一般音訊資料之音訊分類，音訊分段及音訊
檢索之研究

A STUDY ON CLASSIFICATION, SEGMENTATION
AND RETRIEVAL FOR GENERIC AUDIO DATA

研究生:林瑞祥

Student: Ruei-Shiang Lin

指導教授:陳玲慧博士

Advisor: Dr. Ling-Hwei Chen

國立交通大學電機資訊學院



A Dissertation Submitted to
Department of Computer and Information Science
College of Electrical Engineering and Computer Science
National Chiao Tung University
In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy
in
Computer and Information Science

September 2004
Hsinchu, Taiwan, R. O. C.

中華民國九十三年九月

一個關於一般音訊資料之音訊分類，音訊分段及音訊檢索之研究

研究生:林瑞祥

指導教授:陳玲慧博士

國立交通大學電機資訊學院 資訊科學系

摘要

近年來由於多媒體資料之大量增長，使得有效管理多媒體資料庫之議題變得十分重要而富挑戰性。因此多媒體資料庫之檢索及儲存便成為一個重要之研究領域。由於音訊資料在多媒體資料當中隨處可見，也扮演著一個重要的特徵，因此音訊資料相關的研究與分析便顯得重要；尤其是基於音訊內涵為主的相關分析更為顯的重要與迫切。

基於音訊內涵為主的相關研究其目前的發展狀況仍是十分有限，目前主要的問題與發展方向主要可歸納為三個方向：音訊分類、音訊分段以及音訊檢索。本論文之主要目的在基於 spectrogram 並運用圖樣識別等相關的理論來發展一些解決上述問題的方法。

一般而言，對於音訊資料的內容分析而言，音訊分類是最為重要的處理步驟；而目前音訊分類的研究其主要的問題乃是音訊的分類種類不足。大多數的分類法都是只將音訊分成語音和音樂兩大類；這樣的分類法比較簡單且容易，然而這樣的分類法並不足以應付目前的多

媒體資料。為了解決這個問題，我們將提出一個新的音訊分類法；除了語音和音樂這兩大類，我們所提出的分類法尚考慮了目前多媒體資料中常見的語音與背景音樂混合、流行歌曲等複合型態音訊資料。這個方法主要的重點在於，利用所提出的新音訊特徵與階層式分類法來達到音訊分類的目的。其系統之設計除了具備以音訊內涵為特徵來處理之功能及特色之外，其處理效率更是一個核心重點。

接著我們會提出一個基於音訊分類的音訊分段法。此方法的主要觀念是基於一個事實，即不同種類的音訊資料其 spectrogram 上蘊含了視覺上可見的特徵；例如音樂性的資料其能量在 spectrogram 上會集中分佈在某些方向，而語音類的資料，其能量的分佈會集中在某些頻帶區間，而隨機性的音訊資料例如雜訊，其能量的分佈則出現在所有方向。基於上述事實，我們利用 Gabor Wavelet 先針對以一秒為單位之音訊資料的 spectrogram 上能量在方向性分佈以及比例進行強化，接著利用強化後的 spectrogram 上能量在方向性分佈以及比例的分析來進一步將音訊資料分類。接著，基於分類後的結果來應用於音訊片段的音訊分段切割處理。

最後，我們將提出一個基於音訊內涵的音訊資料檢索方法。此方法將針對使用者所提供的音訊查詢片段進行音訊檢索，其檢索能力範圍包括資料庫中相似的音訊片段，樂曲中重複的音訊片段及旋律相同

但表達方式不同的樂曲，例如不同語言或者不同人等。此方法的主要觀念也是運用音訊資料其 spectrogram 上所蘊含的視覺上可見的有效特徵，並利用 Gabor Wavelets 針對音訊資料的 spectrogram 上能量在方向性分佈以及比例進行強化，並利用強化後的 spectrogram 其傅立葉頻譜的反應值來找出最有效率的 spectrogram。最後利用特徵選擇以及圖樣識別理論找出所需要的特徵以提供音訊檢索之用。

本論文中所提出之方法可應用於多媒體資料檢索，音訊瀏覽及數位圖書館系統之設計。



A STUDY ON CLASSIFICATION, SEGMENTATION AND RETRIEVAL FOR GENERIC AUDIO DATA

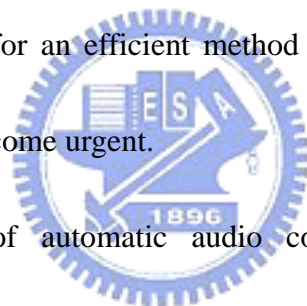
Student: Ruei-Shiang Lin

Advisor: Dr. Ling-Hwei Chen

Institute of Computer and Information Science
National Chiao Tung University

ABSTRACT

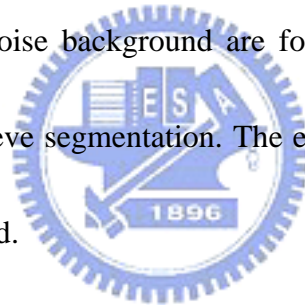
The recent emerging of multimedia and the tremendous growth of multimedia data archives have made the effective management of multimedia databases become a very important and challenging task. Digital audio is an important and integral part of many multimedia applications such as the construction of digital libraries. Thus, the demand for an efficient method to automatically analyze audio signal based on its content become urgent.



The major problems of automatic audio content analysis include audio classification, segmentation and retrieval etc. In this dissertation, based on spectrogram, we will propose three methods to address the problems of audio classification, segmentation and content-based retrieval. 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. These categories are the basic sets needed in the content analysis of audiovisual data. First, a hierarchical audio classification method will be presented to classify audio signals into the aforementioned basic audio types.

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. The experimental results of the proposed method are quite encouraging.

Next, based on the Gabor wavelet features, we will propose a non-hierarchical audio classification and segmentation method. The proposed method will first divide an audio stream into clips, each of which contains one-second audio information. Then, each clip is classified as one of two classes or five classes. Two classes contain speech and music; pure speech, pure music, song, speech with music background, and speech with environmental noise background are for five classes. Finally, a merge technique is provided to achieve segmentation. The experimental results demonstrate the effectiveness of the method.



Finally, we will propose a method for content-based retrieval of perceptually similar music pieces in audio documents. It allows the user to select a reference passage within an audio file and retrieve perceptually similar passages such as repeating phrases within a music piece, similar music clips in a database or one song sung by different persons or in different languages. The experimental results demonstrate the effectiveness of the method. The methods proposed in this dissertation can be used as the basic components when developing an audio content analysis system or a system used in a digital library application.

TABLE OF CONTENTS

ABSTRACT IN CHINESE.....	I
ABSTRACT IN ENGLISH	IV
TABLE OF CONTENTS	VI
LIST OF TABLES	VIII
LIST OF FIGURES	IX
ABBREVIATION	XI
ACKNOWLEDGEMENTS	XII
CHAPTER 1 INTRODUCTION	1
1.1 Motivation and Applications.....	1
1.2 State of the Problems and Research Scope	2
1.2.1 Some Problems of Audio Classification and Segmentation	3
1.2.2 Some Problems of Audio Retrieval.....	8
1.3 Audio Representation.....	9
1.3.1 Short Time Fourier Transform	10
1.3.2 Multi-resolution Short Time Fourier Transform.....	11
1.4 Synopsis of the Dissertation.....	13
CHAPTER 2 A NEW APPROACH FOR CLASSIFICATION OF GENERIC AUDIO DATA	15
2.1. Introduction.....	15
2.2 The Proposed Method	18
2.2.1 Feature Extraction Phased.....	21
2.2.1.1 The Energy Distribution Model	22
2.2.1.2 The Horizontal Profile Analysis.....	25
2.2.1.3 The Temporal Intervals	27
2.2.2 Audio Classification.....	33
2.2.2.1 The Coarse-Level Classification.....	33
2.2.2.2 The Fine-Level Classification.....	34
2.3. Experimental Results	35
2.3.1 Classification Results.....	36
2.4. Summary	38
CHAPTER 3 A NEW APPROACH FOR AUDIO CLASSIFICATION AND SEGMENTATION USING GABOR WAVELETS AND FISHER LINEAR DISCRIMINATOR	39

3.1 Introduction.....	39
3.2 The Proposed Method	42
3.2.1 Initial Feature Extraction	44
3.2.1.1 Gabor Wavelet Functions and Filters Design.....	45
3.2.1.2 Feature Estimation and Representation	47
3.2.2 Feature Selection and Audio Classification	48
3.2.3 Segmentation.....	52
3.3 Experimental Results	54
3.3.1 Classification Results.....	55
3.3.2 Segmentation Results.....	57
3.4. Summary	59
CHAPTER 4 CONTENT-BASED AUDIO RETRIEVAL BASED ON GABOR WAVELETS	61
4.1 Introduction.....	61
4.2 The Proposed Method	65
4.2.1 Initial Feature Extraction	66
4.2.1.1 Feature Estimation	68
4.2.1.2 Feature Selection and Representation	71
4.2.2 Audio Retrieval and Similarity Measurement.....	72
4.2.2.1 Similarity Measure	73
4.2.2.2 Retrieval	74
4.3 Experimental Results	76
4.3.1 Retrieval Results	76
4.4. Summary	80
CHAPTER 5 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS	81
5.1 Conclusions.....	81
5.2 Future Research Directions.....	83
REFERENCES	84
PUBLICATION LIST	88

LIST OF TABLES

TABLE 2.1	COARSE-LEVEL CLASSIFICATION RESULTS	37
TABLE 2.2	FINAL CLASSIFICATION RESULTS	37
TABLE 3.1	TWO-WAY CLASSIFICATION RESULTS.	56
TABLE 3.2	FIVE-WAY CLASSIFICATION RESULTS.	56
TABLE 3.3	SEGMENTATION RESULTS.	59
TABLE 4.1	THE AVERAGE RECALL RATES OF THE FIRST EXPERIMENT	78
TABLE 4.2	THE AVERAGE RECALL RATES OF THE SECOND EXPERIMENT.	78



LIST OF FIGURES

Fig. 1.1	The hierarchical classification scheme.....	5
Fig. 1.2	Two examples to show some possible different kinds of patterns in a spectrogram. (a) Line tracks corresponding to tones in a music spectrogram. (b) Clicks, noise burst and frequency sweeps in a song spectrogram.....	6
Fig. 1.3	Temporal segmentation.....	7
Fig. 1.4	Short Time Fourier Transform.....	10
Fig. 1.5	An example of tiling in the time-frequency plane.....	12
Fig. 1.6	A schematic diagram of the TFD generating details.....	12
Fig. 2.1	Block diagram of the proposed system.....	19
Fig. 2.2	Five spectrogram examples. (a) Music. (b) Speech with music background. (c) Song. (d) Speech. (e) Speech with environmental noise background.....	20
Fig. 2.3	Two examples of the energy distribution models. (a) Unimodel (the histogram of the energy distribution of Fig. 2.2 (a)). (b) Bimodel (the histogram of the energy distribution of Fig. 2.2 (c))...	23
Fig. 2.4	Five examples of the horizontal profiles. (a) – (e) are the horizontal profiles of Figs. 2(a) - 2(e), respectively.....	26
Fig. 2.5	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).....	29
Fig. 2.6	An example of the re-merged process. (a) Initial temporal intervals. (b) Result after re-merged process.....	32
Fig. 3.1	Block diagram of the proposed method, where “MB” and “NB” are the abbreviations for “music background” and “noise background”, respectively.....	43
Fig. 3.2	Two examples to show some possible different kinds of patterns in a spectrogram. (a) Line tracks corresponding to tones in a music spectrogram. (b) Clicks, noise burst and frequency sweeps in a song spectrogram.....	44

Fig. 3.3	An example of using FLD for two-way speech/music discriminator.....	51
Fig. 3.4	A block diagram of feature selection and classification using FLD.	52
Fig. 3.5	Demonstration of audio segmentation and indexing, where “SMB” and “SNB” are the abbreviations for “speech with music background” and “speech with noise background”, respectively. (a) Original result. (b) Final results after applying the segmenting algorithm to (a).....	58
Fig. 4.1	Block diagram of the proposed method.....	65
Fig. 4.2	Two examples to show some possible different kinds of patterns in a spectrogram. (a) Line tracks corresponding to tones in a musical instrument spectrogram. (b) Clicks, noise burst, tones, and frequency sweeps in a song spectrogram.....	67
Fig. 4.3	An example to show the enhancement process performing in a spectrogram. (a) The Gabor-wavelet filtered spectrogram with the maximum contrast. (b) Enhanced spectrogram.....	70



ABBREVIATION

ASA	Auditory Scene Analysis
QBH	Query by Humming
QBE	Query-by-Example
TFD	Time-Frequency Distribution
STFT	Short Time Fourier Transform
MFCC	Mel-Frequency Cepstral Coefficients
MSTFT	Multi-resolution Short Time Fourier Transform
SVD	Singular Value Decomposition
HAS	Human Auditory System
FLD	Fisher Linear Discriminator



ACKNOWLEDGEMENTS

I would like to extend my sincere thanks and appreciation to Professor Ling-Hwei Chen, my dissertation advisor, for her kind patience, teaching, guidance and advice throughout this research project. I have benefited tremendously from her experience and knowledge. Without her persistent support and encouragement, this dissertation would not have come into existence. I also thank Professor Yu-Tai Ching and Professor Sheng-Jyh Wang, the member of my proposal committee, for their valuable comments and constructive suggestions.

I also consider myself very fortunate in having many faculty members in the AIP group for broadening my academic background and making my life a lot more fun and interesting for the past six years. Finally, I will dedicate this dissertation to my parents and my girlfriend for their endless love and constant concern during these years.

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION AND APPLICATIONS

The recent emerging of multimedia and the tremendous growth of multimedia data archives have made the effective management of multimedia databases become a very important and challenging task. Therefore, developing efficient analysis techniques for multimedia data based on its content become very important and have drawn lots of attentions recently.



Audio is an important and integral part of many multimedia applications such as professional media production, audio archive management, commercial music usage, content-based audio/video retrieval, and so on. Thus, developing some techniques to treat audio signal can help analyze multimedia data. For example, audio classification and segmentation techniques can be used to support video scene-change detection and video classification. In general, neighboring scenes in video 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. Furthermore, the availability of large multimedia data archives has made content-based information retrieval become a very

popular research topic. Audio and especially music collections are also deemed as one of the most important features when performing content-based information retrieval.

In recent years, due to the importance of audio analysis, techniques for audio content analysis have started emerging as research prototypes [1-24] and devoting to solve the audio related problems called Auditory Scene Analysis (ASA), such as audio segmentation, audio classification, audio indexing and content-based audio retrieval etc. Those are the fundamental processes of any type of auditory analysis. In this dissertation, we will propose methods to deal with above-mentioned audio related problems and will be described in the following.



1.2. STATE OF THE PROBLEMS AND RESEARCH SCOPE

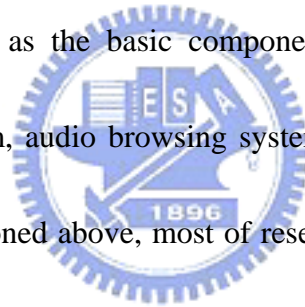
In this dissertation, we will propose three methods to deal with the problems of audio classification, audio segmentation and content-based audio retrieval. These three problems are defined as follows:

- (1) ***Audio classification***: given an audio clip, to develop a method to classify it into one of the common categories: music, speech, song, and etc. It is the most important process for auditory analysis since different audio types require different processing and have different significance to different applications.

(2) **Audio segmentation**: given an audio stream, to develop a method to automatically detect when there are abrupt changes (segmentation boundaries) in the stream. For example the change from music to speech is the example of the segmentation boundary.

(3) **Audio content-based retrieval**: given an audio clip as query sample, to develop a method to retrieve its perceptually similar clips in audio documents.

Actually, the three problems are logically sequenced. And the solutions to these three problems can be used as the basic components when developing an audio content-based analysis system, audio browsing system or a system used in a digital library application. As mentioned above, most of research efforts have been spent on these three problems. However, some points still remain to be solved and will be described in the following.



1.2.1 Some Problems of Audio Classification and Segmentation

One problem of audio classification is the audio categories. Traditional approaches for audio classification tend to roughly divide audio signals into two major distinct categories: speech and music (two-way classification) [3-5]. 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-way classification for audio data is not enough in many applications, such as content-based video retrieval [12, 14]. 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. This indicates the need to take other kinds of audio into consideration.

Some problems are in those existing classification methods for more than two audio categories. For example, Zhang and Kuo [14] provided a classifier, which extracts 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. 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. In the dissertation, we will propose two audio classification methods to address the above-mentioned shortcomings.

The first proposed method is a hierarchical audio classifier, which will 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. From the hierarchical viewpoint, these five categories are first roughly divided into two major distinct categories: single-type and hybrid-type, i.e., with or without background components. Then, the single-type sounds are further classified into speech and music; the hybrid-type sounds are classified into speech with environmental noise background, speech with music background and song. Fig. 1.1 shows this hierarchical classification scheme. In the first proposed method, we will use lesser features with high differentiating power to achieve the classification purpose. However, the first proposed method is not suitable for classification-based audio segmentation since the features are extracted from the audio clips with larger length. To address this shortcoming, in the dissertation, we will propose the other audio classification methods to support classification-based audio segmentation.

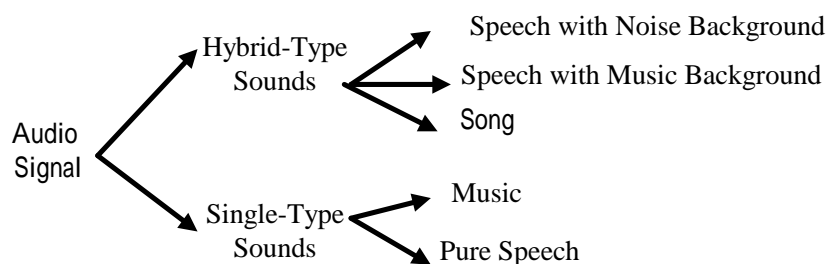


Fig. 1.1. The hierarchical classification scheme.

The second proposed method is a non-hierarchical audio classifier, which will first divide an audio stream into clips, each of which contains one-second audio information. Based on the classified clips with smaller length, the proposed method is suitable and can be used to support classification-based audio segmentation.

Generally speaking, the spectrogram is a good representation for an audio signal since it is often visually interpretable. By observing a spectrogram, we can find that the energy is not uniformly distributed, but tends to cluster to some patterns. All curve-like patterns are called tracks. Fig. 1.2(a) shows that for a music signal, some line tracks corresponding to tones will exist on its spectrogram. Fig. 1.2(b) shows some patterns including clicks (broadband, short time), noise burst (energy spread over both time and frequency), and frequency sweeps in a song spectrogram. Thus, if

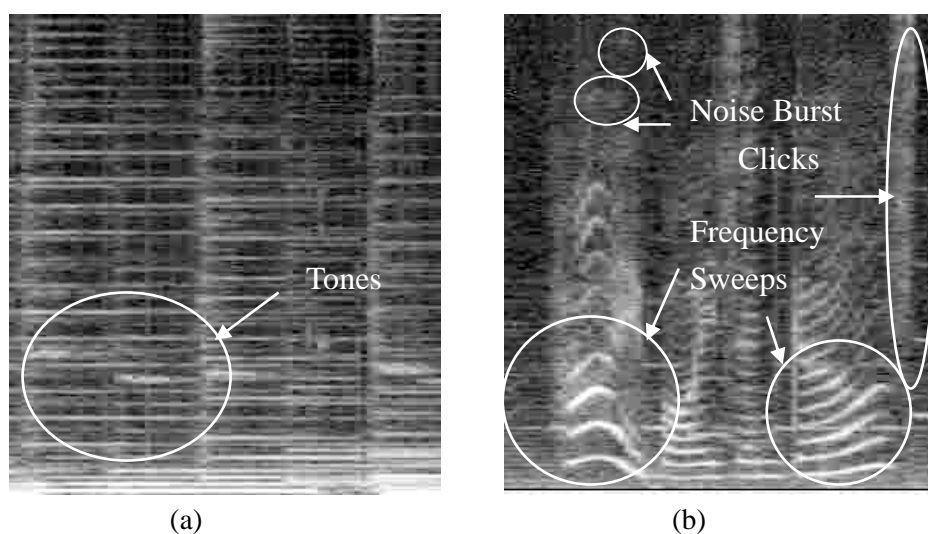


Fig. 1.2. Two examples to show some possible different kinds of patterns in a spectrogram. (a) Line tracks corresponding to tones in a music spectrogram. (b) Clicks, noise burst and frequency sweeps in a song spectrogram.

we can extract some features from a spectrogram to represent these patterns, the classification should be easy. Based on these phenomena, the proposed method will adopt feature selection process to explore the features with the highest discriminative ability to achieve classification purposes and will be used to do audio segmentation.

As for the audio segmentation, most of the existing approaches for audio segmentation can be classified into two major paradigms: temporal segmentation and classification-based segmentation. Temporal segmentation (see Fig. 1.3) is a more primitive process than classification-based segmentation since it does not try to interpret the data. By contrast, the classification-based segmentation divides an audio sequence into semantic scenes called “audio scene” and to index them as different audio classes. That is, the approaches via classification usually adopt classification results to achieve segmentation purpose and the performance is dependent on the classification result. In this dissertation, based on the proposed above-mentioned classification method, we will present one classification-based segmentation method to achieve segmentation purpose.

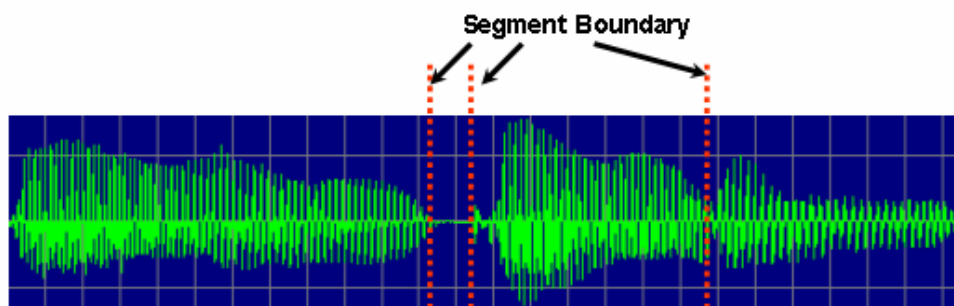


Fig. 1.3. Temporal segmentation.

1.2.2 Some Problems of Audio Retrieval

In recent years, techniques for audio information retrieval have started emerging as research prototypes. These systems can be classified into two major paradigms [22, 34]. In the first paradigm, the user sings a melody and similar audio files containing that melody are retrieved. This kind of approaches [18] is called “Query by Humming” (QBH). It has the disadvantage of being applicable only when the audio data is stored in symbolic form such as MIDI files. The conversion of generic audio signals to symbolic form, called polyphonic transcription, is still an open research problem in its infancy [16]. Another problem with QBH is that it is not applicable to several musical genres such as Dance music where there is no singable melody that can be used as a query. The second paradigm [9, 15-17, 19-24] is called “Query-by-Example” (QBE), a reference audio file is used as the query and audio files with similar content are returned and ranked by their similarity degree. In order to search and retrieve general audio signals such as the raw audio files (e.g. mp3, wave, etc.) on the web or databases, only the QBE paradigm is currently applicable.

There are some disadvantages in the existing QBE audio-retrieval methods. For example, the method proposed by Wold, et al. [9] is only suitable for sounds with a single timbre. It is supervised and not adequate to index general audio content. An approach provided in [15] has accuracy varying considerably for different types of

recording, and the audio segment to be searched should be known *a priori* in this algorithm. Two MFCC-based (Mel-frequency cepstral coefficients) approaches [16, 21] are not suitable for melody retrieval (e.g. music) since the MFCC-based features do not capture enough information about the pitch content, they characterize the broad shape of the spectrum. Besides, most of the current works only deal with the monophonic sources. Polyphonic music is more common, but it is also more difficult to represent. To solve the above-mentioned shortcomings, in the dissertation, we will present one method for content-based audio retrieval and will also consider polyphonic music.

In the dissertation, we will develop our methods based on spectrogram. In the following, we will give a brief review of the generation of the spectrogram.



1.3 AUDIO REPRESENTATION

We will develop our methods based on spectrogram that is a commonly used representation of an acoustic signal in a three-dimensional (time, frequency, intensity) space known as a time-frequency distribution (TFD) [29]. 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). To construct a spectrogram, the Short Time Fourier Transform (STFT) is applied. In the following, we will give a brief review of the theories of the Short Time Fourier Transform.

1.3.1 Short Time Fourier Transform

In general, 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 (see Fig.1.4). Let $s(t)$ denote the audio signal and $STFT(\tau, \omega)$ be the result of STFT, that is

$$STFT(\tau, \omega) = \sum_{t=0}^{n-1} s(t + \frac{\tau}{2}n) r^*(t) e^{-j\omega t}, \quad (1.1)$$

where $r^*(t)$ is the window function, τ stands for the frame number, n is the window size and ω is the frequency parameter. Then, the spectrogram, $S(\tau, \omega)$, is

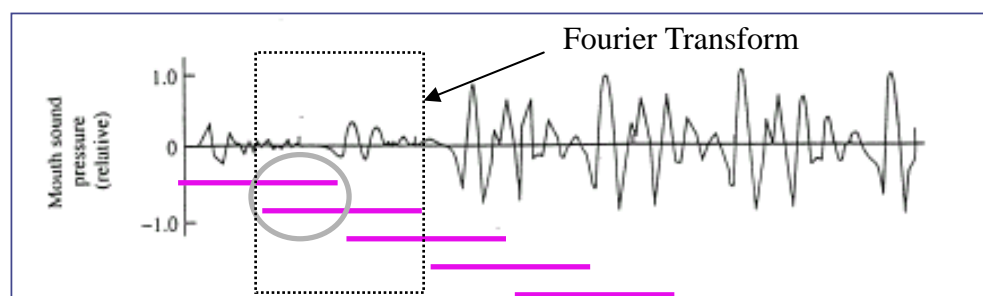


Fig. 1.4. Short Time Fourier Transform.

the energy distribution associated with the Short Time Fourier Transform, that is,

$$S(\tau, \omega) = 10 \log_{10} \left(\frac{|STFT(\tau, \omega)|^2}{M} \right), \quad (1.2)$$

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

1.3.2 Multi-resolution Short Time Fourier Transform

Conventionally, in the Short Time Fourier Transform, the TFD is sampled uniformly in time and frequency. However, it is not suitable for the auditory model because the frequency resolution within the human psycho-acoustic system is not constant but varies with frequency [29]. By contrast, in the Multi-resolution Short Time Fourier Transform (MSTFT), the TFD is perceptually tuned, mimicking the time-frequency resolution of the ear. That is, the TFD consists of axes that are non-uniformly sampled. Frequency resolution is coarse and temporal resolution is fine at high frequencies while temporal resolution is coarse and frequency resolution is fine at low frequencies [19].

One example of the tiling in the time-frequency plane is shown in Fig. 1.5 and Fig. 1.6 shows a schematic diagram of the TFD generating using the Multi-resolution Short Time Fourier Transform.

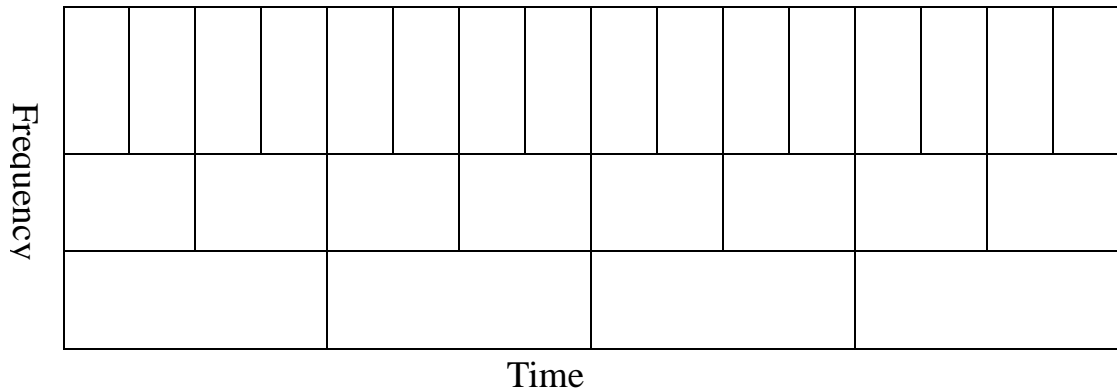


Fig. 1.5. An example of tiling in the time-frequency plane.

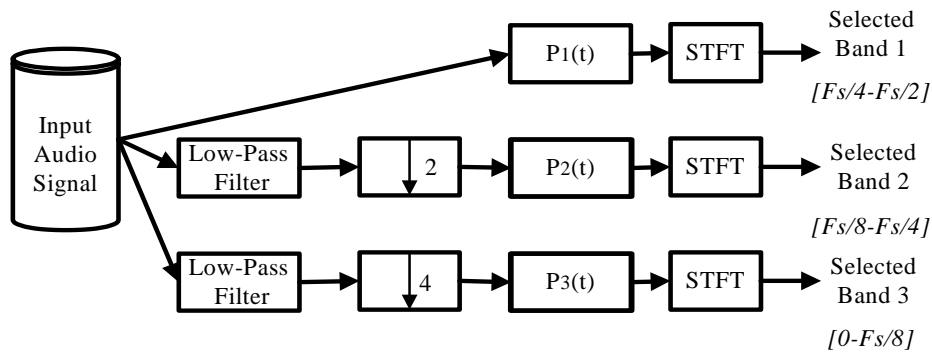


Fig. 1.6. A schematic diagram of the TFD generating details.

There are three parts in the TFD generating. In the first part, the N -point STFT is applied to the original audio signal $P_1(t)$ to obtain a spectrogram $S_1(x, y)$. In the second part, a low-pass filter is first applied to $P_1(t)$ and then the filtered result is downsampled half size to obtain signal $P_2(t)$ and the N -point STFT is applied to $P_2(t)$ to obtain a spectrogram $S_2(x, y)$. In the third part, a low-pass filter is first applied to $P_1(t)$ and then the filtered result is downsampled quarter size to obtain signal $P_3(t)$ and the N -point STFT is applied to $P_3(t)$ to obtain a spectrogram

$S_3(x, y)$. The frequency resolution Δf_j and the analysis time interval T_j in

$S_j(x, y)$ can be calculated as follows:

$$\Delta f_j = \frac{1}{2^{j-1}} \cdot \frac{F_s}{N} = \frac{1}{T_j}, \quad j = 1, 2, 3. \quad (1.3)$$

Note that the window center at the k th time block in $S_j(x, y)$, t_j^k , is given by

$$t_j^k = \frac{k}{2} T_j, \quad j = 1, 2, 3. \quad (1.4)$$

Finally, based on $S_1(x, y)$, $S_2(x, y)$, and $S_3(x, y)$, a spectrogram $I(x, y)$ is

obtained according to the following equation:

$$I(x, y) = \begin{cases} S_1(x, y), & \text{if } y \in [F_S/4, F_S/2], x = 0, 1, \dots, N_f - 1; \\ S_2(2i, y), & \text{if } y \in [F_S/8, F_S/4], x = 2i, 2i+1, \dots, 2i+2, \quad i = 0, 1, \dots, N_f/2 - 1; \\ S_3(4i, y), & \text{if } y \in [0, F_S/8], x = 4i, \dots, 4i+3, \quad i = 0, 1, \dots, N_f/4 - 1; \end{cases} \quad (1.5)$$

where N_f is the frame number of $P_1(t)$. From Eq. (1.3), we can see that in $I(x, y)$,

frequency resolution is coarse and temporal resolution is fine at high frequencies

while temporal resolution is coarse and frequency resolution is fine at low frequencies.

This means that $I(x, y)$ meets the human psycho-acoustic system.

1.4 SYNOPSIS OF THE DISSERTATION

The rest of the dissertation is organized as follows. Chapter 2 describes the proposed hierarchical audio classification method. The non-hierarchical audio

classification and segmentation method based on Gabor wavelets is proposed in Chapter 3. The proposed method of audio retrieval based on Gabor wavelets is described in Chapter 4. Some conclusions and future research directions are drawn in Chapter 5.



CHAPTER 2

A NEW APPROACH FOR CLASSIFICATION OF GENERIC AUDIO DATA

2.1. INTRODUCTION

Audio classification [1-14] 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 proposed. These methods tend to roughly divide audio signals into two major distinct categories: speech and music. Scherier and Slaney [3] 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 [5] 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

[11]. Recently, video retrieval has become an important research topic. To raise the retrieval speed and precision, a video is usually segmented into several scenes [11,14]. 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 [7] presented a method to classify audio signals into “music,” “speech,” and “others.” The method was developed for the parsing of news stories. In [8], audio signals are classified into speech, silence, laughter, and non-speech sounds for the purpose of segmenting discussion recordings in meetings. The above-mentioned approaches are developed for specific scenarios, only some special audio types are considered. The research in [12-14] has taken more general types of audio data into account. In [12], 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

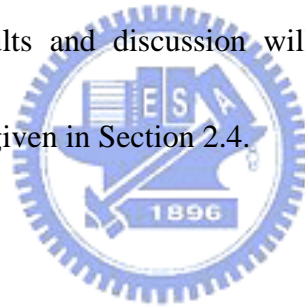
is reported. Zhang and Kuo [14] first extracted some audio features including the short-time fundamental frequency and the spectral tracks by detecting the peaks from the spectrum. The spectrum is generated by autoregressive model (AR model) coefficients, which are estimated from the autocorrelation 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 chapter, 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 [15]. 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. Experimental results show that the proposed method achieves an accuracy rate of more than 96% in audio classification.

The chapter is organized as follows. In Section 2.2, the proposed method will be described. Experimental results and discussion will be presented in Section 2.3. Finally, the summary will be given in Section 2.4.



2.2. THE PROPOSED METHOD

The system diagram of the proposed audio classification method is shown in Fig. 2.1. It is based on the spectrogram and consists of three phases: feature extraction, the coarse-level classification and the fine-level classification. First, an input audio clip is transformed to a spectrogram as mentioned in *Short Time Fourier Transform* section (Chapter 1, Section 1.3.1) and four effective audio features are extracted. Figs. 2.2(a) – 2.2(e) show five examples of the spectrograms of music, speech with music

background, song, pure speech, and speech with environmental noise background, respectively. 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.

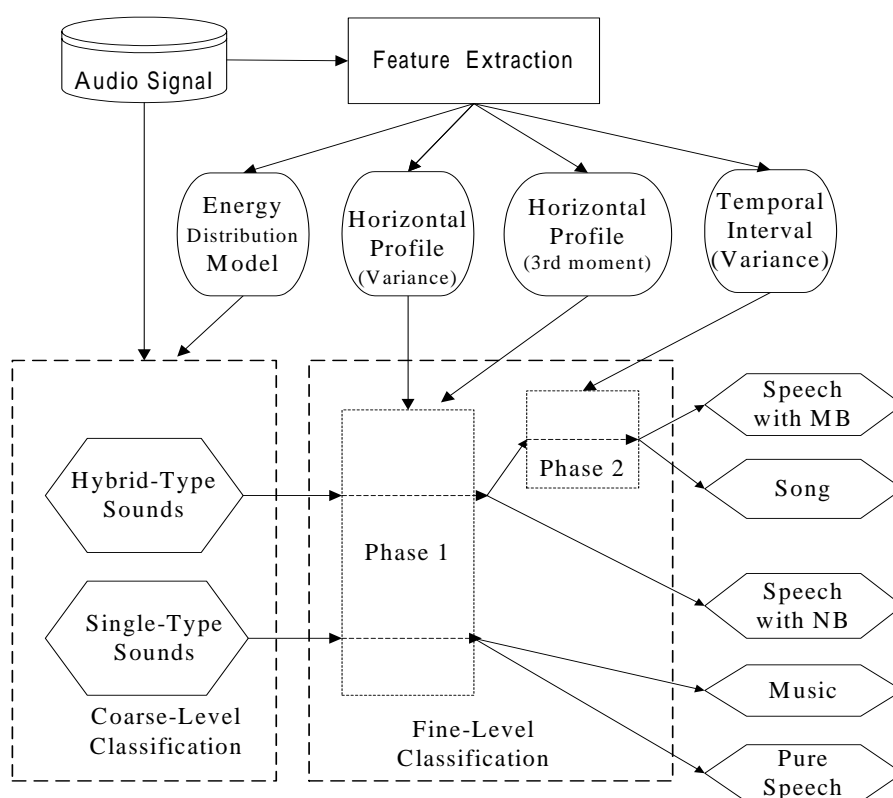
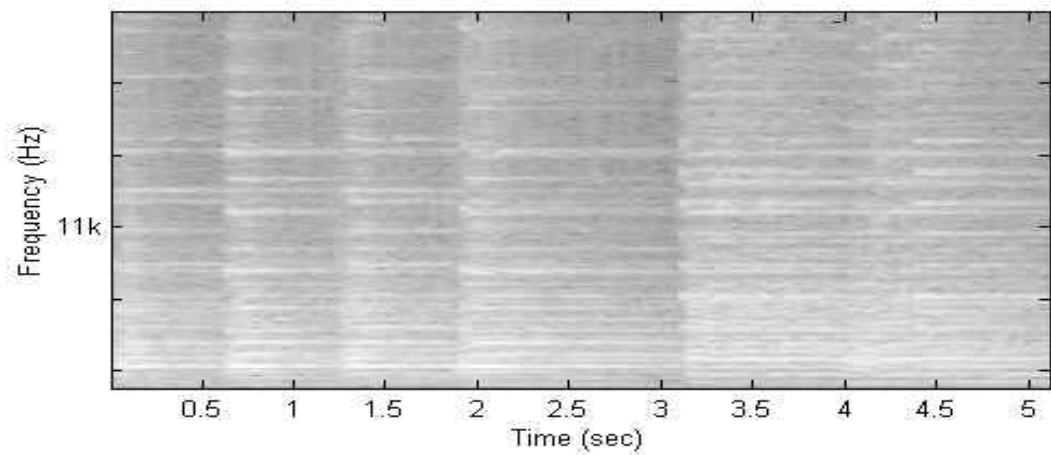
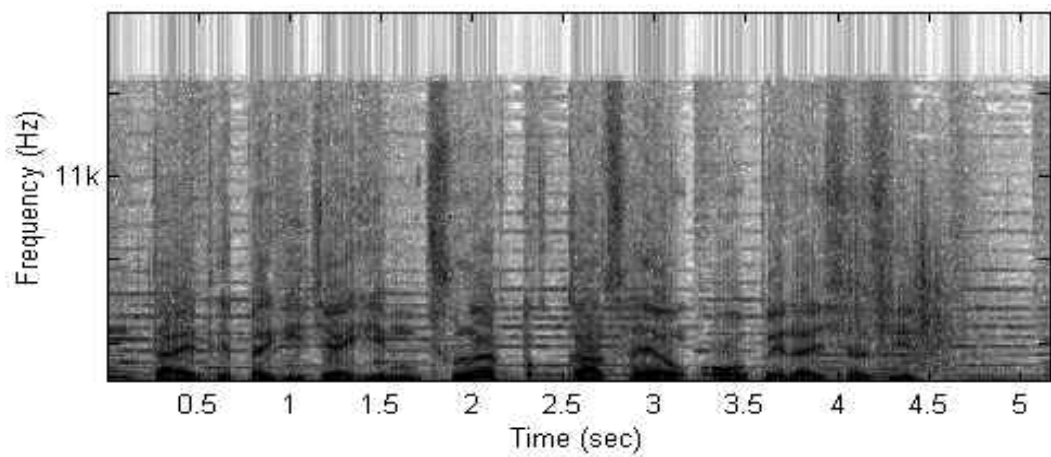


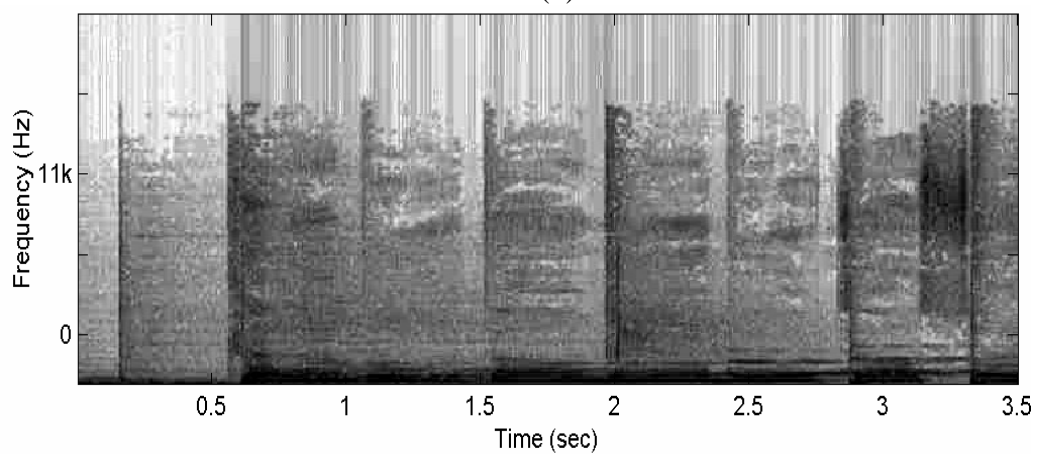
Fig. 2.1. Block diagram of the proposed system, where “MB” and “NB” are the abbreviations for “music background” and “noise background”, respectively.



(a)



(b)



(c)

Fig. 2.2. Five spectrogram examples. (a) Music. (b) Speech with music background. (c) Song. (d) Speech. (e) Speech with environmental noise background.

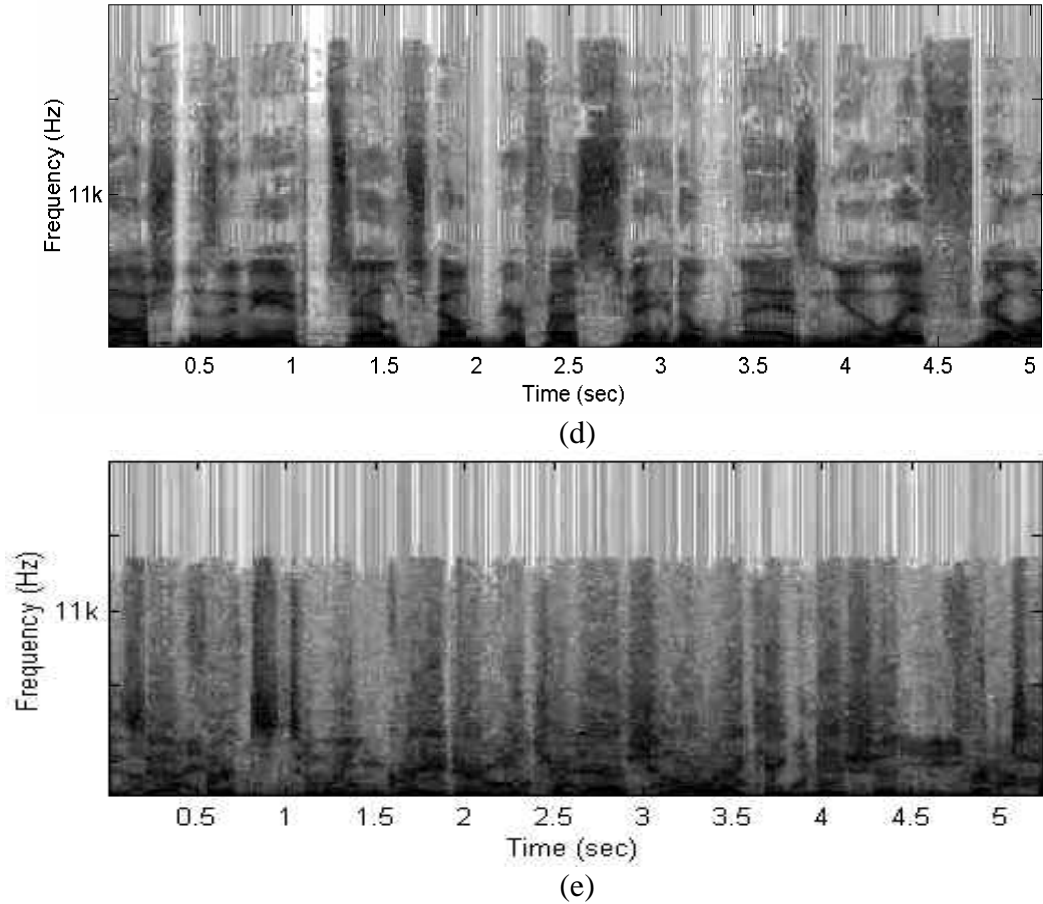


Fig. 2.2. Five spectrogram examples. (a) Music. (b) Speech with music background. (c) Song. (d) Speech. (e) Speech with environmental noise background. (Continued)

2.2.1. 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 audio signal is constructed first. Based on the spectrogram, these four features are extracted

and described as follows.

2.2.1.1 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. 2.3 (a) and 2.3 (b)), in audio signals. In Fig. 2.3, 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. 2.3 (a)). On the other hand, for a bimodel, dividing $h(i)$ into two parts from μ , each part will be unimodel (see Fig. 2.3 (b)). Thus, if we find a local 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.

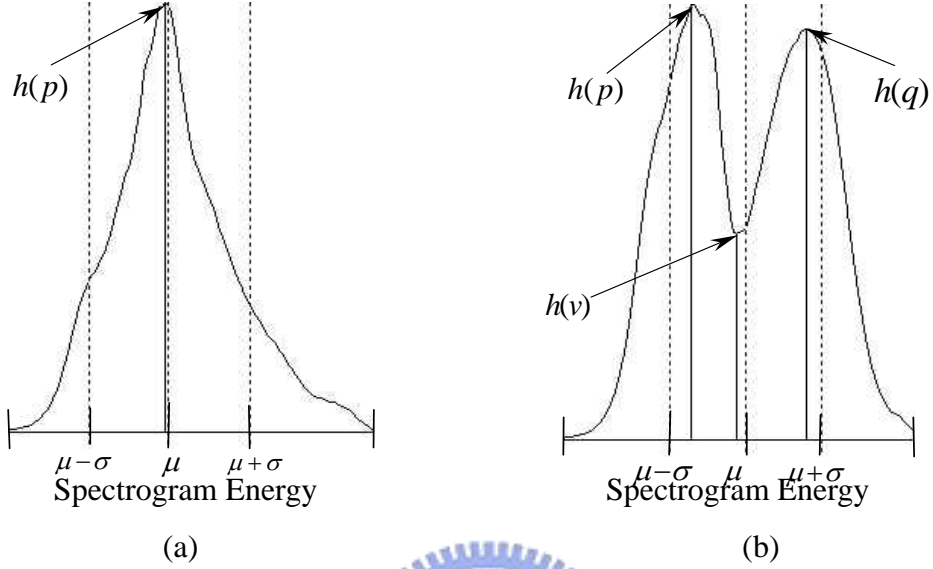


Fig. 2.3. Two examples of the energy distribution models. (a) Unimodel (the histogram of the energy distribution of Fig. 2.2 (a)). (b) Bimodel (the histogram of the energy distribution of Fig. 2.2 (c)).

Algorithm 2.1. 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, $h(i), i = 0, \dots, 255$, 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 $|p - \mu| \leq 5$, $T = \text{unimodel}$, go to *Step 9*.

Else

Use μ to set the search range \mathfrak{R}_p as follows:

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

End if.

Step 5. Find the position q of the highest peak $h(q)$ within \mathfrak{R}_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 $T = \text{bimodel}$ 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 $T = \text{unimodel}$.

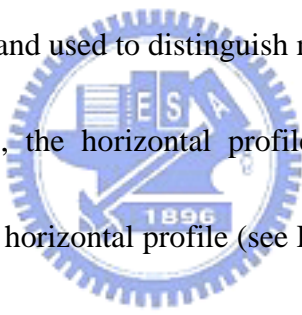
Step 9. Output T and assign μ to $T1$, $\mu + \sigma$ to $T2$.

End of **Algorithm 2.1**.

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.

2.2.1.2 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. 2.2 (a) – 2.2 (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. 2.2 (d) – 2.2 (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. 2.4 (a) – 2.4 (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. 2.4 (a) – 2.4 (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. 2.4 (d) – 2.4 (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 $F_s/4$ are discarded, where F_s 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.

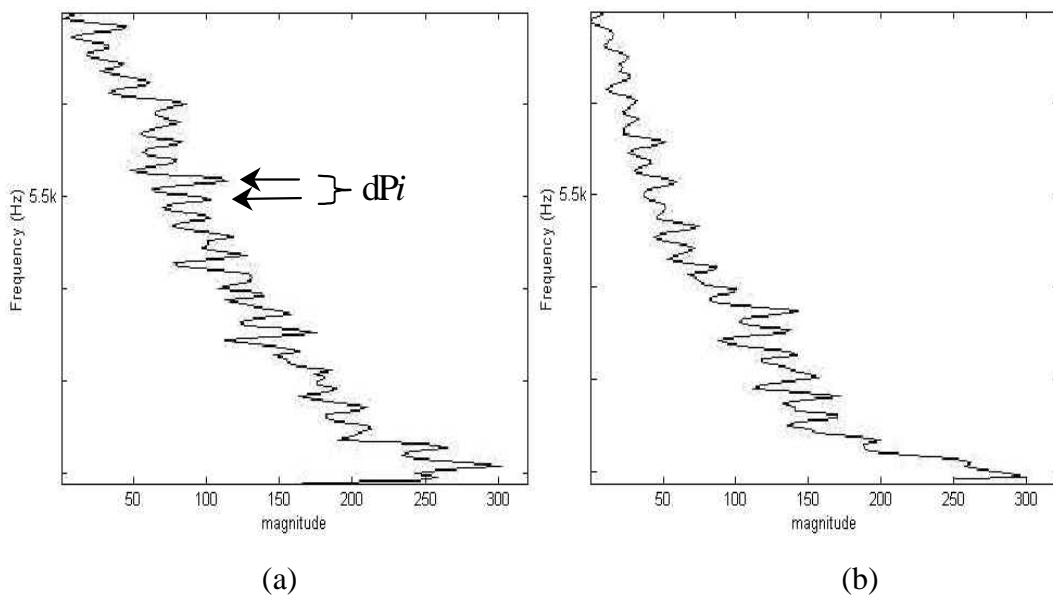
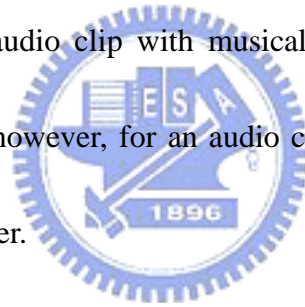


Fig. 2.4. Five examples of the horizontal profiles. (a) – (e) are the horizontal profiles of Figs. 2(a) - 2(e), respectively.

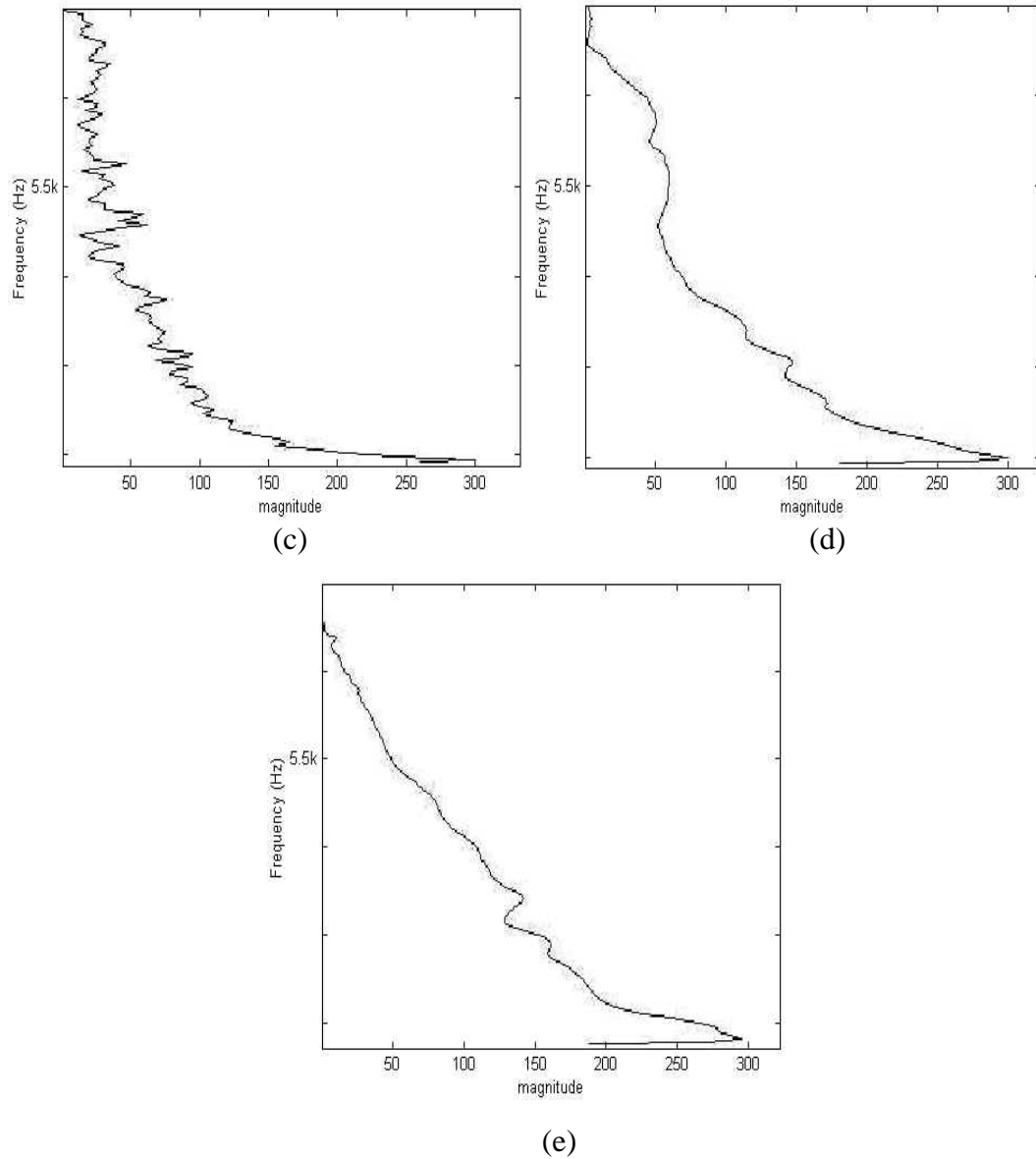


Fig. 2.4. Five examples of the horizontal profiles. (a) – (e) are the horizontal profiles of Figs. 2(a) - 2(e), respectively. (Continued)

2.2.1.3 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.

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 is 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. 2.5 (c)). Conversely, song is usually melodic and rhythmic, the difference between any two adjacent duration of music-voice in song is small and near constant (see Fig. 2.5 (a)).

By observing the spectrogram in different frequency bands, we can see that music-voice (i.e. speech and music appears 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. 2.5.

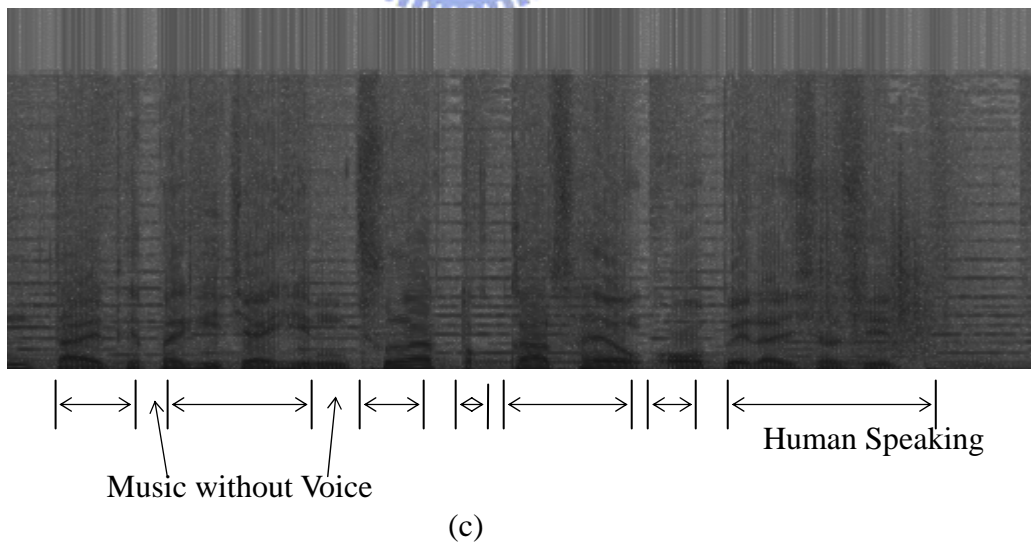
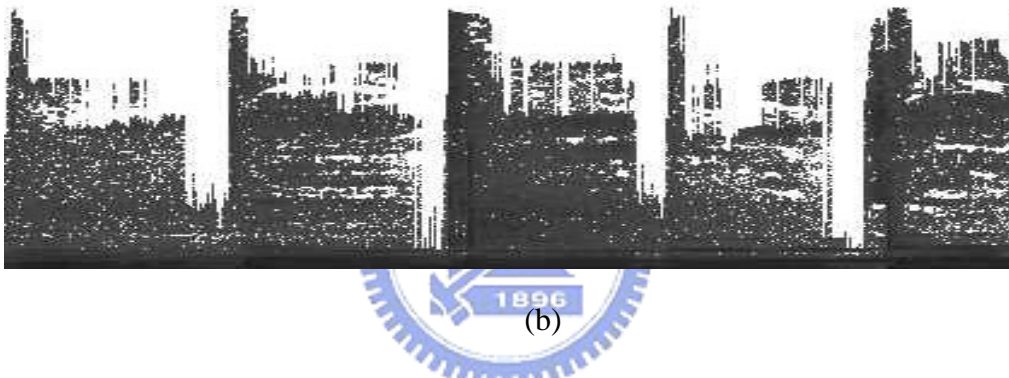
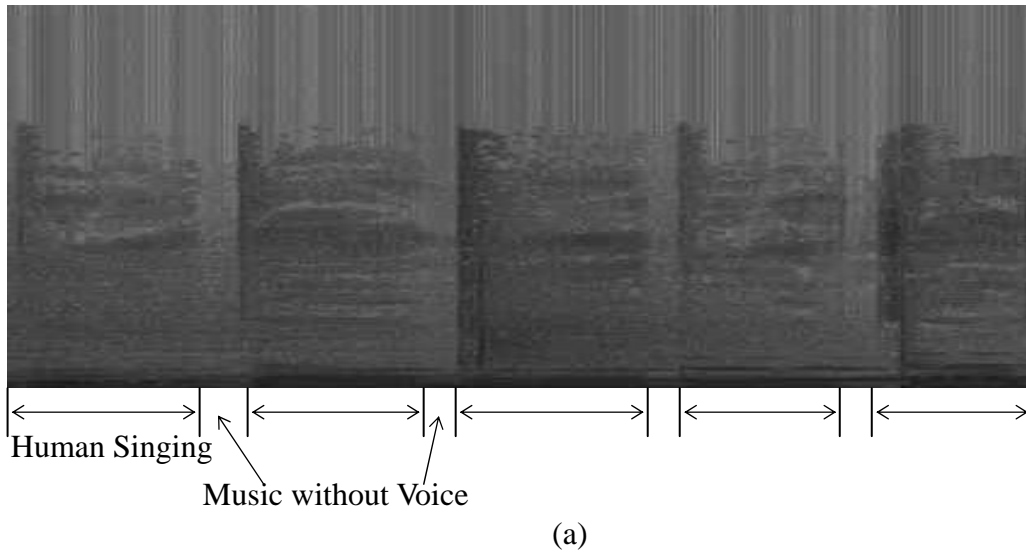
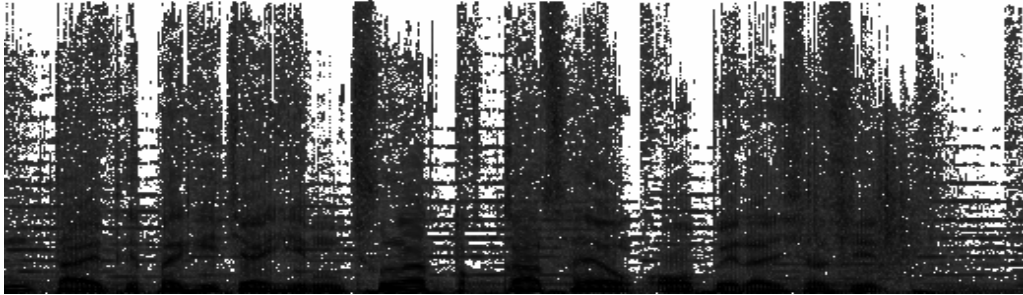


Fig. 2.5. Two examples of the filtered spectrogram. (a) The spectrogram of song. (b) The filtered spectrogram of (a). (c) The spectrogram of speech with music background. (d) The filtered spectrogram of (c).



(d)

Fig. 2.5. 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). (Continued)

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. The temporal 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:*

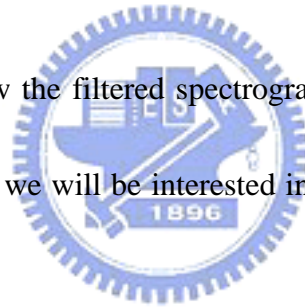
For the spectrogram of each frame τ , $S(\tau, \omega)$, find the highest

frequency ω_h with $S(\tau, \omega_h) > T2$. Set $\hat{S}(\tau, \omega) = 0, \forall \omega > \omega_h$.

2) *Filter other components:*

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

Figs. 2.5 (b) and 2.5 (d) show the filtered spectrograms of Figs. 2.5 (a) and 2.5 (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 of ones will separate two temporal intervals (see Fig. 2.5 (a)). To extract temporal intervals, the entire frequency band $[0, Fs/2]$ is first divided into two subbands of unequal width: $[0, Fs/8]$ and $[Fs/8, Fs/2]$. Next, for each frame, evaluate the ratio of the non-zero part in each subband to the total non-zero 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 re-merged process is then applied to the raw temporal intervals. During the re-merged process, two neighboring intervals are merged if the distance between them is less than a threshold. Fig. 2.6 shows an example of the re-merge 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.

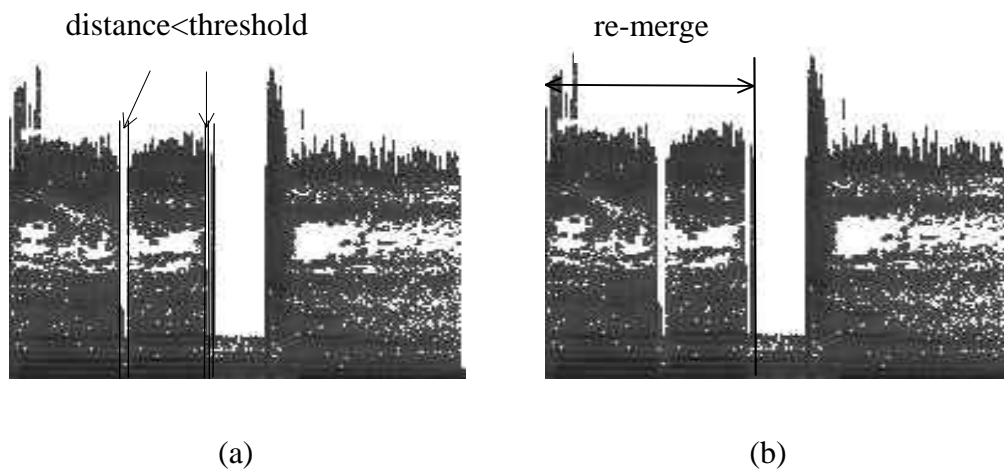


Fig. 2.6. An example of the re-merged process. (a) Initial temporal intervals. (b) Result after re-merged process.

2.2.2. 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 on-line classification, features described above are computed on the fly with incoming audio data.

2.2.2.1 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.

2.2.2.2 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 conducted. 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 [15] for class w_k , $d_k(X)$ has the form:

$$d_k(X) = \ln P(w_k) - \frac{1}{2} \ln |C_k| - \frac{1}{2} (X - m_k)^T C_k^{-1} (X - m_k) , \quad (2.3)$$

where m_k and C_k are the mean vector and covariance matrix of X , and $P(w_k)$ is the 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 that 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 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. To do 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.

2.3. EXPERIMENTAL RESULTS



In order to do comparison, we have collected a set of 700 generic audio pieces of different types of sound according to the collection rule described in [14] as the testing database. Care was taken to obtain a wide variation in each category, and most of clips are taken from MPEG-7 content set [14, 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.

2.3.1 Classification Results

Tables I and II show the results of the coarse-level classification and the final classification results, respectively. From Table II, 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, it needs less than one twentieth of the time required to play the audio clip for processing an audio clip. The only computational expensive part is the spectrogram, and the other processing is simple by comparison (e.g. variances, peak finding, etc). In order to do comparison, we also like to cite the efficiency of the existing system described in [14], which also includes the five audio classes considered in our method and uses similar database to ours. The authors of [14] 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 2.1
COARSE-LEVEL CLASSIFICATION RESULTS.

Audio Type		Number	Correct Rates
Single-Type	Pure Speech	200	100%
Sounds	Pure Music	200	100%
Hybrid-type	Song	200	100%
Sounds	Speech with MB	50	100%
	Speech with NB	50	100%

TABLE 2.2
FINAL CLASSIFICATION RESULTS.

Audio Type		Number	Correct Rates
Single-Type	Pure Speech	200	100%
Sounds	Pure Music	200	97.6%
Hybrid-type	Song	200	98.53%
Sounds	Speech with MB	50	96.5%
	Speech with NB	50	100%

2.4. SUMMARY

In this chapter, 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.

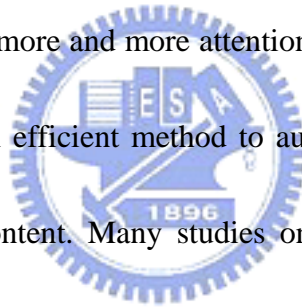
Besides the general audio types such as music and speech tested in existing work, we have taken hybrid-type sounds (speech with music background, speech with environmental noise background, and song) into account. While current existing approaches for audio content analysis are normally developed for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

CHAPTER 3

A NEW APPROACH FOR AUDIO CLASSIFICATION AND SEGMENTATION USING GABOR WAVELETS AND FISHER LINEAR DISCRIMINATOR

3.1. INTRODUCTION

In recent years, audio, as an important and integral part of many multimedia applications, has been gained more and more attentions. Rapid increase in the amount of audio data demands for an efficient method to automatically segment or classify audio stream based on its content. Many studies on audio content analysis [1-14] haven been proposed.



A speech/music discriminator was provided in [3], based on thirteen features including cepstral coefficients, four multidimensional classification frameworks are compared to achieve better performance. The approach presented by Saunders [5] takes a simple feature space, it is performed by exploiting lopsidedness of the distribution of zero-crossing rate, where speech signals show a marked rise that is not common for music signals. In general, for speech and music, it is not hard to reach a relatively high level of discrimination accuracy since they have quite different

properties in both time and frequency domains.

Besides speech and music, it is necessary to take other kinds of sounds into consideration in many applications. The classifier proposed by Wyse and Smoliar [7] classifies audio signals into “music,” “speech,” and “others.” It was developed for the parsing of news stories. In [8], audio signals are classified into speech, silence, laughter, and non-speech sounds for the purpose of segmenting discussion recordings in meetings. However, the accuracy of the segmentation resulted using this method varies considerably for different types of recording. Besides the commonly studied audio types such as speech and music, the research in [12-14] has taken into account hybrid-type sounds, e.g., the speech signal with the music background and the singing of a person, which contain more than one basic audio type and usually appear in documentaries or commercials. In [12], 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. Zhang and Kuo [14] 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. Accuracy of above 90% is reported. However, this method is complex and time-consuming due to the computation of autocorrelation function. Besides, the thresholds used in this approach are empirical, they are improper when the source of audio signals is changed.

In this chapter, we will provide two classifiers, one is for speech and music (called two-way); the other is for five classes (called five-way) that are pure speech, music, song, speech with music background, and speech with environmental noise background. Based on the classification results, we will propose a merging algorithm to divide an audio stream into some segments of different classes.

One basic issue for content-based classification of audio sound is feature selection. The selected features should be able to represent the most significant properties of audio sounds, and they are also robust under various circumstances and general enough to describe various sound classes. The issue in the proposed method is addressed in the following: first, some perceptual features based on the Gabor wavelet filters [15-16] are extracted as initial features, then Fisher Linear Discriminator (FLD) [17] is applied to these initial features to explore the features with the highest discriminative ability.

Note that FLD is a tool for multigroup data classification and dimensionality

reduction. It maximizes the ratio of between-class variance to within-class variance in any particular data set to guarantee maximal separability. Experimental results show that the proposed method can achieve an accuracy rate of discrimination over 98% for a two-way speech/music discriminator, and more than 95% for a five-way classifier which uses the same database as that used in the two-way discrimination. Based on the classification result, we can also identify scene breaks in audio sequence quite accurately. Experimental results show that our method can detect more than 95% of audio type changes. These results demonstrate the capability of the proposed audio features for characterizing the perceptual content of an audio sequence.

The rest of the chapter is organized as follows. In Section 3.2, the proposed method is described in details. Experimental results and discussion are presented in Section 3.3. Finally, in Section 3.4, we give a summary.

3.2. THE PROPOSED METHOD

The block diagram of the proposed method is shown in Fig. 3.1. It is based on the spectrogram and consists of five phases: time-frequency distribution (TFD) generation, initial feature extraction, feature selection, classification and segmentation. First, the input audio is transformed to a spectrogram, $I(x, y)$, as mentioned in

Multi-resolution Short Time Fourier Transform section (Chapter 1, Section 1.3.2).

Second, for each clip with one-second window, some Gabor wavelet filters will be applied to the resulting spectrogram to extract a set of initial features. Third, based on the extracted initial features, the Fisher Linear Discriminator (FLD) is used to select the features with the best discriminative ability and also to reduce feature dimension. Fourth, based on the selected features, classification method is then provided to classify each clip. Finally, based on the classified clips, a segmentation technique is presented to identify scene breaks in each audio stream. In what follows, we will describe the details of the proposed method.

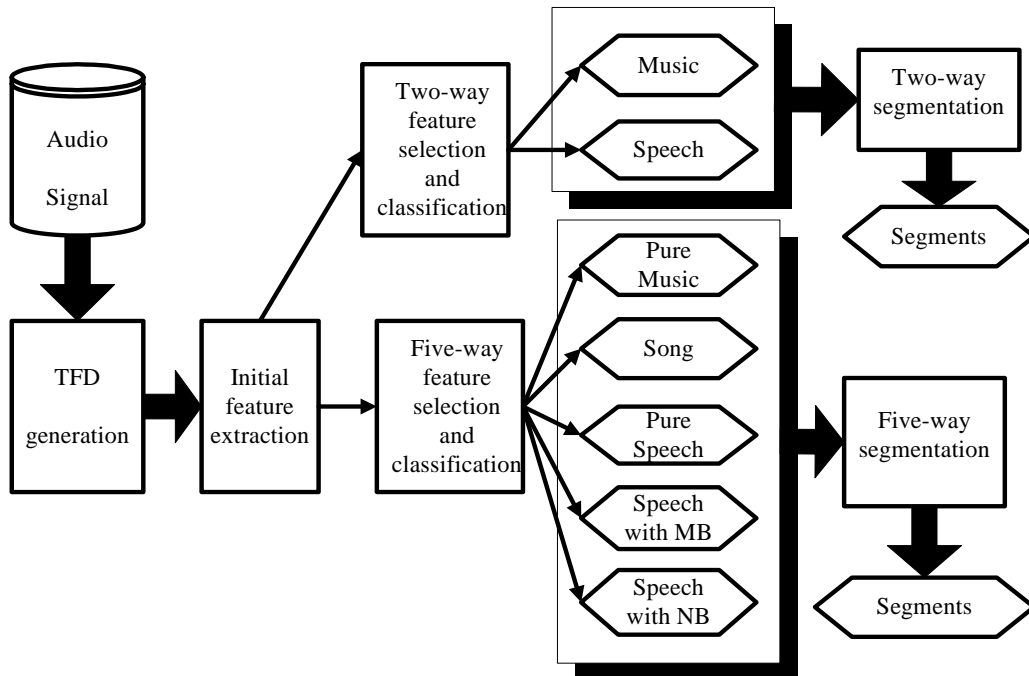


Fig. 3.1. Block diagram of the proposed method, where “MB” and “NB” are the abbreviations for “music background” and “noise background”, respectively.

3.2.1 Initial Feature Extraction

Generally speaking, the spectrogram is a good representation for the audio since it is often visually interpretable. By observing a spectrogram, we can find that the energy is not uniformly distributed, but tends to cluster to some patterns (see Fig. 3.2 (a), 3.2 (b)). All curve-like patterns are called tracks [31]. Fig. 3.2 (a) shows that for a music signal, some line tracks corresponding to tones will exist on its spectrogram. Fig. 3.2 (b) shows some patterns including clicks (broadband, short time), noise burst (energy spread over both time and frequency), and frequency sweeps in a song spectrogram.

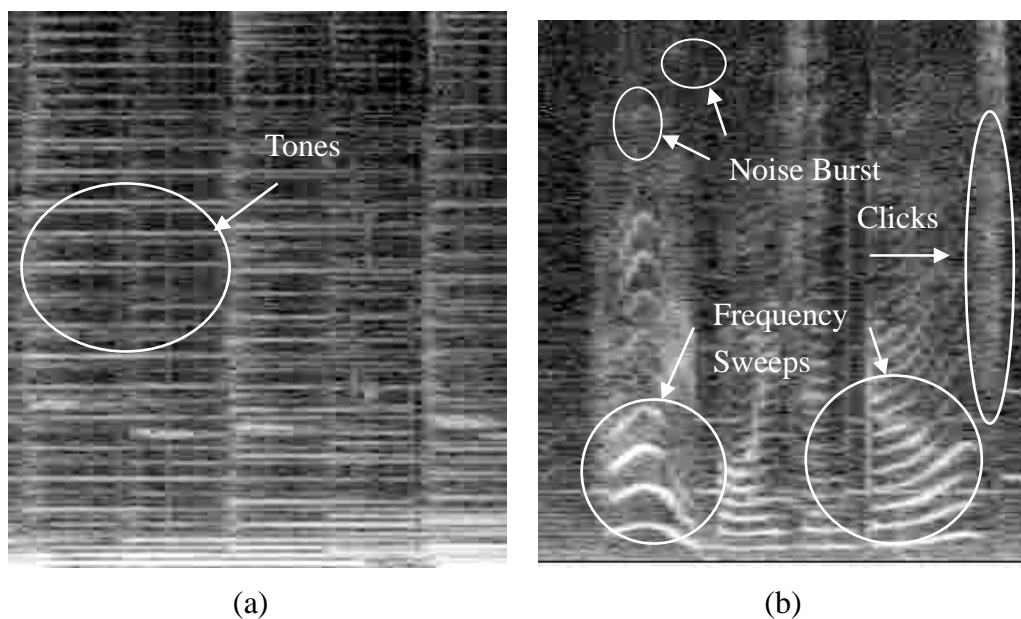
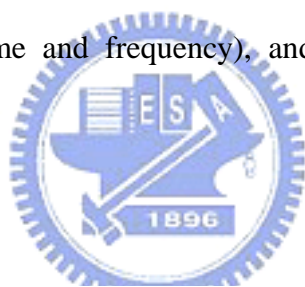


Fig. 3.2. Two examples to show some possible different kinds of patterns in a spectrogram. (a) Line tracks corresponding to tones in a music spectrogram. (b) Clicks, noise burst and frequency sweeps in a song spectrogram.

Thus, if we can extract some features from a spectrogram to represent these patterns, the classification should be easy. Smith and Serra [32] proposed a method to extract tracks from a STFT spectrogram. Once the tracks are extracted, each track is classified. However, tracks are not well suited for describing some kinds of patterns such as clicks, noise burst and so on. To treat all kinds of patterns, a richer representation is required. In fact, these patterns contain various orientations and spatial scales. For example, each pattern formed by lines (see Fig. 3.2 (a)) will have a particular line direction (corresponding to orientation) and width (corresponding to spatial scale) between two adjacent lines; each pattern formed by curves (see Fig. 3.2 (b)) contains multiple line directions and a particular width between two neighboring curves. Since Gabor wavelet transform provides an optimal way to extract those orientations and scales [27], in this chapter, we will use the Gabor wavelet functions to extract some initial features to represent those patterns. The detail will be described in the following section.



3.2.1.1 Gabor Wavelet Functions and Filters Design

Two-dimensional Gabor kernels are sinusoidally modulated Gaussian Functions. Let $g(x, y)$ be the Gabor kernel, its Fourier Transform $G(u, v)$ can be defined as

follows [28]:

$$g(x, y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right) \exp\left[\frac{-1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi j\omega x\right], \quad (3.1)$$

$$G(u, v) = \exp\left(\frac{-1}{2}\left[\frac{(u-\omega)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right]\right), \quad (3.2)$$

where $\sigma_u = \frac{1}{2\pi\sigma_x}$ and $\sigma_v = \frac{1}{2\pi\sigma_y}$ and ω is the center frequency.

Gabor wavelets are sets of Gabor kernels which will be applied to different subbands with different orientations. It can be obtained by appropriate dilations and rotations of $g(x, y)$ through the following generating functions [28]:

$$g_{mn}(x, y) = a^{-m} g(x', y'), \quad a > 1, \quad m, n = \text{integer},$$

$$x' = a^{-m}(x \cos \theta + y \sin \theta), \quad \text{and} \quad y' = a^{-m}(-x \sin \theta + y \cos \theta), \quad (3.3)$$

$$a = \left(\frac{\omega_h}{\omega_l}\right)^{\frac{1}{S-1}}, \quad (3.4)$$

$$\sigma_u = ((a-1)\omega_h) / ((a+1)\sqrt{2\ln 2}), \quad (3.5)$$

$$\sigma_v = \tan\left(\frac{\pi}{2k}\right) \left[\omega_h - 2\ln 2 \left(\frac{\sigma_u^2}{\omega_h}\right)\right] \left[2\ln 2 - \frac{(2\ln 2)^2 - \sigma_u^2}{\omega_h^2}\right]^{\frac{-1}{2}}, \quad (3.6)$$

where $\theta = \frac{n\pi}{K}$, $n = 0, 1, \dots, K-1$, $m = 0, 1, \dots, S-1$, K is the total number of orientations, S is the number of scales in the multi-resolution decomposition, ω_h and ω_l are the highest and the lowest center frequency, respectively. In this chapter, we set $\omega_l = \frac{3}{64}$, $\omega_h = \frac{3}{4}$, $K = 6$ and $S = 7$.

3.2.1.2 Feature Estimation and Representation

To extract the audio features, each Gabor wavelet filter, $g_{mn}(x, y)$, is first applied to the spectrogram $I(x, y)$ to get a filtered spectrogram, $W_{mn}(x, y)$, as

$$W_{mn}(x, y) = \int I(x - x_1, y - y_1) g_{mn}^*(x_1, y_1) dx_1 dy_1, \quad (3.7)$$

where * indicates the complex conjugate. The above filtering process is executed by *FFT* (fast Fourier Transform). That is

$$W_{mn}(x, y) = F^{-1} \{ F \{ g_{mn}(x, y) \} \cdot F \{ I(x, y) \} \}. \quad (3.8)$$

Since peripheral frequency analysis in the ear system roughly follows a logarithmic axis, in order to keep with this way, the entire frequency band $[0, F_s/2]$ is divided into six subbands of unequal width: $F_1=[0, F_s/64]$, $F_2=[F_s/64, F_s/32]$, $F_3=[F_s/32, F_s/16]$, $F_4=[F_s/16, F_s/8]$, $F_5=[F_s/8, F_s/4]$, and $F_6=[F_s/4, F_s/2]$. In our experiments, high frequency components above $F_s/4$ (i.e., subband $[F_s/4, F_s/2]$) are discarded to avoid the influence of noise. Then, for each interested subband F_i , the directional histogram, $H_i(m, n)$, is defined to be

$$H_i(m, n) = \frac{N_i(m, n)}{\sum_{n=0}^5 N_i(m, n)}, \quad i = 0, \dots, 4, \quad (3.9)$$

$$W_{mn}^i(x, y) = \begin{cases} 1, & \text{if } W_{mn}(x, y) > T_m \text{ and } y \in F_i \\ 0, & \text{otherwise} \end{cases}, \quad (3.10)$$

$$N_i(m, n) = \sum_x \sum_y W_{mn}^i(x, y), \quad (3.11)$$

where $m = 0, \dots, 6$. and $n = 0, \dots, 5$. Note that $N_i(m, n)$ is the number of pixels in the filtered spectrogram $W_{mn}(x, y)$ at subband F_i , scale m and direction n with value larger than threshold T_m . T_m is set as

$$T_m = \mu_m + \sigma_m. \quad (3.12)$$

where

$$\mu_m = \frac{\sum_{n=0}^5 \sum_x \sum_y W_{mn}(x, y)}{N_m}, \quad \sigma_m = \left(\frac{\sum_{n=0}^5 \sum_x \sum_y (W_{mn}(x, y) - \mu_m)^2 / N_m}{N_m} \right)^{\frac{1}{2}},$$

and N_m is the number of pixels over all the 6 filtered spectrogram $W_{mn}(x, y)$ with scale m .



An initial feature vector, f , is now constructed using $H_i(m, n)$ as feature components. Recall that in our experiments, we use seven scales ($S=7$), six orientations ($K=6$) and five subbands, this will result in a $7 \times 6 \times 5$ dimensional initial feature vector

$$f = [H_0(0,0), H_0(0,1), \dots, H_4(6,5)]^T. \quad (3.13)$$

3.2.1.3 Feature Selection and Audio Classification

The initial features are not used directly for classification since some features give poor separability among different classes and inclusion of these features will

lower down classification performance. In addition, some features are highly correlated so that redundancy will be introduced. To remove these disadvantages, in this chapter, the Fisher Linear Discriminator (FLD) is applied to the initial features to find those uncorrected features with the highest separability. Before describing FLD, two matrices, between-class scatter and within-class scatter, will first be introduced. The within-class scatter matrix measures the amount of scatter between items in the same class and the between-class scatter matrix measures the amount of scatter between classes.

For the i^{th} class, the within-class scatter matrix S_w^i is defined as

$$S_w^i = \sum_{x_k^i \in X_i} (x_k^i - \mu_i)(x_k^i - \mu_i)^T, \quad (3.14)$$

the total within-class scatter matrix S_w is defined as

$$S_w = \sum_{i=1}^C S_w^i, \quad (3.15)$$

and the between-class scatter matrix S_b is defined as

$$S_b = \sum_{i=1}^C N_i (\mu_i - \mu)(\mu_i - \mu)^T, \quad (3.16)$$

where μ_i is the mean of class X_i , N_i is the number of samples in class X_i , x_k^i is the k^{th} sample in X_i , and C is the number of classes.

In FLD, a matrix $V_{opt} = \{v_1, v_2, \dots, v_{C-1}\}$ is first chosen, it satisfies the following equation:

$$V_{opt} = \arg \max_V \left| \frac{V^T S_b V}{V^T S_w V} \right|. \quad (3.17)$$

In fact, $\{v_1, v_2, \dots, v_{C-1}\}$ is the set of generalized eigenvectors of S_b and S_w corresponding to the $C-1$ largest generalized eigenvalues $\{\lambda_i | i = 1, 2, \dots, C-1\}$ [17], i.e.,

$$S_b v_i = \lambda_i S_w v_i. \quad (3.18)$$

Note that in this chapter, two classes and five classes (i.e., $C = 2$ and $C = 5$) are used and one-second audio clip is taken as the basic classification unit.

Based on V_{opt} , the initial feature vector for each one-second audio clip in the training data and testing data is projected to the space generated by V_{opt} to get a new feature vector f' with dimension $C-1$. f' is then used to stand for the audio clip. Before classification, it is important to give a good similarity measure. In our experiments, the Euclidean distance worked better than others (e.g., Mahalanobis, covariance, etc.). For each test sample, x_j with feature vector f'_j , the Euclidean distance between the test sample and the class center of each class in the space generated by V_{opt} is evaluated. Then the sample is assigned to the class with minimum distance. That is, x_j is assigned as class C'_j according to the following criterion:

$$C'_j = \arg \min_i \|f'_j - \mu'_i\|, i = 1, 2, \dots, C, \quad (3.19)$$

where μ'_i is the mean vector of the projected vectors of all test samples in class i .

Fig. 3.3 shows an example of using a two-way speech/music discriminator. In the figure, “x” stands for the projected result of an music signal, “o” stands for the projected result of a speech signal. From this figure, we can see that through FLD, music and speech samples can be easily separated. Fig. 3.4 outlines the process of feature selection and classification.

Two problems arise when using Fisher discriminator. First, the matrices needed for computation are very large. Second, since we may have fewer training samples than the number of features in each sample, the data matrix is rank deficient. To avoid the problems described above, it is possible to solve the eigenvectors and eigenvalues of a rank deficient matrix by using a generalized singular value decomposition routine. One simple and speedup solution [33] is taken in this chapter.

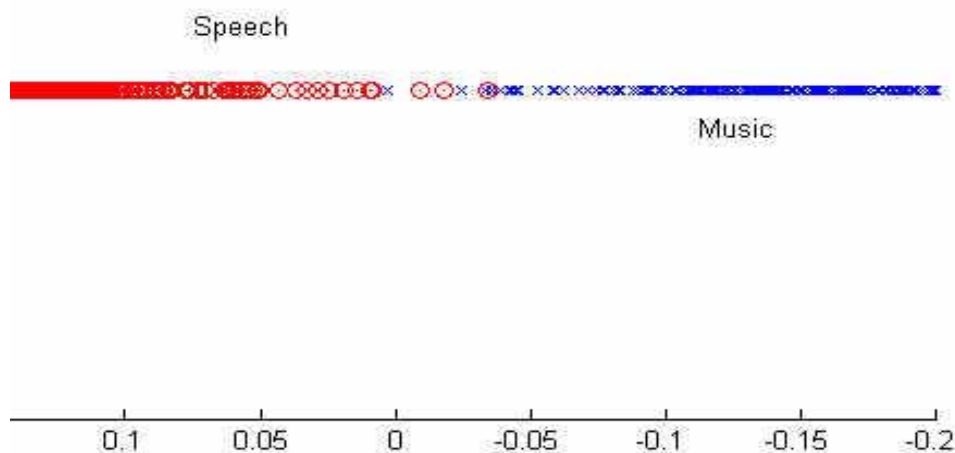


Fig. 3.3. An example of using FLD for two-way speech/music discriminator.

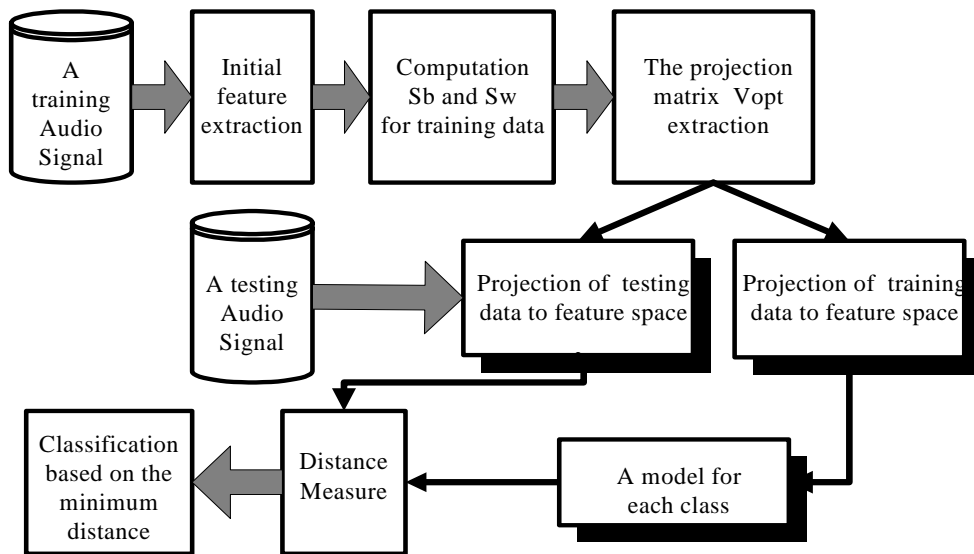


Fig. 3.4. A block diagram of feature selection and classification using

3.2.1.4 Segmentation



The segmentation is to divide an audio sequence into semantic scenes called “audio scene” and to index them as different audio classes. Due to some classification errors, a reassigning algorithm is first provided to rectify these classification errors.

For example, if we detect a pattern like speech-music-speech, and the music subpattern lasts a very short time, we can conclude that the music subpattern should be speech. First, for each one-second audio clip, the similarity measure between the audio clip and the center of its class is defined as

$$Similarity = 1 - \frac{dist_{\min}}{\sum_{j=1}^5 dist_j}, \quad dist_{\min} = \min_j dist_j, \quad (3.20)$$

where $dist_j$ is the Euclidean distance between the clip and the j^{th} class center in

the feature space. If the similarity measure is less than 0.9, mark the clip as ambiguous. Note that ambiguous clips often arise in transition periods. For example, if a transition happens when speech stops and music starts, then each clip in the transition will contain both speech and music information. Then, each ambiguous clip will be reassigned as the class of the nearest unambiguous clip. After the reassignment is completed, all neighboring clips with the same class are merged into a segment. Finally, for each audio segment, the length is evaluated. If the length is shorter than the threshold T ($T=3$ second), each clip in the segment is reassigned as the class of one of its two neighboring audio segments with the least Euclidean distance between the clip and the center of class of the selected neighboring segment.

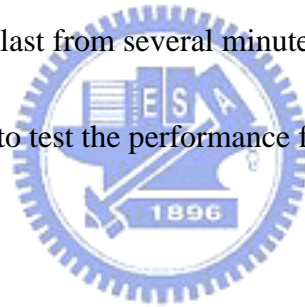


3.3. EXPERIMENTAL RESULTS

In order to do comparison, we have collected a set of 700 generic audio pieces (with duration from several seconds to no more than one minute) of different types of sound according to the collection rule described in [14] as the testing database. Care was taken to obtain a wide variation in each category, and some of clips are taken from MPEG-7 content set [23]. The database contains 100 pieces of classical music played with varied instruments, 100 other music pieces of different styles (jazz, blues,

light music, etc.), 200 pieces of pure speech in different languages (English, Chinese, Japanese, etc.), 200 pieces of song sung by male, female, or children, 50 pieces of speech with background music (e.g. commercials, documentaries, etc.), and 50 pieces of speech with environmental noise (e.g. sport broadcast, news interview, etc.). These shorter audio clips are stored as 16-bit per sample with 44.1 kHz sampling rate in the WAV file format and are used to test the audio classification performance. Note that we take one-second audio signal as a test unit.

We also collected a set of 15 longer audio pieces recorded from movies, radio or video programs. These pieces last from several minutes to an hour and contain various types of audio. They are used to test the performance for audio segmentation.

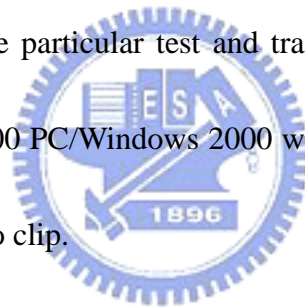


3.3.1 Audio Classification Results

In order to examine the robust use for a variety of the audio source and the accuracy for audio classification, we present two experiments. One is two-way discrimination and the other is five-way discrimination. Concerning the two-way discrimination, we try to classify the audio set into two categories: music and speech. As for the five-way discrimination, the audio set will be classified into five categories: pure speech, pure music, song, speech with music background, and speech with

environmental noise background.

Tables 3.1 and 3.2 show the results of the classification. From these tables, we can see that the proposed classification approach for generic audio data can achieve an over 98% accuracy rate for the speech/music discrimination, and more than 95% for the five-way classification. Both classifiers use the same testing database. It is worth mentioning that 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 is stable and independent on the particular test and training sets. The experiments are carried out on a Pentium II 400 PC/Windows 2000 with less than one-eleventh of the time required to play the audio clip.



In our experiments, there are several misclassifications. From Table 3.2, we can see that most errors occur in the speech with music background category. This is due to that the music or speech component is weak. In order to do comparison, we also like to cite the efficiency of the existing system described in [14], which also includes the five audio classes considered in our method and uses similar database to ours. The authors of [14] 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 3.1
TWO-WAY CLASSIFICATION RESULTS

Audio Type	Number	Correct Rate
Speech	300	98.17%
Music	400	98.79%

TABLE 3.2
FIVE-WAY CLASSIFICATION RESULTS

Audio Type	Number	Discrimination Results				
		Pure Music	Song	Pure Speech	Speech with MB	Speech with NB
Pure Music	200	94.67%	3.21%	1.05%	1.07%	0%
Song	200	0.8%	96.43%	0%	1.97%	0.8%
Pure Speech	200	0%	0.14%	98.40%	0.11%	1.35%
Speech with MB	50	1.01%	4.2%	3.10%	89.62%	2.07%
Speech with NB	50	0.15%	0.71%	1.28%	0.63%	97.23%

3.3.2 Audio Segmentation Results

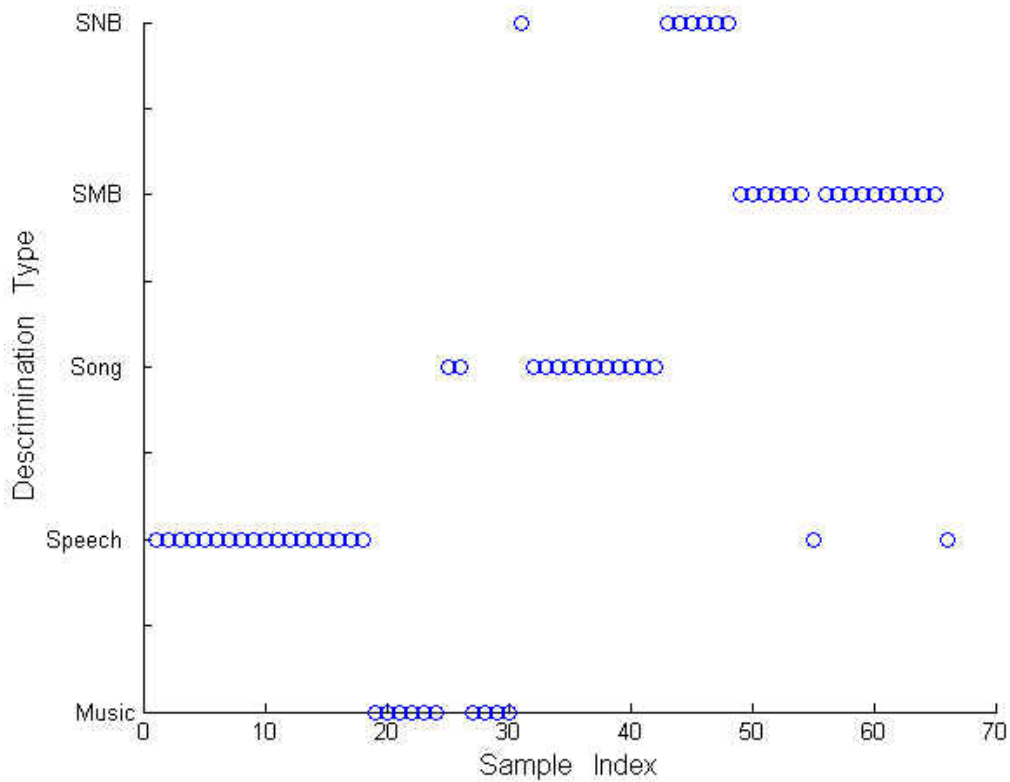
We tested our segmentation procedure with audio pieces recorded from radio, movies, and video programs. We made a demonstration program for online audio

segmentation and indexing as shown in Fig. 3.5. Fig. 3.5 (a) shows the classification result for a 66 second audio piece recorded from MPEG-7 data set CD19 that is a Spanish cartoon video called “Don Quijote de la Mancha.” Fig. 3.5 (b) shows the result of applying the segmentation method to Fig. 3.5 (a). Besides the above example, we also performed experiments on other audio pieces.

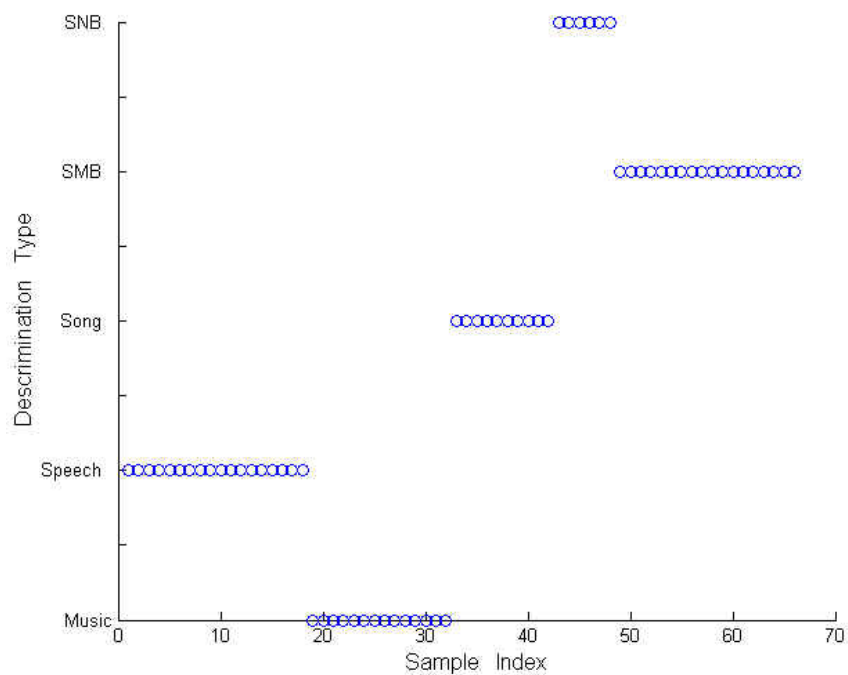
Listed in Table 3.3 is the result of the audio segmentation, where miss-rate and over-rate are defined as the ratio between the number of miss-segmented ones and the actual number of segments, and the ratio between the number of over-segmented ones and the actual number of segments in audio streams, respectively. Besides, error rate is defined as the ratio between the number of segments indexed in errors and the actual number of segments in audio stream.



The first column shows the segmentation result without applying the reassignment process to the classification result, and the second column shows the segmentation result using the reassignment process. The experiments have shown that the proposed scheme achieves satisfactory segmentation and indexing. Using human judgement as the ground truth, our method can detect more than 95% of audio type changes.



(a)



(b)

Fig. 3.5. Demonstration of audio segmentation and indexing, where “SMB” and “SNB” are the abbreviations for “speech with music background” and “speech with noise background”, respectively. (a) Original result. (b) Final results after applying the segmenting algorithm to (a).

TABLE 3.3
SEGMENTATION RESULTS.

	Without Using Reassignment	Using Reassignment
Miss-Rate	0%	1.1%
Over-Rate	5.2%	1.8%
Error-Rate	2.5%	1.3%

3.4. SUMMARY



In this chapter, we have presented a new method for the automatic classification and segmentation of generic audio data. An accurate classification rate higher than 95% was achieved. The proposed scheme can treat a wide range of audio types. Furthermore, the complexity is low due to the easy computing of audio features, and this makes online processing possible. The experimental results indicate that the extracted audio features are quite robust.

Besides the general audio types such as music and speech tested in existing work, we have taken into account other different types of sounds including hybrid-type

sounds (e.g. speech with music background, speech with environmental noise background, and song). While current existing approaches for audio content analysis are normally developed for specific scenarios, the proposed method is generic and model free. Thus, it can be widely applied to many applications.



CHAPTER 4

CONTENT-BASED AUDIO RETRIEVAL BASED ON GABOR WAVELETS

4.1. INTRODUCTION

The recent emerging of multimedia and the tremendous growth of multimedia data archives have made the effective management of multimedia databases become a very important and challenging task. Therefore, developing an efficient searching and indexing technique for multimedia databases become very important and have drawn lots of attention recently. As many research works were done on the content-based retrieval of image and video data, less attention was received to the content-based retrieval of audio data.

In recent years, techniques for audio information retrieval have started emerging as research prototypes [9-24]. These systems can be classified into two major paradigms [22, 34]. In the first paradigm, the user sings a melody and similar audio files containing that melody are retrieved. This kind of approaches [18] is called “Query by Humming” (QBH). It has the disadvantage of being applicable only when the audio data is stored in symbolic form such as MIDI files. The conversion of

generic audio signals to symbolic form, called polyphonic transcription, is still an open research problem in its infancy [22]. Another problem with QBH is that it is not applicable to several musical genres such as Dance music where there is no singable melody that can be used as a query. The second paradigm [9-10, 13, 16-24] is called “Query-by-Example” (QBE), a reference audio file is used as the query and audio files with similar content are returned and ranked by their similarity degree. In order to search and retrieve general audio signals such as the raw audio files (e.g. mp3, wave, etc.) on the web or databases, only the QBE paradigm is currently applicable.

In this dissertation, we will develop a QBE system that will work directly on real world raw audio data without attempting to transcribe the music.



Wold, et al. [9] proposed an approach to retrieve the audio objects based on their content in waveform. In this approach, an N-vector for a given sound is constructed according to the acoustical features including loudness, pitch, brightness, bandwidth, and harmonicity. The N-vector is then used to classify sounds for similar searching. This method is only suitable for sounds with a single timbre. Besides, the method is supervised and not adequate to index general audio content. An approach based on the histogram model of the zero-crossing features for searching quickly through broadcast audio data was provided in [15]. In this approach, a certain reference template is defined and applied on each audio stream to find whether it contains the desired

reference sound. The accuracy of the result using this method varies considerably for different types of recording. Besides, the audio segment to be searched should be known *a priori* in this algorithm.

Footo [16] proposed a data driven approach for audio data retrieval by computing the Mel-frequency cepstral coefficients (MFCCs) of an audio signal first. Then a learning algorithm is applied on these MFCCs to generate a quantization tree. Each kind of audio signals is inserted into the corresponding bin in the quantization tree. Cosine measurement or Euclidean distance can be used to measure the similarity between two bins. A QBE system called “SoundSpotter” [21] provides a sound classification tools to classify a large database into several categories and finds the best matches to the selected query sound using state-path histograms. It is also based on the MFCCs representation. Both of the above-mentioned two MFCC-based approaches are not suitable for melody retrieval (e.g. music) since the MFCC-based features do not capture enough information about the pitch content, rather, they characterize the broad shape of the spectrum. In [24], local peaks in spectrogram are identified and a spectral vector is extracted near each peak. Since the parameters used in the peak identification algorithm are too many and empirical, they are improper when the source of audio signals is changed.

In this chapter, based on the Gabor wavelet features, we will propose a method

for content-based retrieval of perceptually similar music pieces in audio documents. It is based on the QBE paradigm and allows the user to select a reference passage within an audio file and retrieve perceptually similar passages such as repeating phrases within a music piece, similar music clips in a database or one song sung by different persons or in different languages. The proposed method consists of four phases: time-frequency distribution (TFD) generation, initial feature extraction, feature selection and similarity measurement. First, the input audio stream is transformed to a spectrogram and divided into clips, each of which contains one-second audio information and will meet the human auditory system (HAS) [29]. Second, for each clip with one-second window, a set of initial frame-based features are extracted based on the Gabor wavelet filters [27-28]. Third, based on the extracted initial features, the Singular Value Decomposition (SVD) [25] is used to perform the feature selection and to reduce the feature dimension. Finally, a similarity measuring technique is provided to perform pattern matching on the resulting sequences of feature vectors.

Experimental results show that the proposed method can achieve over 96% accuracy rate for audio retrieval and the complexity is low enough to allow operation on today's personal computers and other cost-effective computing platforms. These results demonstrate the capability of the proposed audio features for characterizing the perceptual content of an audio sequence. The rest of the chapter is organized as

follows. In Section 4.2, the proposed method is described in details. Experimental results and discussion are presented in Section 4.3. Finally, in Section 4.4, we give a summary.

4.2. THE PROPOSED METHOD

The block diagram of the proposed method is shown in Fig. 4.1. It is based on the spectrogram and consists of four phases: time-frequency distribution (TFD) generation, initial feature extraction, feature selection and similarity measurement. First, the input audio is transformed to a spectrogram, $I(x, y)$, as mentioned in *Multi-resolution Short Time Fourier Transform* section (Chapter 1, Section 1.3.2). Second, for each clip with one-second window, some Gabor wavelet filters will be applied to the resulting spectrogram to extract a set of initial features. Third, based on the extracted initial features, the Singular Value Decomposition (SVD) [25] is used to

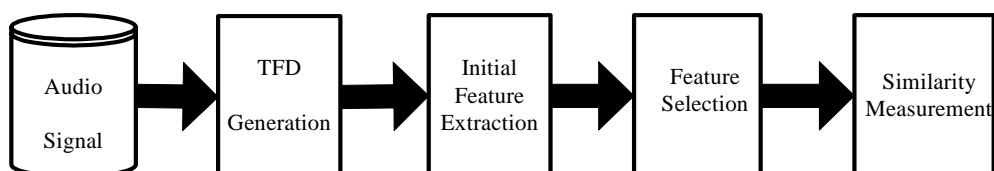


Fig. 4.1. Block diagram of the proposed method.

perform the feature selection and to reduce the feature dimension. Finally, based on the selected features, a similarity measure is provided to measure the similarity of audio data. In what follows, we will describe the details of the proposed method.

4.2.1. Initial Feature Extraction

Generally speaking, the spectrogram is a good representation for the audio since it is often visually interpretable. By observing a spectrogram, we can find that the energy is not uniformly distributed, but tends to cluster to some patterns. All curve-like patterns are called tracks [31]. Fig. 4.2 (a) shows that for a musical instrument signal, some line tracks corresponding to tones will exist on its spectrogram. Fig. 4.2 (b) shows some patterns including clicks (broadband, short time), noise burst (energy spread over both time and frequency), tones, and frequency sweeps in a song spectrogram. Thus, if we can extract some features from a spectrogram to represent these patterns, the retrieval should be easy. Smith and Serra [32] proposed a method to extract tracks from a STFT spectrogram. Once the tracks are extracted, each track is classified. However, tracks are not well suited for describing some kinds of patterns such as clicks, noise burst and so on. To treat all kinds of patterns, a richer representation is required. In fact, these patterns contain

various orientations and spatial scales. For example, each pattern formed by lines (see Fig. 4.2 (a)) will have a particular line direction (corresponding to orientation) and width (corresponding to spatial scale) between two adjacent lines; each pattern formed by curves (see Fig. 4.2 (b)) contains multiple line directions and a particular width between two neighboring curves. Since Gabor wavelet transform provides an optimal way to extract those orientations and scales [29], in this chapter, we will use the Gabor wavelet functions to extract some initial features to represent the needed patterns. The detail will be described in the following section.

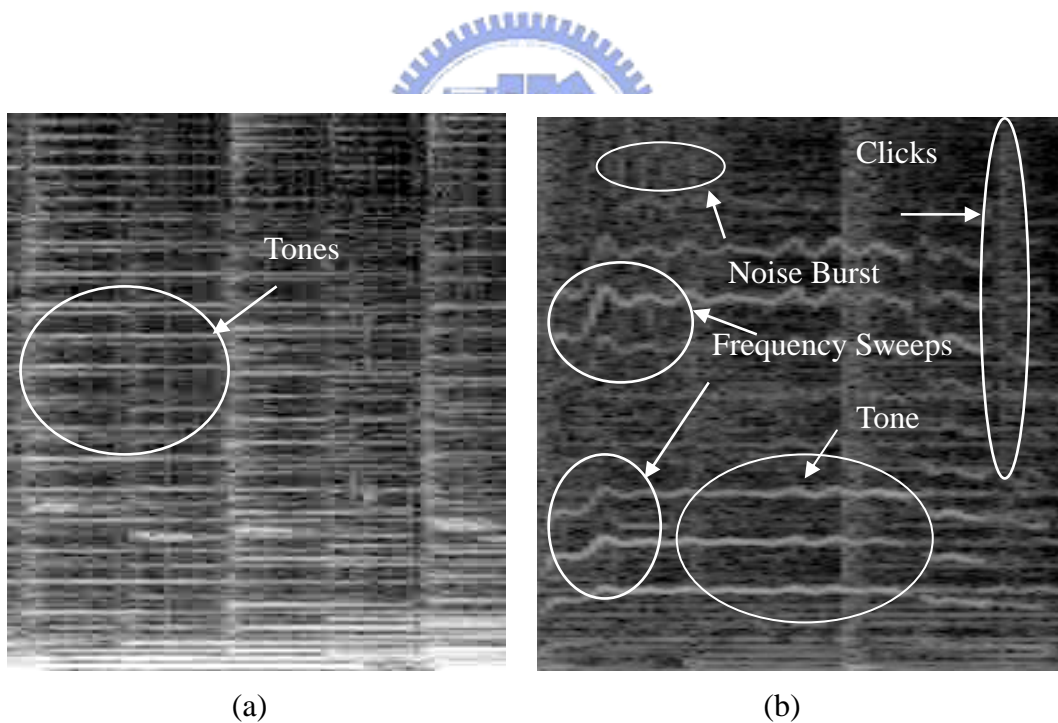


Fig. 4.2. Two examples to show some possible different kinds of patterns in a spectrogram. (a) Line tracks corresponding to tones in a musical instrument spectrogram. (b) Clicks, noise burst, tones, and frequency sweeps in a song spectrogram.

4.2.1.1 Feature Estimation

In this chapter, we will deal with musical audio signal including musical instrument and song. Most of the current works only deal with the monophonic sources, in this chapter we will also consider polyphonic music. Polyphonic music is more common, but it is also more difficult to represent. The most meaningful feeling of human perception for the music data is primarily the pitch and timbre. Both of them are correlated with the tones. For example, the fundamental tone decides the pitch that we hear, and the harmonics decide the timbre. Based on the above-observation for the spectrogram (see Fig. 4.2 (a) and 4.2 (b)), we find that some line tracks corresponding to tones will exist in the spectrogram. Thus, if we can extract the features about tones, the retrieval should be easy.

Since through our observation, most prominent tracks are near horizontal, in this chapter, we only take one orientation that is horizontal. Thus, each Gabor wavelet filter as mentioned in *Gabor Wavelet Functions and Filters Design* section (Chapter 3, Section 3.2.1.1), $g_m(x, y)$, can be briefly represented by $g_m(x, y)$. Note that in our experiments, we set $\omega_l = \frac{3}{64}$, $\omega_h = \frac{3}{4}$, $K = 1$ and $S = 7$. To extract the audio features, each Gabor wavelet filter, $g_m(x, y)$, is first applied to the spectrogram $I(x, y)$ to get a filtered spectrogram, the spectrum of which is represented by

$W_m(u, v)$ called spectrogram spectrum. That is

$$W_m(u, v) = F\{g_m(x, y)\} \cdot F\{I(x, y)\}, \quad (4.1)$$

where $F\{\cdot\}$ is a fast Fourier Transform.

Up to now, there are S spectrogram spectrum with scale m , $W_m(u, v)$, to be available. Since, in each audio signal, those tracks appear in the corresponding spectrum have a certain scale, not all these spectrogram spectrum are used to perform the feature estimation, only the one with the maximum contrast (which corresponds to the track scale) is used. To reach this goal, the vertical profile of the spectrum, $P_m(u)$

($m = 1, 2, \dots, S$), is constructed as follows:

$$P_m(u) = \sum_v W_m(u, v). \quad (4.2)$$

Let M_p be the number of the local peaks (u_1, u_2, \dots, u_{M_p}) in $P_m(u)$, $P_m(u_i)$

($i = 1, 2, \dots, M_p$) be the magnitudes of these peak points, and

$$P_m^{\max} = \max_{u_i} P_m(u_i). \quad (4.3)$$

Then the contrast is defined as

$$\text{contrast}_m = P_m^{\max} - \frac{1}{M_p} \sum_{i=1}^{M_p} P_m(u_i). \quad (4.4)$$

Let

$$mc = \arg \text{contrast}_m, \quad (4.5)$$

then the spectrogram spectrum, $W_{mc}(u, v)$, and the corresponding spectrogram,

$w_{mc}(x, y)$, are used to do initial feature extraction.

Fig. 4.3 (a) shows an example of the Gabor-wavelet filtered spectrogram with the maximum contrast, $w_{mc}(x, y)$. From Fig. 4.3 (a), we can see that the tracks in the figure are somewhat obscured, to remove this phenomenon, an enhancement process [27] is applied as follows:

$$w_f(x, y) = F^{-1} \left\{ W_{mc}(u, v) \cdot |W_{mc}(u, v)|^\alpha \right\}. \quad (4.6)$$

where α is set as 1.4 and $w_f(x, y)$ is the enhanced spectrogram. Fig. 4.3 (b) shows the result of the enhancement process for Fig. 4.3 (a).

An initial feature vector, \mathbf{f} , is now constructed using $w_f(x, y)$ as feature components. Recall that in our experiments, for each clip with one-second window (M frames) is used for constructing spectrogram. Besides, high frequency components above $F_s/4$ are discarded to avoid the influence of noise. These will result in a $M \times N$ dimensional initial feature vector

$$\mathbf{f} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M]^t, \quad (4.7)$$

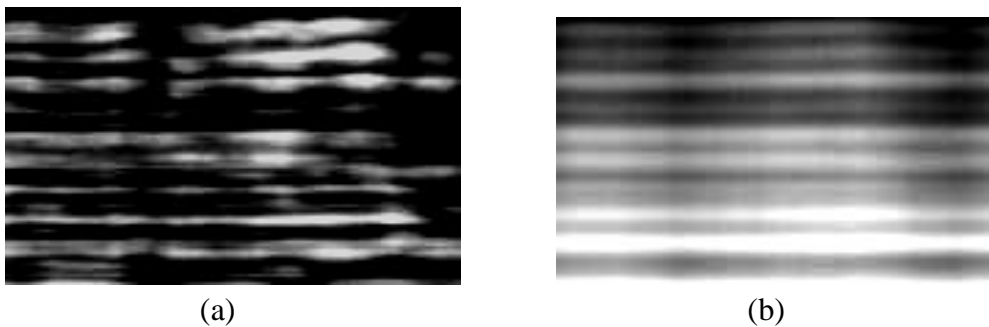


Fig. 4.3. An example to show the enhancement process performing in a spectrogram. (a) The Gabor-wavelet filtered spectrogram with the maximum contrast. (b) Enhanced spectrogram.

where $\mathbf{x}_i = [w_f(i,1), w_f(i,2), \dots, w_f(i,N)]$ ($i = 1, 2, \dots, M$) is the spectral vector of each frame in $w_f(x, y)$.

4.2.1.2 Feature Selection and Representation

The initial features are not used directly for similarity measurement since some features give poor separability among different objects and inclusion of these features will lower down the system performance. In addition, some features are highly correlated so that redundancy will be introduced. To remove these disadvantages, in this chapter, the Singular Value Decomposition (SVD) [23] is applied to the initial features to find those uncorrected features with the highest separability.



As for the SVD, it is a well-known technique for reducing the dimensionality of data while retaining maximum information content. It decomposes the data into a sum of vector outer products with vectors representing both the basis function (eigenvectors) and the projected features (eigen coefficients). A subset of the complete basis is selected to reduce data dimensionality. The loss of information is minimized because the basis functions are ordered by statistical salience; thus, functions with low information content are discarded.

Based on SVD, the initial feature vector, \mathbf{f} , for each one-second audio clip can

be decomposed into the form [28]:

$$\mathbf{f} = \mathbf{USV}^t, \quad (4.8)$$

where \mathbf{S} is a diagonal matrix containing the singular values of \mathbf{f} along its diagonal, and the columns of \mathbf{U} and \mathbf{V} are the eigenvectors (the basis function) of \mathbf{ff}^t , and $\mathbf{f}^t\mathbf{f}$ respectively. Then the basis, \mathbf{V} , is reduced by retaining only the first k basis functions. That is

$$\mathbf{V}_k = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k]. \quad (4.9)$$

And the initial feature vector \mathbf{f} is projected to the space generated by \mathbf{V}_k to get a

new feature vector \mathbf{f}' with the reduced dimension. \mathbf{f}' is then used to stand for the audio clip as follows:

$$\mathbf{f}' = [\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_M]^t = \mathbf{fV}_k, \quad (4.10)$$

where \mathbf{x}'_i ($i = 1, 2, \dots, M$) is a k -dimensional vector. Note that we will call \mathbf{V}_k as the basis of \mathbf{f}' .

4.2.2. Audio Retrieval and Similarity Measurement

In general, audio (multimedia) data searching can be classified into two different strategies: “a-whole-object search”, and “in-object search”. “A-whole-object search” approach searches for data that is globally similar to the query input; on the other

hand, an “in-object search” approach searches for a large piece of data containing a fragment that is similar to the query. A method of using the latter searching strategy can reach the aim of the first searching strategy but not vice versa. Thus, in this chapter, the retrieval is performed based on the latter searching strategy. Based on the feature vector introduced in the previous section, the similar audio clip retrieval will be conducted. Before retrieval, it is important to give a good similarity measure. Here, a distance measure is first proposed to evaluate the similarity between two audio clips. In our experiments, the Euclidean distance worked better than others (e.g., Mahalanobis, covariance, etc.) in the space generated by \mathbf{V}_k .



4.2.2.1 Similarity Measure

For the candidate audio sequence, \mathbf{y}_c with feature vector $\mathbf{f}'_j = [\mathbf{x}'_{j,1}, \mathbf{x}'_{j,2}, \dots, \mathbf{x}'_{j,M}]$ ($j = 1, 2, \dots, l$), where l is the number of the one-second clips in the audio sequence.

That is, \mathbf{y}_c is divided into one-second clips:

$$\mathbf{y}_c = [y_1, y_2, \dots, y_l], \quad (4.11)$$

where y_j has feature vector \mathbf{f}'_j .

For every queried one-second clip, y_q , before computing the distance between y_q and each of the candidate clip y_j , y_q should be projected to the basis of y_j to

get the corresponding feature $\mathbf{f}'_q = [\mathbf{x}'_{q,1}, \mathbf{x}'_{q,2}, \dots, \mathbf{x}'_{q,M}]^t$. Then the distance between one-second clips y_q and y_j is evaluated as follows:

$$Dist_{q,j} = \left(\sum_{i=1}^M |\mathbf{x}'_{q,i} - \mathbf{x}'_{j,i}|^2 \right)^{\frac{1}{2}}, \quad (4.12)$$

where $j = 1, 2, \dots, l$ and $|\mathbf{x}'_{q,i} - \mathbf{x}'_{j,i}|$ stands for the Euclidean distance between two vectors: $\mathbf{x}'_{q,i}$ and $\mathbf{x}'_{j,i}$. Then for all j , sort $Dist_{q,j}$ in an increasing order. For the top g clips, we define their grades, $Gd_{q,j}$, as $g, g-1, g-2, \dots$, and 1, respectively.

The clip with the least distance will have the highest grade and be considered as the most similar one. In addition, $Gd_{q,j}$ of all other clips are defined as zero. Note that

in this chapter, one-second audio clip is taken as the basic distance measurement unit.



4.2.2.2 Retrieval

For a query audio sequence, \mathbf{y}_q , with length p -seconds, it is first divided into p successive one-second clips. That is

$$\mathbf{y}_q = [y_q^1, y_q^2, \dots, y_q^p]. \quad (4.13)$$

Next, for each clip y_q^i ($i = 1, 2, \dots, p$) and a candidate audio sequence \mathbf{y}_c , the similarity measure is first performed and the corresponding grades, $Gd_{q,j}^i$ ($i = 1, 2, \dots, p$ and $j = 1, 2, \dots, l$), are evaluated based on Eq. (4.12). According to these grades, the total grade of the candidate clip y_j ($j = 1, 2, \dots, l$), $Gd_T_{q,j}$, is defined

to be

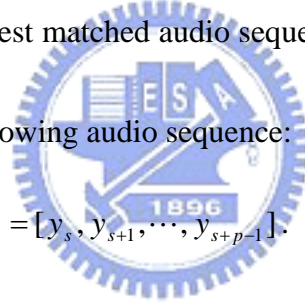
$$Gd_T_{q,j} = \sum_{i=1}^p Gd_{q,j+i-1}^i. \quad (4.14)$$

where $j = 1, 2, \dots, l$. Finally, based on the set of total grades, $Gd_T_{q,j}$ ($j = 1, 2, \dots, l$), sort $Gd_T_{q,j}$ in an increasing order. According to this measure, the top k clips with the higher similarity to the query one can be retrieved. For example, the matching clip with the highest similarity to the query one can be retrieved according to the following criterion:

$$s = \arg \max_j Gd_T_{q,j}, \quad (4.16)$$

where $j = 1, 2, \dots, l$ and the best matched audio sequence, \mathbf{y}_o , in the candidate audio sequence will result in the following audio sequence:

$$\mathbf{y}_o = [y_s, y_{s+1}, \dots, y_{s+p-1}]. \quad (4.17)$$



4.3. EXPERIMENTAL RESULTS

In order to show the efficiency of the proposed method, we have collected a set of 150 musical pieces (50 musical instruments, 100 songs) with total length about three hours and 10000 phrases as the testing database. Care was taken to obtain a wide variation in each type such as varied instruments, different languages (English, Chinese, Japanese, etc.), different singers (male, female, or children), and different

style (jazz, rock, folk, etc.). These audio clips are stored as 16-bit per sample with 44.1 kHz sampling rate in the WAV file format and are used to test the audio retrieval performance. Note that in order to do comparison, the testing database includes the dataset described in [17, 18], and some of clips are taken from MPEG-7 content set [29].

4.3.1 Experiment Results

There are two major factors affecting the performance of the proposed approach, i.e., the number of the basis functions used and the length of the query example. In order to examine the performance of the proposed method, we present two experiments. In the first experiment, for each music object in the database, we use its refrain as the query example to retrieve all repeating phrases similar to this refrain. Therefore, 150 queries are performed. This experiment is presented to examine the quality of the proposed retrieval approach with two above-mentioned major factors. As for the second experiment, for each song, there will have two versions which are sung in different languages or by different persons in the database. We use its refrain in a certain version (e.g. the Chinese version) as the query example to retrieve all repeating phrases similar to this refrain in other version (e.g. the English version). This experiment is presented to examine the robust of the proposed retrieval approach.

In this chapter, the performance is evaluated by the precision rates (P_r) and the recall rates (R_e) [30]. Note that the recall rate, R_e , and the precision rate, P_r , are defined as

$$R_e = \frac{N}{T} \quad \text{and} \quad P_r = \frac{N}{K}, \quad (4.18)$$

where N is the number of relevant items retrieved (i.e. correctly retrieved items), T is the total number of relevant items (i.e. correctly retrieved items and the relevant items that have not been retrieved) and K is the total number of the retrieved items.

The recall rate is typically used in conjunction with the precision rate, which measures the fraction of the retrieved patterns that is relevant. The precision and recall rate can often be traded-off. That is one can achieve high precision rate and low recall rate or the other way round.



Tables 4.1 and 4.2 show the results of two experiments presented in this chapter.

TABLE 4.1
THE AVERAGE RECALL RATES OF THE FIRST EXPERIMENT

Basic Function Numbers	Query Sample Length		
	One second	Two seconds	Three seconds
5	29%	71%	74%
10	31%	75%	75%
15	40%	98%	98%

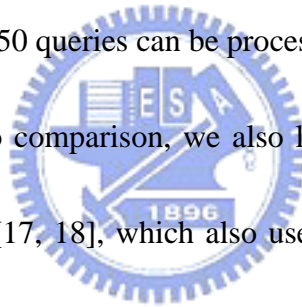
TABLE 4.2
THE AVERAGE RECALL RATES OF THE SECOND EXPERIMENT

Basic Function	Query Sample Length		
Numbers	One second	Two seconds	Three seconds
5	31%	71%	72%
10	31%	71%	74%
15	38%	94%	94%

In our experiments, the number of retrieved patterns was adjusted to the number of relevant patterns, so the precision rate and recall rate are the same. From Table 4.1, we can see that the above-mentioned two factors affect the performance of the proposed approach. The more basis functions are used, the higher the recall rate will be. And the longer length of the query sample is used, the higher the recall rate will be. Based on the first experiment, we can see that it is best to perform retrieval using 15 basis functions and two-second length of query sample. From Table 4.2, we can also see the same phenomena as Table 4.1 except for the lower recall rate.

Besides, by examining the occurrence of missing in the experiments based on human judgement as the ground truth, we found two major factors. First, for the first experiment, we find that some errors occur in those searched clips containing a

transition, which is made due to that we simply segment an audio object into several one-second clips uniformly against pre-dividing the audio object into sequences of audio phrases. As a matter of fact, this kind of errors can be reduced by increasing the length of query sequence (i.e., clip number) to get more related information or performing the pre-dividing for the audio phrases. Secondly, we find that some errors occur due to that the refrains of some songs are performed at different tempo. From these tables, we can see that the proposed retrieval approach for music data can achieve an over 96% accuracy rate. The experiments are carried out on a Pentium II 400 PC/Windows 2000. The 150 queries can be processed in less than five seconds for 10000 phrases. In order to do comparison, we also like to cite the efficiency of the existing system described in [17, 18], which also uses similar database to ours. The authors reported that their accuracy rates are more than 90%.



4.4 SUMMARY

Digital audio signals, especially for music are an important type of media. However, few works were focused on the music databases. In this chapter, we have presented a new method for content-based music retrieval to retrieve perceptually similar music pieces in audio documents. In the proposed method, based on the Gabor

wavelet filters, the extracted perceptual features are general enough to meet the human auditory system. An accurate retrieval rate higher than 96% was achieved. Furthermore, the complexity is low due to the easy computing of audio features, and this makes online processing possible.

There are several related tasks to be conducted in the future. First, we will work on the other type of audio source such as sound effects and the compression domain. Second, we will work on developing an automatic segmentation technique to divide the musical objects into sequences of phrase.

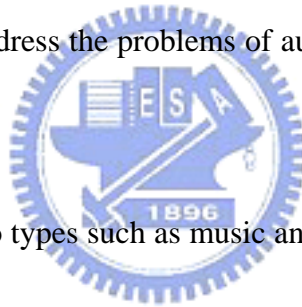


CHAPTER 5

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

5.1 Conclusions

Rapid increase in the amount of audio data demands for an efficient method to automatically analysis audio signal based on its content. In this dissertation, we have presented three methods to address the problems of audio segmentation, classification and content-based retrieval.



Besides the general audio types such as music and speech tested in existing work, in this dissertation, we have taken hybrid-type sounds (speech with music background, speech with environmental noise background, and song) into account. First, we have proposed a hierarchical audio classification method 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. An accurate classification rate higher than 96% was achieved. The experimental results indicate that the extracted audio features are quite robust.

We also propose a classification-based audio segmentation method based on Gabor wavelets. The proposed method provides two classifiers, one is for speech and music (called two-way); the other is for five classes (called five-way) that are 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 Fisher Linear Discriminator to obtain features with the highest discriminative ability. Based on the classification results, a merging algorithm is provided to divide an audio stream into some segments of different classes to achieve segmentation. Experimental results show that the proposed method can achieve over 98% accuracy rate for speech and music discrimination, and more than 95% for a five-way discrimination. By checking the class types of adjacent clips, we also can identify more than 95% audio scene breaks in audio sequence.

Two important and distinguishing features compared with previous work in the above two proposed schemes are the complexity and running time. Although the proposed schemes 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. Thus, the proposed methods can be widely applied to many audiovisual analysis applications such as content-based video retrieval.

Finally, we have presented a new method for content-based music retrieval to

retrieve perceptually similar music pieces in audio documents. It is based on the QBE paradigm and allows the user to select a reference passage within an audio file and retrieve perceptually similar passages such as repeating phrases within a music piece, similar music clips in a database or one song sung by different persons or in different languages. First, an audio stream is divided into clips and the frame-based features of each clip are extracted based on the Gabor wavelet filters. Then, a similarity measuring technique is provided to perform pattern matching on the resulting sequences of feature vectors. The experimental results demonstrate the capability of the proposed audio features for characterizing the perceptual content of an audio sequence.



5.2 Future Research Directions

Content-based audio analysis is still a new area that is not well explored. There are some possible future research directions. For example, in audio classification and segmentation, we will work on the other type of audio source such as sound effects and the compression domain. In the content-based audio retrieval, we will emphasize in query by humming (QBH).

REFERENCES

- [1] S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic audio content analysis," in *Proc. ACM Multimedia'96*, Boston, MA, April 1996, pp. 21-30.
- [2] J. Foote, "An overview of audio information retrieval," *ACM Multimedia Systems*, vol. 7, no. 1, pp. 2-11, January 1999 .
- [3] E. Scherier and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'97*, Munich, Germany, April 1997, pp. 1331-1334.
- [4] S. Rossignol, X. Rodet, and J. Soumagne et al., "Feature extraction and temporal segmentation of acoustic signals," in *Proc. ICMC 98*, Ann Arbor, Michigan, 1998, pp. 199-202.
- [5] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'96*, vol. 2, Atlanta, GA, May 1996, pp. 993-996.
- [6] I. Fujinaga, "Machine recognition of timbre using steady-state tone of acoustic instruments," in *Proc. ICMC 98*, Ann Arbor, Michigan, 1998, pp. 207-210.
- [7] L. Wyse and S. Smoliar, "Toward content-based audio indexing and retrieval and a new speaker discrimination technique," in *Proc. ICJAI'95*, Singapore, December 1995.
- [8] D. Kimber and L. D. Wilcox, "Acoustic segmentation for audio browsers," in *Proc. Interface Conf.*, Sydney, Australia, July 1996.
- [9] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia Mag.*, vol. 3, no. 3, pp. 27-36, Fall 1996.

- [10] L. Guojun and T. Hankinson, "A technique towards automatic audio classification and retrieval," in *Proc. Int. Conf. Signal Processing'98*, vol. 2, 1998, pp. 1142-1145.
- [11] 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*, Seattle, May 1998, pp. 3741-3744.
- [12] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533-544, April 2001.
- [13] 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.
- [14] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 441-457, May 2001.
- [15] G. Smith, H. Murase, and H. Kashino, "Quick audio retrieval using active search," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'98*, Seattle, WA, May 1998, pp.3777-3780.
- [16] J. Foote, "Content-based retrieval of music and audio," in *Proc. SPIE, Multimedia Storage and Archiving systems II*, Vol.3229, 1997, pp. 138-147.
- [17] T. Zhang and C.-C. J. Kuo, "Content-based classification and retrieval of audio," in *Proc. SPIE, Conf. Advanced Signal Processing Algorithm, Architectures, and Implementations VIII*, vol.3461, San Diego, July 1998.
- [18] A. Ghias, J. Logan, D. Chamberlin, and B. Smith, "Query by humming: musical information retrieval in an audio database," in *Proc. Int. Conf. ACM Multimedia*, 1995, pp. 213-236.

- [19] G. Tzanetakis and P. Cook, "Audio information retrieval (AIR) tools," In *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, 2000.
- [20] K. Martin, E. Scheirer, and B. Vercoe, "Musical content analysis through models of audition," In *Proc. ACM Multimedia Workshop on Content-Based Processing of Music*, Bristol, UK, 1998.
- [21] C. Spevak and E. Favreau, "Soundspotter-a prototype system for contest-based audio retrieval," in *Proc. Int. Conf. Digital Audio Effects*, September 2002, pp. 27-32.
- [22] G. Tzanetakis, *Manipulation, Analysis and Retrieval System for Audio Signals*. Ph.D. thesis, Princeton University, 2002.
- [23] C. Yang, "MACS: music database retrieval based on spectral similarity," In *IEEE Workshop on Applications of Signal Processing*, 2001.
- [24] C. Yang, *Music database retrieval based on spectral similarity*. Stanford University Database Group Technical Report 2001-14, 2001.
- [25] S.-T. Bow, *Pattern Recognition and Image Preprocessing*. Marcel Dekker, 1992.
- [26] MPEG Requirements Group, "Description of MPEG-7 content set," Doc. ISO/MPEG N2467, MPEG Atlantic City Meeting, October 1998.
- [27] D. Gabor, "Theory of communication," *Journal of the Institute for Electrical Engineers*, vol. 93, pp. 429-439, 1946.
- [28] B. S. Manjunath and W.Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 173-188, 1992.
- [29] S. Qian and D. Chen, *Joint Time-Frequency Analysis Methods and Applications*. Upper Saddle River, NJ: Prentice-Hall, 1966.
- [30] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*. Springer, 1990.
- [31] F. David Rosenthal, *Computational Auditory Scene Analysis*. Lawrence Erlbaum

Associates, Inc., 1998.

- [32] J. Smith M. and X. Serra, “An analysis/resynthesis program for non-harmonic sounds based on a sinusoidal representation,” in *Proc. ICMC 87*, Ann Arbor, Michigan, 1987, pp. 290ff.
- [33] N. Peter Belhumeur, and David J. Kriegman, “Eigenfaces vs. fishfaces: recognition using class specific linear projection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, July 1997.
- [34] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley, 2002.
- [35] M. Casey, “MPEG-7 sound recognition tools,” *IEEE Transactions on Circuits and Systems Video Technology*, special issue on MPEG-7, IEEE, vol. 11, no. 6, pp. 737-747, 2001.
- [36] ISO/IEC JTC1/SC29/WG11 Moving Pictures Expert Group, “Information technology–multimedia content description interface – part 4: Audio. Committee Draft 15938-4,” ISO/IEC, 2000.
- [37] ISO/IEC JTC1/SC29/WG11 Moving Pictures Expert Group, “Introduction to MPEG-7,” available from <http://www.cselt.it/mpeg>.
- [38] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.
- [39] A. J. Willis and L. Myers, “A cost-effective fingerprint recognition system for use with low-quality prints and damaged fingertips” *Pattern Recognition*, Vol. 34, No. 2, pp. 255-270, February 2001.

PUBLICATION LIST

We summarize the publication status of the proposed methods and our research status in the following.

- (1) Ruei-Shiang Lin and Ling-Hwei Chen, “**A New Approach for Classification of Generic Audio Data,**” accepted by International Journal of Pattern Recognition and Artificial Intelligence.
- (2) Ruei-Shiang Lin and Ling-Hwei Chen, “**A New Approach for Audio Classification and Segmentation Using Gabor Wavelet Filtering and Fisher Linear Discriminator,**” accepted by International Journal of Pattern Recognition and Artificial Intelligence.
- (3) Ruei-Shiang Lin and Ling-Hwei Chen, “**Content-based Retrieval of Audio Based on Gabor Wavelet Filtering,**” accepted by International Journal of Pattern Recognition and Artificial Intelligence.

