Content-based audio classification and retrieval

Liu, Ming Chun

2005

Liu, M. C. (2005). Content-based audio classification and retrieval. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/4738>

<https://doi.org/10.32657/10356/4738>

Nanyang Technological University

Downloaded on 26 Apr 2023 21:20:02 SGT

Content-based Audio Classification and Retrieval

Liu Ming Chun

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2005

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

.....

Date

.....

Liu Ming Chun

*To my family and Xiaoyan,
for their encouragement and love.*

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Chunru Wan, for his invaluable support and guidance throughout my candidature of PhD period and the course of this dissertation. His extensive knowledge, enlightened direction, and continuous encouragement make my thesis work smoothly and positively.

Secondly, I am benefited very much from Dr. Lipo Wang's profound thought and influence in academic research.

Thirdly, I'd like to thank the anonymous reviewers of my thesis for their valuable comments and suggestions.

Last but not the least, I am very grateful to my University, Nanyang Technological University, for its facilities and financial supports.

In particular, I want to thank my family and my wife Zhu Xiaoyan. Without their persistent emotional support, the achievement of my thesis is impossible.

Summary

With the development of Internet and multimedia technologies, the demands for search and retrieval multimedia contents from online databases are expanding. On the other hand, the existing web search engines are designed for text documents. However, these search engines are not efficient when adopted for searching multimedia files, not only because the manual annotation of multimedia data is a tedious task, but also because such annotation may not be possible due to the huge data repositories and ambiguous nature of multimedia when described in texts. Therefore, content-based technology has emerged to provide an alternative solution for retrieval of multimedia database. Compared with content-based retrieval of image and video, less work has been done on audio, which is partly due to the difficulties involved in representing and classification of non-speech audio. Until recent years, audio, as an important media, has gained more and more attention.

In this thesis, a comprehensive literature survey of content-based audio signal processing has been carried out. Then, we focus on the content-based audio classification and retrieval. The purpose of this research is to build an efficient content-based audio classification and retrieval system for a variety of applications. The procedure of content-based audio retrieval system normally includes several sequential steps such as audio feature extraction and selection, audio segmentation, classification and indexing, user query processing and so on.

Firstly, audio features from time, frequency and coefficient domains have been extracted and normalized automatically. Then, a Sequential Forward Selection (SFS) scheme together with several representative statistical and neural network classifiers is used to select a small set of right features to achieve higher

classification accuracies in different applications. The merits of this approach have been demonstrated in various experiments of audio classification and musical instrumental sounds classification.

When the audios are classified and indexed, it makes retrieval easier. A Probabilistic Neural Network (PNN) combined with distance measurement scheme is chosen for hierarchical retrieval of the audio databases. The relevance feedback techniques to improve retrieval performance by taking advantage of user interaction are also discussed.

The fuzzy logic technique has been found useful in both audio classification and retrieval. A hierarchical Fuzzy Inference System (FIS) is built and utilized to form a “fuzzy tree” as an audio directory for coarse-to-detailed audio browse and search. In case of enhancing the searching ability, the fuzzy logic is introduced in Boolean query as well.

In addition, a novel method of audio-visual based story segmentation for news video is presented. By using this method, news video can be segmented into stories based on anchorperson with simple computation.

Finally, we conclude that the content-based audio classification and retrieval techniques can be applied in the fast growing Internet, multimedia digital libraries, MPEG-7 and MPEG-21 standards. Recommendation for the future work includes further studies in the existing areas especially the audio feature extraction for high level audio understanding. Several relatively new fields such as audio ontology creation can be investigated and integrated into the next generation machine-processable semantic web or multimedia digital libraries.

Contents

Acknowledgements	i
Summary	ii
List of Figures	ix
List of Tables	xiii
1 Introduction	1
1.1 Motivation	2
1.2 Objectives	3
1.3 Major Contributions of the Thesis	4
1.4 Organization of the Thesis	6
2 Literature Survey	7
2.1 Spoken Document Retrieval and Indexing	8
2.2 Music Indexing, Music Information Retrieval and Content Understanding	10
2.3 Audio Content Analysis	13

2.4	Content-based Audio Segmentation and Annotation	14
2.5	Content-based Audio Classification and Indexing	16
2.6	Content-based Audio Retrieval	19
2.7	Content-based Audio Aided Video Processing	22
2.8	Audio Ontology Development	24
2.9	Applications	27
2.10	Discussions	30
3	Content-based Audio Classification	34
3.1	The Databases	34
3.2	Audio Feature Extraction, Normalization and Selection	36
3.2.1	Time Domain Features	41
3.2.2	Frequency Domain Features	45
3.2.3	Coefficient Domain Features	52
3.2.4	Feature Normalization	54
3.2.5	Feature Selection	54
3.3	Major Approaches for General Audio Classification	56
3.3.1	Heuristic Rule-based Approach	57
3.3.2	Statistical Approach	58
3.3.2.1	Nearest Neighbor Classification (NN)	59
3.3.2.2	k-Nearest Neighbor Classification (k-NN)	60
3.3.2.3	Gaussian Mixture Model (GMM)	61
3.3.3	Neural Network Approach	63

<i>Contents</i>	vi
3.3.4 Experimental Results	65
3.4 Automatic Classification of Musical Instrument Sounds	73
3.4.1 Introduction	73
3.4.2 The Music Database	74
3.4.3 Feature Extraction and Selection	74
3.4.4 Experiments	77
3.4.5 Summary of Musical Instrument Classification	77
3.5 Discussion	78
4 Content-Based Audio Retrieval	82
4.1 Introduction	83
4.2 Query-by-Example	84
4.3 Probabilistic Neural Network for Content-based Audio Retrieval	85
4.3.1 Performance Measurement	86
4.3.2 Experimental Results	87
4.4 Content-based Audio Retrieval With Relevance Feedback	93
4.4.1 Relevance Feedback in the Related Research Field	93
4.4.2 Relevance Feedback for Audio Retrieval	94
4.4.3 Relevance Feedback Algorithm I	95
4.4.4 Relevance Feedback Algorithm II	98
4.4.5 Query Refining and Other Issues	99
4.4.6 Experimental Results	100
4.5 Discussions	105

5	Audio Classification and Retrieval Using a Fuzzy Inference System	106
5.1	Introduction	107
5.2	Fuzzy Inference System	108
5.2.1	Selecting Features as Inputs	109
5.2.2	Membership Function and Rule Construction	110
5.2.3	Tuning the FIS	118
5.3	FIS Classification and Its Application in Audio Retrieval	119
5.3.1	FIS Classification Results	119
5.3.2	Fuzzy-tree for Audio Retrieval	120
5.4	Boolean Query Using Fuzzy Logic	121
5.4.1	The Fuzzy Logic and Boolean Query	121
5.4.2	Similarity Measurement and Membership Function	122
5.4.3	The Fuzzy Expert System for Boolean Query	123
5.5	Boolean Query Experiments	125
5.5.1	Experiments on AND Boolean Queries from the Same Class	125
5.5.2	Experiments on AND Boolean Queries from Different Classes	126
5.5.3	Experiments on OR Boolean Queries	127
5.5.4	Experiments on Mixed Boolean Queries	128
5.6	Discussions	129
6	Audio-visual Anchorperson-based News Video Story Segmentation	132
6.1	Audio-visual Based News Video Story Segmentation	133

<i>Contents</i>	viii
6.2 News Video Segmentation Procedure	134
6.2.1 Overview of the Process	134
6.2.2 Audio Segmentation Based on Spectral Flux	135
6.2.3 Speech/non-speech Discrimination	137
6.2.4 Anchorperson Detection	138
6.3 Experiments	145
6.3.1 Experiments Setup	145
6.3.2 Experimental Results	146
6.4 Discussions	146
7 Conclusions and Recommendations	149
7.1 Conclusions	149
7.2 Recommendations for Further Research	151
Author's Publications	157
Bibliography	159

List of Figures

2.1	The content-based music information processing.	10
2.2	A hierarchical audio classification.	17
2.3	The content-based audio retrieval.	20
2.4	A female speech ontology example in the audio domain.	26
3.1	A clip of female speech “university”.	37
3.2	A clip from the song “beauty and beast”.	38
3.3	A clip of the sound of heavy rain.	38
3.4	The power spectrum of a clip of female speech “university”. . .	39
3.5	Speech ZCR.	43
3.6	Music ZCR.	43
3.7	Sound ZCR.	44
3.8	The trajectory of pitch from an audio melody.	47
3.9	The LP spectrum and Fourier spectrum of a speech waveform. .	49
3.10	The male speech utterance of “caves”.	49
3.11	Frequency trajectories of the first two formants of a male speech.	50
3.12	Amplitude trajectories of the first two formants of a male speech.	50

3.13	The spectrogram of a female speech “university”.	52
3.14	The procedure of MFCC calculation.	53
3.15	Feature correlation matrix.	56
3.16	The correlation coefficients ρ of the feature RMS.	57
3.17	A rule-based audio classification procedure.	59
3.18	The PNN network architecture (cited from MathWorks [1]). . .	64
3.19	Performance of audio classification experiment 1.	67
3.20	Performance of audio classification experiment 2.	67
3.21	Performance of audio classification experiment 3.	68
3.22	Performance of audio classification experiment 4.	68
3.23	The audio classification accuracy versus feature dimension. . . .	79
4.1	The procedure of content-based audio retrieval.	84
4.2	One typical precision-recall (P-R) curve of audio retrieval using database I.	89
4.3	One typical precision-recall (P-R) curve of audio retrieval using database II.	89
4.4	The mean precision-recall (P-R) curve of audio retrieval using database I.	90
4.5	The mean precision-recall (P-R) curve of audio retrieval using database II.	90
4.6	The GUI of the audio retrieval system.	96
4.7	The difference in average-precision of retrieval performance with total files considered.	103

4.8	The difference in average-precision of retrieval performance with Top T files considered (T=15).	103
5.1	The hierarchical fuzzy tree for audio retrieval.	109
5.2	The feature distribution of mean pitch for female and male.	112
5.3	The simulated feature distribution of mean pitch for female and male using Gaussian membership function.	112
5.4	The feature distribution of mean pitch salience ratio for female and male.	113
5.5	The simulated feature distribution of mean pitch salience ratio for female and male using Gaussian membership function.	113
5.6	The Gaussian membership function simulating the output for female and male.	114
5.7	The FIS input-output diagram for female and male classifier.	114
5.8	The feature distribution of mean pitch salience ratio for telephone ring and other sounds.	115
5.9	The simulated feature distribution of pitch salience ratio for telephone ring and other sounds using Gaussian membership function.	115
5.10	The Gaussian membership function simulating the output for telephone ring and other sounds.	116
5.11	The FIS input-output diagram for telephone ring classifier.	116
5.12	Precision-recall curve of AND Boolean query.	128
6.1	Overview of the story segmentation process.	135
6.2	The Total Spectral Flux (TSF) ratio of part of audio streams.	137
6.3	The anchorperson sample frames.	140

6.4	The combined color histograms of female anchorperson frame. . .	140
6.5	The combined color histograms of male anchorperson frame. . .	141
6.6	The combined color histograms of both female and male anchor- person frame.	141
6.7	The non-anchorperson sample frames.	142
6.8	The combined color histograms of logo frame.	142
6.9	The combined color histograms of field report frame.	143
6.10	The combined color histograms of advertisement frame.	143
7.1	The ‘digital item’ structure.	154
7.2	The framework for audio retrieval using MPEG-21 concept “au- dio item” with ontology.	155

List of Tables

2.1	The major existing approaches for content-based audio signal processing.	31
2.2	The emerging fields of content-based audio processing in chronological order.	32
2.3	The frequently cited researchers and their research works related to content-based audio processing.	33
2.4	The challenging fields in audio signal processing.	33
3.1	The structure of database I.	35
3.2	The structure of database II.	36
3.3	The mean file size and range of file sizes for both databases. . .	37
3.4	The structure of extracted features.	40
3.5	First two formant-frequencies of man, woman, and child during speaking and singing.	48
3.6	The overall and individual classification accuracy of experiment 1.	69
3.7	The overall and individual classification accuracy of experiment 2.	69
3.8	The overall and individual classification accuracy of experiment 3.	70
3.9	The overall and individual classification accuracy of experiment 4.	71

3.10	The musical instrument collection.	75
3.11	The descriptions of audio features for music instrument classification.	76
3.12	The highest audio classification accuracies achieved based on different feature sets.	78
4.1	Experimental result of audio retrieval on database I.	91
4.2	Experimental result of audio retrieval on database II.	92
4.3	The overall average-precision of the relevance feedback.	102
4.4	The overall average-precision of relevance feedback based on Top T (T=15) files considered.	102
4.5	The comparison with Musclefish LLC in audio retrieval.	104
5.1	Fuzzy classifiers and their input features.	110
5.2	The detailed methods used in fuzzy inference system.	111
5.3	The performance of Fuzzy Inference System (FIS) for audio classification.	120
5.4	The ‘Titanic’ audio database for Boolean query.	126
5.5	The retrieval result of AND Boolean queries from same class. . .	127
5.6	The retrieval result of AND Boolean queries from different classes.	129
5.7	The retrieval result of OR Boolean queries.	130
5.8	The retrieval result of AND, OR mixed Boolean queries.	131
6.1	The manual segmentation results of the news video.	147
6.2	The automatic story segmentation results of the news video. . .	147

Chapter 1

Introduction

With the rapid development of multimedia technology and wide spread of wide-band Internet, users can access much more multimedia contents through their networks. However, unlike text documents, most of these audio-visual documents are not well organized and structured for machine processing. Normally, they can only be accessed by external properties, such as their file names, authors, lengths, publishers and formats, rather than directly by their intrinsic contents, such as the genres of a clip of music audio, the scenes of a segment of movie video and so on.

Although the content characteristics can be annotated or indexed manually, this kind of job is tedious and time-consuming. Also due to the annotation, some emotional and environmental information can be lost. For example, listening to speech can impart much more information than reading from the transcription. In addition, because of the ambiguity of the nature of multimedia contents, it is sometimes not easy to describe them precisely using words. Thus, it may cause further problems during searching and retrieval.

To overcome this problem, research has been carried out to index the multimedia contents automatically. In this dissertation, we focus on the automatic classification, segmentation and retrieval of audio together with audio-aided video processing.

1.1 Motivation

With the growth of the Internet and the advance of multimedia technologies, the demand for search and retrieval multimedia content from large databases are expanding. Although there are powerful text search engines, such as Google, which are frequently accessed by users to search for the text webpages of interest, their multimedia search ability is very limited or problematic. Mostly, the audiovisual documents are indexed by their associated external properties like titles or surrounding text descriptions. In this way, the resulting multimedia data retrieval turns to a conventional text search problem again.

However, users could benefit from the ability to have direct access to those multimedia content, which contain rich information but couldn't be precisely annotated by words. On the other hand, the existing web search engines are designed for retrieval of text documents. These web search engines are not efficient for searching multimedia files, not only because the manual annotation of either image, audio or video files is a tedious task, but also because such annotation may not be possible due to the huge data sets, or can be misleading because of the ambiguous nature of multimedia files when described in texts. Therefore, automatic content-based technology should be adopted in designing multimedia searching engines.

Nowadays, available audio collections are rapidly increasing from the fast growing Internet and digitized libraries. Determining how to effectively classify and retrieve such a huge audio database is a challenging task. A complete design of a content-based audio retrieval system involves many issues such as audio feature extraction and selection, audio representation, classification, segmentation, retrieval and performance evaluation. With various research conducted in the literature, those issues involved in building a comprehensive audio search engine have been investigated. However, the performance can be further improved and new emerging applications, such as audio retrieval in digital libraries, audio description in MPEG-7 and MPEG-21 can be found and explored.

1.2 Objectives

The objectives of this thesis form a coherent study in the context of content-based audio signal processing, including audio classification, audio retrieval, and audio-aided video processing, which can be divided into five parts.

1. To get a full understanding of content-based audio classification, audio retrieval, and audio-aided video processing.
2. To build a prototype content-based audio classification and retrieval system for a variety of applications such as digitized libraries. It aims to be able to classify broad classes of sound accurately and retrieve the audio database effectively and efficiently.
3. To introduce human interaction in audio retrieval for performance improvement using relevance feedback techniques.

4. To find out novel solutions for audio classification and retrieval, such as fuzzy logic approach.
5. To investigate how to utilize the audio contents to aid integrated audio-visual signal processing.

1.3 Major Contributions of the Thesis

In order to realize these objectives, in this thesis, we have made our contributions along the above mentioned five tracks.

1. With many research works already accomplished and closely linked with our topic, a broad literature survey in the context of content-based audio signal processing, including audio classification, audio retrieval, audio-aided video processing, together with spoken document retrieval and music information retrieval is carried out. Thus, a big picture in the context of content-based audio signal processing is painted.
2. A test-bed content-based audio classification and retrieval system has been built. One common database in the literature and a more diversified database built by the author have been collected. There are a variety of audio features extracted from time, frequency and coefficient domains including several new ones identified for different applications. A sequential feature selection method is adopted to select the best small feature set to form a feature vector. For audio classification, we have proposed three statistical classifiers and one neural network classifier to achieve better performance using fewer features as compared to existing classifiers tested

on the common database. In addition, a small set of features have been selected for music instrument classification using statistical classifiers. For audio retrieval, our proposed probabilistic neural network (PNN) hierarchical retrieval scheme can achieve better overall performance than a simple distance method in both the recall-precision relation and average precision.

3. Two relevance feedback techniques have been studied to enhance retrieval performance by taking advantage of human interaction. The first approach is a modified approach originally used in image retrieval with relevant feedback. The second approach is a novel technique based on constrained optimization. This new proposed technique can handle both the positive and negative feedback within the same framework.
4. A novel Fuzzy Inference System (FIS) has been applied for audio classification. Furthermore, the concept of fuzzy logic has also been developed in Boolean query for audio retrieval. The fuzzy logic has been found very helpful in both audio classification and retrieval. For examples, a hierarchical FIS has been designed to classify audios from coarse classes to detailed classes. The so-called hierarchical “fuzzy tree” can be employed for top-down audio search as well. The fuzzy logic introduced for Boolean queries may further enhance the searching abilities.
5. An audio-visual anchorperson-based news video story segmentation is proposed. The benefit of using audio in companion video processing has been demonstrated.

1.4 Organization of the Thesis

The rest of thesis is organized as follows. A literature review of content-based audio signal processing is given in Chapter 2. Then, the audio feature extraction and different approaches of audio classification are described in Chapter 3. The audio features are extracted from time, frequency, and coefficient domains and normalized for sequential forward feature selection. The audio classifiers from statistical and neural network approaches together with statistical music instrument classifiers have been described subsequently in the same chapter.

The audio retrieval using query-by-example and relevance feedback techniques are presented in Chapter 4. In Chapter 5, we reveal the direct connection between fuzzy logic and human judgment in content-based audio processing. The Fuzzy Inference System (FIS) is thus introduced for both hierarchical audio classification and Boolean query.

To utilize audio in its associated multimedia signal processing, the audio-aided video segmentation is illustrated in Chapter 6. Chapter 7 concludes the thesis. Finally, recommendations for future work in content-based audio processing, including one relatively new field of audio ontology development, are discussed.

Chapter 2

Literature Survey

One booming technology today is the Internet, due to its fast growing number of users and rich contents. With huge data storage and broadband networks becoming available, multimedia content like image, video, and audio are fast increasing and accessible. In order to search and index these media effectively, various content-based multimedia retrieval systems have been studied.

Starting from the early 1990s, content-based audio signal processing has raised great interest in the research and industry communities. Although compared with its counterparts, such as image and video, there has been less work done for content-based audio processing, this is partly due to the difficulties involved in representing and classifying non-speech audio. However, as audio is a compulsory part in integrated multimedia scenarios like the MPEG, digital library and entertainment industries, more efforts need to be placed in the audio field for a well-balanced multimedia system or for a full-fledged audio database system alone. On the other hand, some of the existing techniques derived for image and video processing can be utilized for audio with necessary modifica-

tions. Therefore, in recent years, audio, as an important media, has gained more attention. The audio objects being studied include speech, music and general sound effects. The existing research on content-based audio signal processing includes audio content analysis, audio classification and segmentation, audio retrieval and audio-aided video processing, such as video indexing, segmentation, annotation and multimedia integration.

2.1 Spoken Document Retrieval and Indexing

Speech signal is the most well studied audio signal in the literature. With the advances of ASR (automatic speech recognition) and IR (information retrieval) techniques, various spoken document retrieval systems have been developed. The definition of Spoken Document Retrieval (SDR) is automatically indexing a collection of spoken documents with speech recognition techniques and retrieving relevant documents in response to a text/speech query. The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 with its purpose to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. A TREC workshop consists of a set tracks, areas of focus in which particular retrieval tasks are defined. The SDR track in TREC represents a significant body of work in this area [2].

The TREC SDR Track has provided an infrastructure for the development and evaluation of spoken document retrieval technology and a common forum for the exchange of knowledge between the speech recognition and information

retrieval research communities. It has also provided objective, demonstrable proof that the technology can be successfully applied to realistic audio collections. The TREC-6 (1997) SDR track was the first formal evaluation of SDR technology. The TREC-9 (2000) SDR Track implemented an evaluation of retrieval of broadcast news excerpts using a combination of automatic speech recognition and information retrieval technologies [3]. A summary of SDR track was given by Garofolo et al. in [4].

Besides SDR track in TREC, there are related works found in other resources. For examples, the Cambridge university spoken document retrieval system was described by Johnson et al. in [5]. The retrieval performance over a wide range of speech transcription error rates was presented and a number of recognition error metrics that more accurately reflecting the impact of transcription errors on retrieval accuracy were defined and computed. Viswanathan et al. [6] proposed a spoken documents retrieval system utilized both content and speaker information together in retrieval by combining the individual content-based and speaker-based retrieval results. Bai et al. [7], instead of using speech transcription in normal spoken document retrieval system, represented a very-large-vocabulary Mandarin voice message file retrieval system using speech queries. By properly utilizing the monosyllabic structure of the Chinese language, the proposed approach performed the statistical similarity estimation between the speech queries and the voice message files, and executed the complete matching process directly at the phonetic level using syllable based statistical information.

Indexing has been found very beneficial for retrieval, because it makes the search process cost less time and produce more meaningful results. Kurimo [8]

presented a method to provide a useful searchable index for spoken audio documents. The idea was to take advantage of the large size of the database and select the best index terms for each document with the help of the other documents close to it using a semantic vector space determined from the training of self-organizing map. Makhoul et al. [9] described a system integrating the requisite speech and language technologies, called Rough'n'Ready, which indexed speech data, created a structural summarization, and provided tools for browsing the stored data.

2.2 Music Indexing, Music Information Retrieval and Content Understanding

Music is another type of audio being extensively studied. The major music information processing includes music indexing, music retrieval, music understanding and music instrument classification as illustrated in Figure 2.1. The major issue is regarded as music retrieval by string matching using query-by-humming technique.

Figure 2.1: The content-based music information processing.

2.2. Music Indexing, Music Information Retrieval and Content Understanding 11

Usually, the extracted melody strings from original music data are adopted to represent the music itself. Hence, the problem of content-based music retrieval is transformed into a string-matching problem. Tseng [10] articulated an algorithm to extract the melody from MIDI files for retrieval. In that system, text strings were input as queries. The MIDI files were converted into text strings and compared to look for certain patterns in the query strings. Liu et al. [11] extracted thematic feature strings, such as melody strings, rhythm strings, and chord strings, from the original music objects and treated as the metadata to represent their content. A new approximate string-matching algorithm was proposed which provided fault tolerance ability according to the music characteristics.

Various forms of query have been used as input to search for files similar to the query example. Ghias et al. [12] presented a system for querying an audio database (of songs) by humming the tune of a song. Kosugi et al. [13] described a retrieval system that enabled a user to hum part of a melody as a query to obtain the name of a desired song from an audio database. The combined use of two musical features, tone transition and tone distribution, enhanced the accuracy of the retrieval system. These features were represented as high dimension vectors to facilitate the retrieval process. Other query methods like query by music segments [14], query by rhythm [15] can be found in song retrievals.

In order to improve the music retrieval performance, several music index techniques have been proposed. Chou et al. [16] implemented a music database system based on the chord-representation model and the PAT-tree index structure with “unstructured search” ability, where a chord was a combination of three or more notes which sounded together in harmony and a PAT-tree was

2.2. Music Indexing, Music Information Retrieval and Content Understanding 12

a Patricia-like tree constructed over all possible substrings of the chord. Lap and Kao [17] studied a number of issues regarding n-gram indexing of musical features using simulated queries. Yang [18] brought out another prototype method of indexing raw-audio music files in a way that facilitated content-based similarity retrieval. Local peaks in signal power were identified in each audio file and a spectral vector was extracted near each peak. Nearby peaks were selectively grouped together to form “characteristic sequences” which were used as the basis for indexing.

Other related issues include music understanding and segmentation [19], music modelling by timbre [20], music content analysis through models of audition [21], musical segmentation using hidden Markov models [22] and (MPEG layer III) digital music management [23]. One particular area, the music instrument identification, has raised notable interest and is discussed in [24], [25], [26], [27].

Recently, the annual International Symposium on Music Information Retrieval (ISMIR), started from year 2000, has become the first established international forum for those involved in the work on accessing digital musical materials. It reflects the tremendous growth of music-related data available either locally or remotely through networks and the consequent need to search this content and retrieve music and musical information efficiently and effectively. A dozens of works can be found among the conference proceedings to tackle the challenges involved in content-based music signal processing include how to calculate similarity during retrieval [28], [29], how to understand music in the way of human being [30], [31], [31], music index [32], [33], and the creation of music digital library [34], [35]. Some early reviews of music information re-

trieval can also be found there [36], [37]. In this thesis, we have selected a small set of features for music instrument classification using statistical classifiers.

2.3 Audio Content Analysis

The audio content analysis or feature extraction is a fundamental process for content-based audio classification and retrieval. The audio contents can be categorized into two groups. On the one hand, there are measurable acoustic properties from the physics point of view and psychoacoustics properties from the human cognition point of view [38], [39]. Acoustics describes how sound behaves within a given physical context. It includes amplitude, loudness, and so on. Psychoacoustics describes how we perceive sound. The psychoacoustic properties of human cognition are subjective terms perceived by the human auditory system including pitch, timbre, and so on. On the other hand, there are high-dimension audio coefficients such as MFCCs (Mel Frequency Cepstral Coefficients) and LPCs (Linear Predict Coefficients) [40]. After analysis, audio features are extracted and available for further processings such as audio classification, segmentation, retrieval and scene analysis, which are based on different combinations of the features. A theoretic framework and applications of automatic audio content analysis were described by Pfeiffer et al. in [41].

Based on audio content analysis, Zhang and Kuo [42] proposed an approach for automatic segmentation and classification of audiovisual data, where simple audio features including the energy function, the average zero-crossing rate, the fundamental frequency and the spectral peak tracks were extracted to ensure the feasibility of real-time processing. Tzanetakis and Cook [43] described a

new method for temporal segmentation based on audio texture. The method was combined with audio analysis techniques and used for hierarchical classification and annotation. They also addressed calculation of features directly from MPEG audio compressed data [44]. Bregman [45] attempted to integrate the existing research on the perceptual organization of sound by connecting it with the “scene analysis” problem encountered in machine vision. Other related techniques such as utilization of spectral flatness features for audio segmentation and retrieval, unified approach of incorporating general features for acoustic modelling, and structured audio representation can be found in [46], [47] and [48] respectively. In this thesis, the audio feature extraction and selection is discussed for automatic audio classification.

2.4 Content-based Audio Segmentation and Annotation

Audio segmentation and classification are both preprocessing in a practical audio retrieval system. Temporal segmentation is a primitive process which divides long audio files into uniform clips. A segment boundary is set when there is a sudden change detected in any of those analyzed content or features. The difference between segmentation and classification is that segmentation is normally required for heterogeneous audio files and the task is to identify boundaries where abrupt changes of features occur. Generally, segmentation doesn’t necessitate the recognition of the audio types between boundaries and is often associated with manual annotation. On the contrary, classification is conducted on short audio clips with homogeneous characteristic and it is obliged to give

label to the audio clip with one of predefined audio classes.

Foote [49] described a method for automatically locating points of significant change in music or audio by analyzing local self-similarity. Tzanetakis and Cook [50] proposed a general methodology for audio segmentation and implemented a prototype audio browsing and annotation tool based on segmentation combined with classification techniques. A new method for audio source type segmentation based on a perceptually derived audio representation has been introduced by Melih and Gonzalez in [51], [52]. Their approach has the advantages of increased search flexibility for content-based retrieval applications and lower processing overheads.

With increasingly different segmentation strategies proposed, Kemp et al. [53] compared model-based, metric-based and energy-based segmentation methods on the same broadcast news testing data. It was shown that the model-based segmenter achieved a very high level of segment boundary precision, yet the metric-based segmenter performed better in terms of segment boundary recall. Thus, a new hybrid algorithm, combining the advantages of both strategies, was introduced. For real time audio segmentation and indexing, simple audio features were adopted by Zhang and Kuo [54] to ensure on-line processing. Morphological and statistical analysis for temporal curves of those features was performed to show differences among different types of audio. A heuristic rule-based procedure was then developed to segment and classify audio recordings into basic types.

A large number of audio resources are stored in MPEG format (MPEG-1 audio Layer 2, and MP3). In order to do segmentation directly on the MPEG encoded sequences so as to avoid computationally intensive decoding proce-

dures, Santo et al. [55] proposed a multi-expert system for the segmentation of the MPEG audio stream. An investigation on MPEG audio segmentation by evolutionary algorithms was also done for detecting music and speech audio segments by Stefano et al. in [56].

2.5 Content-based Audio Classification and Indexing

It is natural to classify audio into a hierarchical directory like the one shown in Figure 2.2. Firstly, the audio can be categorized into three broad classes which are speech, music and sound. Then, speech can be further classified to male and female speech or voiced and unvoiced speech according to different criteria. Music can be grouped into different genres and sound can be sorted into different environmental sounds, sound effects and so on. However, this audio directory is hardly perfect and complete for all circumstances and it is often application-dependent. Research has been conducted towards automatically building such a tree-structure audio directory into different levels and branches according to applications. This way, audio can be indexed not only by those low-level features introduced previously, but also by this high-level class label with semantic meanings.

On the first level of the directory, one simple yet important task of audio classification is to discriminate speech and music in audio clips. Different processings such as ASR (Automatic Speech Recognition) and string matching can be further applied to the discriminated speech and music segments separately. Lu and Hankinson [57] proposed a procedure for speech/music discrimination using

Figure 2.2: A hierarchical audio classification.

four features including silence ratio, spectral centroid, harmony and pitch. The preliminary experiment using silence ratio showed promising results of classifying audio into speech, music and noise. They further conducted investigations based on silence ratio, the variation of zero-crossing rate and their combination in [58]. Liu and Huang [59] described an approach to discriminate news reports from commercials and music in broadcast news programs based on audio information. Fuzzy membership functions associated with the audio features were introduced to improve the performance of the classifier. Foote [60] presented a classifier that distinguished speech from music and non-vocal sounds using statistics generated by a maximum mutual information (MMI) supervised vector quantizer as a measure of audio similarity. The real-time discrimination of speech/music can be found in [61] and [62].

At the second level of the directory, Qi and Hunt [63] adopted a multi-layer feed-forward network based on hybrid features to perform voiced-unvoiced-silence classification for speech signal. Lambrou et al. [64] conducted a study to distinguish three different musical styles of rock, piano and jazz, using wavelet transform analysis in conjunction with statistical pattern recognition techniques.

For general audio data, Li et al. [65] explored a couple of critical issues in an

audio classification system and classified audio segments into seven categories. The environmental sound as one special kind of general sound has raised increasing attention because it can provide many contextual cues that enable us to recognize important aspects of our surroundings. A simple classification of five pre-defined classes of environmental sounds based on several extracted discriminating features was reported in [66]. Zhang and Kuo [67], [68] utilized a hidden Markov model (HMM) to classify environmental sound into applause, explosion, bird's sound, and so on, in a hierarchical system based on the time-frequency analysis of audio. The classification of one typical kind of sound effect, the ringing sound (e.g. telephone, alarm clock, etc), based on frequency-domain production model was addressed in [69].

Corresponding to increasing audio types explored and their greater availability on the Internet, Zhang and Kuo [70] presented an online audio classification and segmentation system, where audio recordings were classified and segmented into speech, music, several types of environmental sounds and silence based on content analysis. Outlines of further classification of audio into finer types and a query-by-example audio retrieval system on top of the coarse classification were also introduced. Another web audio classification scheme using the technique of Fisher kernels was presented in [71], where audio files could be classified to one of three categories: speech, music and other.

In the case of conducting audio classification in compressed MPEG format, the normal feature extraction method can only be applied after the decompression process, which can be extremely time-consuming. As the MPEG audio becomes one major source of audio available, Nakajima et al. [72] proposed a fast and accurate audio data classification method on the MPEG coded domain

to ease the computation load. In this thesis, statistical, neural network and fuzzy logic approaches are implemented and compared for audio classification task.

2.6 Content-based Audio Retrieval

The content-based audio retrieval remains an essential part of content-based audio processing. Its goal is to find documents from an audio database which satisfy certain requirements regarding the query posted by users. A typical situation is to search for audio sounds similar to the proposed query example based on similarity measurement. A desirable audio search engine would retrieve similar sounds on top of the similarity ranking list while leaving the dissimilar ones at the bottom. The audio retrieval is less mature compared with image and video retrieval, due to the difficulty of extracting appropriate audio features that can clearly distinguish different sound sources. A number of issues in audio retrieval are illustrated in Figure 2.3. The issues with dash line arrows are tackled by the author in detail in this thesis.

An early overview of audio information retrieval was written by Foote in [73]. He reviewed the state-of-the-art in audio information retrieval, and focused on advances in automatic speech recognition, word spotting, speaker and music identification and audio similarity measurement with a view towards making audio less “opaque”.

The real pioneering work for retrieval of general audio database was done by Wold et al. [74], where they claimed that many audio and multimedia applications would benefit from the ability to classify and search for audio based on

Figure 2.3: The content-based audio retrieval.

the characteristics of the audio contents rather than by resorting exclusively to keywords. They built such a prototype audio classification and retrieval system which led the research along this direction [75]. In their system, sounds were reduced to perceptual and acoustical features, which let users search or retrieve sounds by different kinds of query. Li [76] presented a new pattern classification method called the nearest feature line (NFL) for equivalent task. Experiments were carried out based on the same database with a lower error rate achieved.

The speed and accuracy are two major concerns of a retrieval system. Liu and Huang [77] proposed a new parametric distance metric to allow faster identification of the audio segments that were similar to the given query example. In that system, an unsupervised method to automatically index the audio content was also given. Both of these efforts can make the online query perform efficiently. Foote [78] presented a pure data-driven audio retrieval system to retrieve audio documents by acoustic similarity. The similarity measurement was based on statistics derived from a supervised vector quantizer. It was effective for very modest requirements in computation and storage and useful for

retrieval simple sounds as well as complex musical excerpts. Kashino et al. [79] proposed a method to search quickly through broadcast audio data to detect and locate known sounds using reference templates, based on active search algorithm and histogram modelling of zero-crossing features. They further improved their method with higher accuracy [80] and extended the method from exact matching to non-exact matching which was applicable even if the feature fluctuated [81].

Depending on applications, there are different strategies to find documents which are exact matches or partial matches to the query. A technique for an exact match for a given piece of cue-audio was described in [82]. In contrast, an in-clip ‘generalized virtual node (GVN)’ algorithm was proposed for searching data fragments that had similar contents of audio recordings for a few measures of a melody [83].

There exist other techniques examined to increase efficiency of searching and browsing audio database. Zhang and Kuo [84] designed a hierarchical system for audio classification and retrieval based on audio content analysis. Their system consisted of three stages, called coarse-level audio classification and segmentation, fine-level classification, and query-by-example audio retrieval. In contrast to existing unstructured audio representations, Melih and Gonzalez [85] designed a perceptually congruent structured audio representation to support content based retrieval and browsing. They also presented a survey of the pertinent issues about the perceptually congruent structures for audio retrieval [86]. In this thesis, the techniques for audio retrieval including relevance feedback, Boolean query processing using fuzzy logic are investigated.

2.7 Content-based Audio Aided Video Processing

It is believed that the performance and ability of video processing can be well enhanced if its accompanying audio signal is utilized. Audio aided video processing can be found in video classification and segmentation, video scene detection and so on. When audio and visual features are combined and integrated together, high-level semantics could be disclosed based on the associated low-level features.

- **HMM-based Audiovisual Processing:** The Hidden Markov Models (HMMs) are effective tools for modelling time-varying patterns. Liu et al. [87] described a video content classifier based on HMM using audio features. There were eight frame-based features to characterize the low-level audio properties and fourteen clip-based features to characterize the high-level audio properties. An overall accuracy of 84.7% was reported in classification of TV broadcast video into commercials, basketball games, football games, news reports and weather forecast. Subsequently, they used four different methods of integrating audio and visual information for video classification based on HMM [88]. Since these multi-modal features could resolve ambiguities that were presented in a single modality, significant performance improvement was shown. Boreczky and Wilcox [89] described another HMM framework for video segmentation using audio and image features. The features for segmentation include an image-based distance between adjacent video frames, an audio distance based on the acoustic difference in intervals just before and after the frames, and an

estimation of motion between the two frames.

- **Video Scene Segmentation:** In video processing, a shot means a set of image frames with similar background and continuous motion, while a scene consists of one or more consecutive shots which are semantically correlated. A scene break is often associated with a simultaneous change of image, motion, and audio characteristics, while a shot break is only accompanied with changes in image or motion or both. Huang et al. proposed using audio information along with image and motion information to accomplish segmentation at different levels with digitized videos from TV programs [90]. Pfeiffer et al. observed that certain basic attributes such as dialogs, similar settings and continuing sounds are consistent indicators of scenes. Therefore, they developed a scheme for identifying scenes by clustering shots according to these basic attributes in [91]. Another novel audio assisted video segmentation scheme was presented by Jiang et al. [92], in which audio and color information were integrated in video scene extraction.
- **Video Processing with Semantics:** Another objective of introducing audio into video processing is to uncover high-level semantic meanings from the audiovisual low-level features during the process of video segmentation, indexing and parsing. Zhang and Kuo [93] presented an approach for the automatic segmentation, indexing and retrieval of audiovisual data associated with semantic meaning based on audio content analysis. David et al. [94] combined an audio segmentation algorithm and a camera shot-break detection algorithm to produce a semantically meaningful segmentation of audio/video content, appropriate for television news retrieval. For

content-based video indexing, Tsekeridou et al. [95] presented an audio-visual content analysis method, where both auditory and visual information sources were analyzed and their inter-relations and coincidence were taken into consideration to extract high-level semantic information. Thereafter, they fully presented their video parsing and indexing method in [96], where integration of the audio and visual mappings constrained by interaction rules led to more detailed video content descriptions and even partial detection of its context.

- **Others:** In order to achieve object detection/localization and tracking in video streams, video scene segmentation using visual and audio cues was described in [97]. For content-based broadcast news video browsing, a system integrating visual, audio and text analysis was developed in [98].

In this thesis, audio-visual based story segmentation for news video is conducted to show the effect of audio-aided video processing.

2.8 Audio Ontology Development

Presently, growing audio databases are becoming available from diversified web resources. Correspondingly, different application dependent and heuristic content-based audio processing systems have been built up. However, it is almost impossible to make a audio processing system always fit its expanding data and the manual implementations of new database systems from scratch for upcoming applications are very time-consuming. Therefore, the re-usability and interoperability with existing domains become crucial and beneficial. The solution is dependent upon the context of establishment of a “semantic web”—an extended

web of machine-readable information and automated services that extend far beyond current capabilities [99]. The semantic web relies heavily on formal ontologies to structure data for comprehensive and transportable machine understanding.

Ontology is a formal, explicit specification of a shared conceptualization [100]. Some of the reasons to develop an ontology are [101]:

- To share common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

In the literature, various ontologies have been created for different domains. For example, in order to represent semantics of images for image content retrieval, a new representation for image contents using visual ontologies and wavelets were proposed in [102]. The visual ontologies were represented by a set of descriptors, relations, logic operators and functions.

With in-depth manual audio catalog and corresponding automatic audio classification, it is feasible to represent the semantics of audio using an audio ontology because there already exist some common audio semantics with high-level meanings derived from various content-based audio systems. Although it is still unlikely to create a universal acceptable machine-processable audio retrieval

system in the very near future, an audio domain ontology can be constructed to move closer to reach this goal.

An audio ontology is a formal explicit description of concepts in audio domain (classes are sometimes called concepts), properties of each concept describing various features and attributes of the concept (slots are sometimes called roles or properties), and restrictions on slots (facets are sometimes called role restrictions). The audio ontology together with a set of individual instances of classes constitutes an audio knowledge base.

For example, all instances of the class **speech**, and its subclass **female speech**, have a slot **speaker**, whose value "Mary" is an instance of the class **author** as shown in Figure 2.4. All instances of the class **author** have a slot **produces** that refers to all the speeches (instances of the class **speech** and its subclasses) that the speaker produces.

Figure 2.4: A female speech ontology example in the audio domain.

In particular, an audio domain-specific ontology was utilized to improve the accuracy (precision and recall) and communication effectiveness of a database system response to a user information request by Khan and McLeod [103].

The ontology was employed along with user profile information, to automatically select and deliver appropriate information units from a multimedia audio database. We propose a framework for audio retrieval using MPEG-21 concept “audio item” with ontology as part of the future work.

2.9 Applications

The technology of content-based audio signal processing has applications in a number of areas. Example applications can be found in real-time segmentation of audio and video, audio database and file system, audio browser and editor, audio information filtering (such as surveillance and violence detection), entertainment industry, digital libraries and so on.

- **Real Time Segmentation of Audio and Video**

The soundtracks as well as audio recordings can be automatically indexed and segmented by using the content-based audio analysis methods. This is accomplished by extracting the trajectories of audio features like brightness, loudness, pitch, and so on. Some of the segmentation and indexing techniques can be applied at real time by looking at transitions and sudden changes of the features for streaming audio and associated video.

- **Audio Databases and File System**

Any database or file system, designed to work with large numbers of audio files, can benefit from content-based indexing and searching capabilities. These require that audio data be represented or supplemented by a so-called metadata record. Advances in digital storage technology and the wide usage of digital audio compression standards like the MPEG standard

have made the management of large archives of audio material possible. Although the creation of high level semantic metadata or indices in a fully automatic way is yet to come, one prototype audio browsing tool that was able to perform semi-automatic audio segmentation and annotation was described by Tzanetakis and Cook in [104].

- **Audio Browser and Editor**

An audio browser with content-based functionality can allow its users to search for their sounds of interest using either query-by-example or query-by-keyword. Such a commercial audio browser called SoundFisher was implemented by MuscleFish LLC. Although most existing audio editors operate directly on the samples of the audio waveform, a more useful editor would include knowledge of the audio content. Thus, a variety of sound classes appropriate for a particular application domain would be developed. For example, editing a concert recording would be aided by classes of audience applause, solo instruments, loud and soft ensemble playing, and so on. Using these classes, the entire concert recording could be initially segmented into regions and indexed, allowing quick access to each musical piece and subsections thereof.

- **Audio Information Filtering**

In contrast to information retrieval where information need is dynamic and information source is stable, information filtering has the characteristics of stable information need and dynamic information source. The information filtering can be viewed as an application of user modelling techniques to facilitate information seeking in dynamic environments. In summary, the design of information filtering systems can be based on two established

lines of research, information retrieval and user modelling. In the “information filtering” process, the user is assumed to be seeking information which addresses a specific long-term interest. Information filtering systems are typically designed to sort through large volumes of dynamically generated information and present the user with sources of information that are likely to satisfy his or her information requirement.

An example application of audio information filtering is surveillance and violence detection. There are a number of police jurisdictions using microphones and video cameras to continuously survey areas where there is a high incidence of criminal activity or a low tolerance of such activity. This kind of surveillance could be made more efficient and easier to monitor with the ability to detect sounds associated with criminal activity such as gunshot or shout. The special requirement for surveillance is that the identification and classification would be done in real-time.

- **Entertainment Industry**

Violence in movies can have a bad influence on children, which is one criterion for movie rating. Although a computer system will never be able to rate movies in a fully automated fashion, especially for mental violence, it can assist in the process. Movie sequences containing violence scenes could be cut out via such a computer-aided film-rating system which identifies explosion and scream shots. A genre classification system of TV sound signals was proposed with hardware implementation to provide a proper timbre automatically to the listeners [105].

- **Digital libraries**

Audio, image, video are major forms of multimedia objects. With the

recent developments in multimedia storage and networking technologies, much multimedia information is becoming increasingly available in areas such as digital libraries and Internet. Content-based retrieval and the issues of representation, storage and retrieval of multimedia objects in digital libraries were addressed briefly in [106].

2.10 Discussions

The ultimate goal of content-based audio processing is to make the audio manageable in a way similar to text documents, where the user can not only hear it at ease but also “read” it inside out. One vision is to create the TOC (Table of Contents) for an audio file automatically and allow random access to its contents.

The audio objects studied are moving from speech and music to more general audio types such as environment sounds. The major existing approaches for processing of those audio objects are listed in Table 2.1. The research interests are changing from low-level feature based auditory scene analysis, audio classification and segmentation to audio retrieval, audiovisual integration, emergence of high-level audio semantics and audio ontology construction. A series of emerging research activities in audio processings are listed in Table 2.2 in chronological order.

Among various works carried out in the context of content-based audio processing, there are several pioneer breakthroughs and novel methods frequently cited in the literature. Table 2.3 lists some of those outstanding papers together with their corresponding authors. Although a series of issues in audio processing

Table 2.1: The major existing approaches for content-based audio signal processing.

Audio Domain	Available Techniques
Speech	Automatic speech recognition, text information retrieval
Music	String matching, music query expression
General Audio and Audiovisual Signal	Feature extraction, statistical classifier and neural networks, heuristic hierarchical classification, HMM

have been tackled and in-depth analysis has been carried out, challenges exist in spoken document retrieval, music understanding and general audio representation, which need further solutions or special treatment. Table 2.4 lists some of those challenging areas to be probed in more detail for the future work.

Table 2.2: The emerging fields of content-based audio processing in chronological order.

Year	Emerging Research and Application Fields
1994	Auditory scene analysis, primary segmentation of auditory scene
1995	Content-based audio retrieval, music instruments identification, query-by-humming
1996	Audio content analysis, audio segmentation and classification, voice message retrieval
1997	Video scene detection, multimedia retrieval, awareness of environmental sound
1998	Video segmentation using audiovisual features, TV programs genre classification, application in digital library
1999	Audio descriptors and descriptor schemes in MPEG-7, video indexing based on joint audiovisual content analysis, spoken document retrieval
2000-present	Audio semantics and audio ontology

Table 2.3: The frequently cited researchers and their research works related to content-based audio processing.

Researchers	Their Research Works and Publication Time
Wold, E. et al.	Content-based classification, search and retrieval (1995-1996) [74], [75]
Silvia.Pfeiffer et al.	Automatic audio content analysis, scene determination based on audio and video features (1996-1998) [41], [91]
Foote, J.	Similarity measure for automatic audio classification, overview of audio information retrieval, segmentation using a measure of audio novelty, content-based retrieval of audio and music (1997-2000) [49], [60], [73], [78]
Liu, Z. et al.	Scene classification, TV programs classification, content-based indexing and retrieval in audio, video segmentation (1997-2000) [59], [77], [87], [88], [90], [107], [108]
Melih, K and Gonzalez, R	Audio retrieval using perceptually based structures, source segmentation for structured audio (1998-2000) [51], [52], [85], [86]
Zhang, T. and Jay Kuo, C.-C.	Audio content analysis for audiovisual data, content-based audio classification and retrieval, segmentation, annotation (1998-2001) [42], [54], [67], [68], [70], [84], [93]
Tzanetakis, G. and Cook, P.	Audio content analysis, computer-assisted annotation of audio, MPEG audio (1999-2000) [43], [44], [50], [104]

Table 2.4: The challenging fields in audio signal processing.

Research Area	Challenge Fields
Spoken document retrieval	Transcription error reduction, integration of both phonetic and text information
Music understanding	Timbre, melody representation, music instrument classification, music genre classification, non-MIDI music feature extraction
General audio processing	psychoacoustic and semantic audio feature expression, audio scene detection, environmental sound classification, audio ontology

Chapter 3

Content-based Audio Classification

Content-based audio classification is a domain-specific classification or pattern recognition problem. In this chapter, we begin with an introduction of two raw audio data sets: one common database used by a number of researchers in the literature, and another more comprehensive database built by the author. Then, we describe the audio feature extraction algorithms together with feature normalization and selection method. Next, we adopt several statistical classifiers and neural network for general audio classification based on the extracted features. Finally, one special case of music instrument classification is discussed.

3.1 The Databases

Two audio databases are used in our experiments. All files have two labels organized in a hierarchical way: a coarse label from the three major classes:

speech, music, and sound; a fine label from more specific classes.

The first small database is the same as in [74], [75] and [76], which has two hierarchies of 16 classes. Among the 16 classes, there are two from speech (female and male speech), seven from music (percussion, oboe, trombone, cello, tubular-bell, violin-bowed, violin-pizzicato), and seven from other environmental sounds and sound effects.

The second database is a much larger one and is an expansion of the database I, which has three hierarchies of 29 classes in a tree-structure built by the author. The audio files in database II are found using searching by Google for audio files with ‘au’ format and downloaded from various resources from Internet. A brief description of the structures of the two databases are given in Table 3.1 and Table 3.2. For both databases, the mean file size and range of file sizes are shown in Table 3.3.

Table 3.1: The structure of database I.

Class name	Number of files	Class name	Number of files
1.Speech	53	Violin-pizzicato(9)	40
Female(1)	36	3.Sound	62
Male(2)	17	Animal(10)	9
2.Music	299	Bell(11)	7
Trombone(3)	13	Crowds(12)	4
Cello(4)	47	Laughter(13)	7
Oboe(5)	32	Machines(14)	11
Percussion(6)	102	Telephone(15)	17
Tubular-bell(7)	20	Water(16)	7
Violin-bowed(8)	45	Total	414

Table 3.2: The structure of database II.

Class name	Number of files	Class name	Number of files
1.Speech	131	Mammals(14)	174
Female(1)	66	3.2Effect	344
Male(2)	54	Alarm(15)	17
Mixture(3)	11	Beat(16)	42
2.Music	458	Bells(17)	50
2.1Instruments	365	Car(18)	12
Brass(4)	44	Clock(19)	15
Keyboard(5)	6	Explosion(20)	44
Percussion(6)	125	Gun(21)	25
String(7)	139	Machines(22)	110
Synthesis(8)	14	Telephone(23)	29
Woodwind(9)	37	3.3Nature	41
2.2Melody	36	Storm(24)	11
Melody(10)	36	Thunder(25)	5
2.3Song	57	Water(26)	20
Cartoon Song(11)	7	Wind(27)	5
Normal Song(12)	50	3.4People	44
3.Sound	649	Applause(28)	8
3.1Animal	220	Laughter(29)	36
Birds(13)	46	Total	1238

3.2 Audio Feature Extraction, Normalization and Selection

The original audio recordings are sequences of waveform. Some typical waveforms from three major classes of audio, namely speech, music, and environmental sound, are shown in Figures 3.1, 3.2, and 3.3 respectively. The Fourier Transformation (FT) of the waveform is a more useful representation in audio content analysis. The power spectrum of the female speech “university” (8KHz sampling frequency) is shown in Figure 3.4. Only with information on the fre-

Table 3.3: The mean file size and range of file sizes for both databases.

Database	Mean file size	Minimum file size	Maximum file size
I	21kb	2kb	141kb
II	26kb	1kb	160kb

quencies can we distinguish between different sounds: every sound we hear is composed of different frequencies and amplitudes whose changing patterns are characteristic. The duration of such patterns is the first basic piece of information for partitioning the audio track into “single sounds”, which are then classifiable. In order to classify different kinds of audios automatically, more features need to be extracted from the raw audio data.

Figure 3.1: A clip of female speech “university”.

In our databases, all audio files are in ‘au’ format, the sample rate of individual file is 8000Hz. The lengths of the sound files usually range from half second to around ten seconds, in that short period, segmentation is omitted. During feature extraction process, each audio file is divided into frames of 256 samples, with 50% overlap at the two adjacent frames. If the energy of an individual frame is below a predefined threshold, the whole frame is marked as a silence

Figure 3.2: A clip from the song “beauty and beast”.

Figure 3.3: A clip of the sound of heavy rain.

frame and is ignored for further processing. After silence reduction, the audio frames are hamming-windowed. Then, the mean and standard deviation of frame-level characteristics are calculated. The audio features are thus extracted from time, frequency and coefficient domains and combined to form the feature vector to represent the individual audio file.

Time domain features include frame energy, RMS (root mean square), ZCR (zero-crossing ratio), VDR (volume dynamic ratio), total energy and silence ratio. Frequency domain features include frequency centroid, bandwidth, four

Figure 3.4: The power spectrum of a clip of female speech “university”.

sub-band energy ratios, pitch, salience of pitch, spectrogram, first two formant frequencies, and formant amplitudes. The first 13 orders of MFCCs (Mel-Frequency Cepstral Coefficients) and LPCs (Linear Prediction Coefficients) are adopted as coefficient features. A feature vector is attached to each sound file in the databases when extraction is completed. A summary of the features are listed in Table 3.4. After feature extraction, the extracted feature vectors are then normalized and ready to be selected for classification and indexing.

Table 3.4: The structure of extracted features.

Feature number	Descriptions
1-2	Mean and standard deviation of frame energy(E): μ_E, σ_E
3-4	Mean and standard deviation of volume root mean square(RMS): μ_{RMS}, σ_{RMS}
5-6	Mean and standard deviation of zero-crossing ratio (ZCR): μ_{ZCR}, σ_{ZCR}
7	Volume dynamic ratio: VDR
8	total energy: TE
9	silence ratio: SR
10-11	Mean and standard deviation of frequency centroid(ω_c): $\mu_{\omega_c}, \sigma_{\omega_c}$
12-13	Mean and standard deviation of bandwidth(B): μ_B, σ_B
14-21	Means and standard deviations of four sub-band energy ratios($ER^{(1-4)}$): $\mu_{ER}^{(1-4)}, \sigma_{ER}^{(1-4)}$
22-23	Mean and standard deviation of pitch(P): μ_P, σ_P
24-25	Mean and standard deviation of salience of pitch(SP): μ_{SP}, σ_{SP}
26-27	Mean and standard deviation of spectrogram(Spec): $\mu_{Spec}, \sigma_{Spec}$
28-31	Mean and standard deviation of first two formant frequencies(F1,F2): $\mu_{F1}, \mu_{F2}, \sigma_{F1}, \sigma_{F2}$
32-35	Mean and standard deviation of first two formant amplitudes(A1,A2): $\mu_{A1}, \mu_{A2}, \sigma_{A1}, \sigma_{A2}$
36-61	Mean and standard deviation of first 13 orders of $MFCC_{1-13}$ (Mel-Frequency Cepstral Coefficients): $\mu_{MFCC_{1-13}}, \sigma_{MFCC_{1-13}}$
62-87	Mean and standard deviation of first 13 orders of LPC_{1-13} (Linear Prediction Coefficients): $\mu_{LPC_{1-13}}, \sigma_{LPC_{1-13}}$

3.2.1 Time Domain Features

Time domain features include frame energy, RMS (root mean square), ZCR (zero-crossing ratio), VDR (volume dynamic ratio) and silence ratio.

- **Frame Energy:** It provides a convenient representation of the amplitude variation over time. Its changing pattern over time may reveal the rhythm and periodicity nature of the underlying sound. The frame energy of the m th frame of the audio signal is defined as:

$$\mathbf{E}_m = \sum_{n=0}^{N-1} [x_m(n)]^2 \quad (3.1)$$

where $x(n)$ is the discrete time audio signal, $x_m(n)$ is the audio signal of m th frame, m is the frame index, n is sample index from 0 to $N-1$, N is the number of samples in each frame. Throughout the thesis, we have chosen $N = 256$. The mean and standard deviation of the frame-level frame energy over all the frames of interest are used as the features for the audio file, which are denoted by

$$\mu_E = \frac{1}{M} \sum_{m=0}^{M-1} E_m \quad (3.2)$$

$$\sigma_E = \sqrt{\frac{1}{M-1} \sum_{m=0}^{M-1} (E_m - \mu_E)^2} \quad (3.3)$$

where the notations of μ and σ are adopted to denote the mean and standard deviation of individual features, M is the number of frames for the audio file concerned.

- **RMS:** It is the measure of loudness of the frame. This feature is unique to segmentation since changes in loudness are important cues for new sound events. It is given by

$$RMS_m = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x_m^2(n)} \quad (3.4)$$

The RMS_m and E_m have the relation of $N * RMS_m^2 = E_m$. The μ_{RMS} and σ_{RMS} are adopted as features.

- **Zero-Crossing Ratio:** A zero-crossing is said to occur if successive samples have different signs. The zero-crossing ratio is defined by

$$Z_m = \frac{1}{2N} \sum_{n=1}^{N-1} |sgn[x_m(n)] - sgn[x_m(n-1)]| \quad (3.5)$$

where $sgn[x] = \begin{cases} 1 & , \quad x \geq 0 \\ -1 & , \quad x < 0 \end{cases}$. The zero-crossing ratio is the number of the time-domain zero-crossings divided by the total number of samples in a frame. Typical ZCRs from the three major audio classes are shown in Figures 3.5, 3.6, and 3.7. The speech ZCR curve has peaks and troughs from unvoiced and voiced components, respectively. Therefore, the curve has a relatively low and stable baseline with high peaks above it, which results in a large variance and a wide range of amplitudes. In contrast, the ZCR curve of music has a much lower variance. It generally has a stable waveform with a relatively small range of amplitudes. Since environmental audio consists of sounds from various origins, their ZCR curves may have very different properties. Generally speaking, we can classify environmental sounds according to properties of their ZCR curves such as

regularity, periodicity, stability, and the range of amplitudes. The μ_{ZCR} and σ_{ZCR} are adopted as features.

Figure 3.5: Speech ZCR.

Figure 3.6: Music ZCR.

- **VDR:** The difference of maximum and minimum RMS normalized by the maximum RMS of the frame audio signal. It is defined by

$$VDR = \frac{\max_m(RMS_m) - \min_m(RMS_m)}{\max_m(RMS_m)} \quad (3.6)$$

Figure 3.7: Sound ZCR.

The magnitude of an audio clip's VDR is dependent on the type of sound present. Audio clips featuring traffic or crowds contain a fairly constant level of unchanging sound volume for the entire duration of the clip and this leads to a low VDR. Conversely, in audio clips containing gunfire, there are staccato periods of large amplitudes alternating with period of silence between the discharging of the gunfire, and this results in a higher VDR.

- **Total Energy:** Total energy of the whole audio file. It is defined as

$$TE = \sum_{m=0}^{M-1} E_m \quad (3.7)$$

- **Silence Ratio:** It is the ratio of silent frames (determined by a preset threshold) and the entire frames. Here, a frame is said to be silent if the frame RMS is less than 10% of the mean RMS of the file. Thus, the silence

ratio can be calculated by

$$SR = \frac{1}{M} \sum_{m=0}^{M-1} \text{sign}[RMS_m - 10\% \mu_{RMS}] \quad (3.8)$$

$$\text{where } \text{sign}[x] = \begin{cases} 1 & , \quad x < 0 \\ 0 & , \quad x \geq 0 \end{cases}$$

3.2.2 Frequency Domain Features

The frequency features include frequency centroid, bandwidth, four sub-band energy ratios, pitch, salience of pitch, spectrogram, first two formant frequencies, and formant amplitudes.

- **Frequency Centroid (Brightness):** The frequency centroid of the m th frame data is defined as

$$\omega_{c,m} = \frac{\int_0^{\omega_0} \omega |F_m(\omega)|^2 d\omega}{\int_0^{\omega_0} |F_m(\omega)|^2 d\omega} \quad (3.9)$$

where $|F_m(\omega)|^2$ is the power spectrum level at the frequency ω , and ω_0 is the half sampling frequency. The μ_{ω_c} and σ_{ω_c} are selected as features.

- **Bandwidth:** The square root of the power-weighted average of the squared difference between the spectral components and the frequency centroid. It can quantitatively express the range of frequencies over which the power or energy density spectrum is concentrated, which is defined by

$$B_m = \sqrt{\frac{\int_0^{\omega_0} (\omega - \omega_{c,m})^2 |F_m(\omega)|^2 d\omega}{\int_0^{\omega_0} |F_m(\omega)|^2 d\omega}} \quad (3.10)$$

The μ_B and σ_B are chosen as features.

- **Sub-Band Energy Ratio:** The frequency spectrum is divided into 4 sub-bands with intervals $[0, \frac{\omega_0}{8}]$, $[\frac{\omega_0}{8}, \frac{\omega_0}{4}]$, $[\frac{\omega_0}{4}, \frac{\omega_0}{2}]$, $[\frac{\omega_0}{2}, \omega_0]$. The sub-band energy ratio $ER_m^{(j)}$ is measured by

$$ER_m^{(j)} = \frac{P_j}{P} \quad (3.11)$$

$$P_j = \int_{L_j}^{H_j} |F_m(\omega)|^2 d\omega \quad (3.12)$$

$$P = \int_0^{\omega_0} |F_m(\omega)|^2 d\omega \quad (3.13)$$

where L_j and H_j are lower and upper bound of sub-band j as given above, and $j=1,2,3,4$. Sub-band energy ratios, when used together, reveal the distribution of spectral energy over the entire frame. The $\mu_{ER}^{(j)}$ and $\sigma_{ER}^{(j)}$ are used as features.

- **Pitch:** Pitch refers to the fundamental period of a human speech waveform. It is an important parameter in the analysis and synthesis of speech. We compute the pitch by finding the non-zero time lag with the largest autocorrelation energy. Figure 3.8 shows a trajectory of pitch from a clip of audio melody. Let P_m be the pitch of the m th frame. The mean and standard deviation of frame pitches, μ_P, σ_P , are used as features.
- **Salience of Pitch:** The salience of pitch equals the ratio of the first peak value at pitch location and the zerolag value of the autocorrelation function. It is defined as

$$SP_m = \frac{\phi_m(P)}{\phi_m(0)} \quad (3.14)$$

Figure 3.8: The trajectory of pitch from an audio melody.

$$\phi_m(P) = \frac{1}{N} \sum_{n=0}^{N-1} x_m(n)x_m(n-P) \quad (3.15)$$

$$\phi_m(0) = \frac{1}{N} \sum_{n=0}^{N-1} x_m^2(n) \quad (3.16)$$

where $\phi_m(P)$ is the pitch value of the m th frame autocorrelation function, $\phi_m(0)$ is the zerolag value of the m th frame autocorrelation function. The μ_{SP} and σ_{SP} are selected as features.

- **First two Formants:** Formant is caused by resonant cavities in the vocal tract of speaker. A formant, when it occurs in natural speech, is just a peak in the spectrum, but in some synthesized speech used in research, it may be an isolated band of frequencies. It appears that these isolated bands are sufficient to induce the perception of vowels or other voiced sounds. The averaged first two formant-frequencies of man, woman and child during speaking and singing are showed in Table 3.5. Note that the data for child singing is not available.

Here, the formant frequencies are computed by peak-picking of the Linear Predictive (LP) spectrum. The basic assumption behind LP is the

Table 3.5: First two formant-frequencies of man, woman, and child during speaking and singing.

	Man	Woman	Child	Singing Man	Sing Woman
F1	270-730	310-850	370-1030	300-700	400-700
F2	840-2290	920-2790	1060-3200	640-1950	800-2250

correlation between the n th sample and the P previous samples of the target signal. For example, in the m frame, the n th signal sample is represented as a linear combination of its previous P samples, plus a residual representing the prediction error:

$$x_m(n) = \sum_{i=1}^P a_m(i)x_m(n-i) + e_m(n) \quad (3.17)$$

where $a_m(i)$ is the i th order coefficients, $e_m(n)$ is the prediction error. The coefficients of P th order linear predictor are found such that the sum of the squares of the errors is minimized. This predictor can be also looked upon as an all-pole linear filter, which has the transfer function:

$$H(z) = \frac{X_m(z)}{E_m(z)} = \frac{1}{1 - \sum_{i=1}^P a_m(i)z^{-i}} \quad (3.18)$$

The LP spectrum of $H(Z)$ is calculated by taking the inverse of the Fourier transformation of the filter coefficients. It can capture the formants as well as track the harmonic content of the audio clip. Such a LP spectrum and Fourier spectrum of a speech waveform are shown in Figure 3.9. The first and second formants are the most important for audio perception. The typical trajectories of the 1st and 2nd formant-frequencies of a male speech, whose utterance is shown in Figure 3.10, are given in Figure 3.11.

The first two formant-frequencies of $F1_m$ and $F2_m$, which are the two highest peaks in the LP spectrum, are selected. Then, the μ_{F1}, σ_{F1} and μ_{F2}, σ_{F2} are selected as features.

Figure 3.9: The LP spectrum and Fourier spectrum of a speech waveform.

- **The two formant amplitudes (A1, A2):** The amplitudes of the spectrum at the first two formant-frequencies of the same speech segment are shown in Figure 3.12. The μ_{A1}, σ_{A1} and μ_{A2}, σ_{A2} are used as features.

Figure 3.10: The male speech utterance of “caves”.

Figure 3.11: Frequency trajectories of the first two formants of a male speech.

Figure 3.12: Amplitude trajectories of the first two formants of a male speech.

- **Spectrogram:** Spectrogram describes the overall energy distribution at different time and frequency. Let $Spec(u, v)$ denote the spectrogram of the sampled auditory input $x(n)$. It is defined as:

$$Spec(u, v) = \frac{1}{N} \sum_{n=0}^{N-1} x(u\Delta\theta + n) \omega(n) e^{-j \frac{2\pi n v}{N}} \quad (3.19)$$

where $u = 0, 1, \dots, [\frac{L-\Delta\theta}{N-\Delta\theta}]$, $v = 0, 1, \dots, N/2$, L is the signal length, N is the window size and set to 256, $\Delta\theta$ is the number of overlapping samples between successive windows and set to 128, $\omega(n)$ is the hamming window function of size N . Each column of the spectrogram contains an estimate of the short-term, time-localized frequency content of the signal. Then, the mean and standard deviation of the absolute value of the spectrogram matrix elements, $\mu_{Spec}, \sigma_{Spec}$, are computed as features.

Most importantly, much helpful information can be derived from the spectrogram, the three-dimension representation of speech in time, frequency, and intensity (darkness). There are two main kinds of spectrogram, wideband (with a bandwidth of 300-500 Hz) and narrowband (with a bandwidth of 45-50 Hz). Wideband spectrograms allow for excellent time resolution, yet frequency components within a 300-500 Hz bandwidth are not easily distinguished. The narrowband spectrogram is able to pick out each individual harmonic, unlike the wideband spectrogram, but its time resolution is not good enough to isolate each individual cycle of vibration. A wideband spectrogram of the female speech “university” is illustrated in Figure 3.13, which shows periodicity in time with vertical striations.

Figure 3.13: The spectrogram of a female speech “university”.

3.2.3 Coefficient Domain Features

The MFCCs (mel-frequency cepstral coefficients) and LPCs (linear prediction coefficients) coefficients, which are widely used in speech recognition, are also adopted for classification of general sounds.

- **Mel-Frequency Cepstral Coefficients:** The mel scale, a scale of pitches judged by listeners to be equal in distance one from another, is a psychological measure of pitch magnitude. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener’s threshold, with a pitch of 1000 mels. The relationship between f hertz and m mel are defined as below:

$$m = 2595 * \log_{10}(1 + f/700) \quad (3.20)$$

$$f = 700(e^{m/1127.01048} - 1) \quad (3.21)$$

MFCCs have been the dominant features used for speech recognition for some time [40]. MFCC calculation involves the following steps: the signal

is divided into short frames (20ms) and hamming windowed, and then a FFT is taken of each frame and its amplitude spectrum is smoothed by combining Fourier bins into outputs of mel-spaced filterbank and converted to a logarithmic scale to approximately model perceived loudness. Finally a discrete cosine transform is applied to extract principal components and reduce the data to typically 13 components per frame. The procedure is represented in Figure 3.14.

Figure 3.14: The procedure of MFCC calculation.

The filter bank consists of $K = 19$ triangular filters, which have a constant mel-frequency interval and cover the frequency range of 0–4000 Hz. Denoting the output of the filter bank by $S_k(k = 1, 2, \dots, K)$, the MFCCs are calculated by

$$c_n = \sqrt{\frac{2}{K}} \sum_{k=1}^K (\log S_k) \cos [n(k - 0.5)\pi/K] \quad n = 1, 2, \dots, L \quad (3.22)$$

where L is the order of the cepstrum and set to 13. The mean and standard deviation of the frame-level 13 order MFCCs $\mu_{MFCC_{1-13}}, \sigma_{MFCC_{1-13}}$, are used as features.

- **Linear Prediction Coefficients:** As mentioned earlier, the LP coefficients are a short-time measure of the speech signal which describe the

signal as the output of an all-pole filter. This all-pole filter provides a good description of the speech articulators. Thus LP analysis is often used in speech recognition and speech coding systems. The first 13 orders of LPC parameters are calculated every 20ms in this implementation. Again, the mean and standard deviation of the frame-level 13 order LPCs $\mu_{LPC_{1-13}}$, $\sigma_{LPC_{1-13}}$, are adopted as features.

3.2.4 Feature Normalization

Normalization can ensure that contributions of all audio feature elements are adequately represented and prevent one feature from dominating the whole feature vector. Each audio feature is normalized over entire files in the database by subtracting its mean and dividing by its standard deviation. The normalization process can be expressed as

$$N_l = \frac{A_l - \mu_l}{\sigma_l} \quad (3.23)$$

where N_l is normalized values of the l th feature, A_l is the original value of l th feature, μ_l is the mean of the l th feature, σ_l is the standard deviation of l th feature. After normalization, different features have a similar distribution over all files in the database. This normalization process will ensure more accurate results during classification. Then, each audio file is fully represented by its normalized feature vector.

3.2.5 Feature Selection

Simply choosing all the features available often doesn't yield the best performance, because some features give poor separability among different classes and

some are highly correlated. These unfavorable features have a negative effect when added into the feature vector. Therefore, before using the extracted features for further processing, the feature correlation matrix is calculated to find the mutual relation among those features. Let $F(j) = [f_1(j), f_2(j), \dots, f_L(j)]^T$ is the feature vector of the j th file, L is the number of features, N_f is the number of files. The $L \times L$ feature correlation matrix R is defined as

$$R = \frac{1}{N_f} \sum_{j=1}^{N_f} F(j)F^T(j) \quad (3.24)$$

The image expression of feature correlation matrix and the correlation coefficients of feature RMS are shown in Figures 3.15 and 3.16, where dark color means high correlation, light color means low correlation in the matrix. From the matrix, we can clearly see that the standard deviation of MFCC coefficients, with feature indices from 49 to 61, are highly correlated with each other and some features in frequency domain are also highly correlated.

Theoretically, we can use the exhausted combination method to select the best feature vector, but the computation complexity is huge. A sub-optimum method, the sequential forward selection (SFS) method, is adopted to select the best feature set. We firstly select the single feature that results in best classification performance and then add one feature at a time in combination with the already selected features that minimizes the classification error rate. We continue to do this until all the features are selected. The classification performance versus dimension of the best feature set for different classifiers will be discussed in next section.

Figure 3.15: Feature correlation matrix.

3.3 Major Approaches for General Audio Classification

In general, there exist three basic approaches for classification or pattern recognition problems, which are observed from syntactic, statistical and neural network perspectives.

Syntactic pattern recognition is used when the significant information in a pattern is the interrelationships of features or primitives, which facilitate structural description of classification.

Statistical pattern recognition attempts to classify pattern based on a set of extracted features and an underlying statistical model. Given a set of training

Figure 3.16: The correlation coefficients ρ of the feature RMS.

patterns from each class, the objective is to establish decision boundaries in the feature space which separate patterns belonging to different classes.

The neural network approach is a non-algorithmic, black box strategy. Due to the availability of efficient learning algorithms for practitioners to use, it is effective when the problem space is high-dimensional and complex interactions exist between problem variables.

3.3.1 Heuristic Rule-based Approach

Although it is not suitable to use the syntactic approach for audio classification, since it is hard to decompose audio into primitives and infer rules from the available training samples, some hierarchy rule-based classification schemes are

used in the literature. Zhang and Kuo [67] developed a heuristic rule based procedure to classify audio into silence, speech, music, song, environmental sound, speech with the music background, and environmental sound with music background by using some simple features. A procedure, shown in Figure 3.17, towards audio classification based on silence ratio, spectral centroid, harmony and pitch was proposed by Lu and Hankinson in [57], where each audio input to be classified went through a number of classification steps until a decision was made. Therefore, it is workable to organize the audio database in a hierarchical way and allow different treatment in each level.

3.3.2 Statistical Approach

The statistical approach has been most intensively studied and used in practice. It has also been employed successfully to design a number of commercial recognition systems. There exist various approaches in statistical pattern recognition. Basically, there are two kinds of classifiers: non-parametric and parametric classifiers. In this thesis, two simple nonparametric approaches, nearest neighbor classifier (NN) and modified k-Nearest Neighbor (k-NN), and one parametric classifier, Gaussian Mixture Model (GMM), are adopted. The simplest and the most intuitive approach to classifier design is based on the concept of similarity. Patterns that are similar should be assigned to the same class, the most straightforward nearest neighbor rule can be conveniently used as a benchmark for all the other classifiers since it appears to provide a reasonable classification performance in most applications.

Figure 3.17: A rule-based audio classification procedure.

3.3.2.1 Nearest Neighbor Classification (NN)

The nearest-neighbor classifier uses all the given training data to implicitly estimate its probability density function (*pdf*). Given a new case, we can assign it to the group of its nearest neighbor in the training database using an appropriate definition of distance. In the experiments, one half of the data is used for training and the other half for validation. Given a feature vector V in the validation set and i th feature vector $T(i)$ in the training set, we define the

Euclidean distance between them as follows:

$$d_{NN}(i) = \|V - T(i)\|_2 = \sqrt{\sum_{j=1}^L (V_j - T_j(i))^2} \quad (3.25)$$

where V_j and $T_j(i)$ refer to the j th element in the feature vectors V and $T(i)$ respectively, L is the feature dimension. Then, the nearest neighbor of V in the training set, $T(k)$, is found. Finally, we assign the file with feature vector V to the same class $T(k)$ belongs to.

3.3.2.2 k-Nearest Neighbor Classification (k-NN)

A variation of NN is the k-NN. Generally speaking, the k-nearest-neighbor classification algorithm is conducted in two steps:

- **Step 1:** For each feature vector in the validation dataset, we locate the k closest feature vectors (the k nearest neighbors) in the training dataset, measured by an appropriate definition of distance, such as the Euclidean Distance introduced previously.
- **Step 2:** We examine the k nearest neighbors to decide the group that most of them belong to. Then, we assign this group to the feature vector being examined.

The computing time increases as k increases, but the advantage is that higher values of k provide stability so that they reduce vulnerability to noise in the training data.

In our experiments, we introduce a modified version of the k-NN. There are also two steps as given below:

- **Step 1:** For each feature vector in the validation dataset, we first find the k nearest neighbors from each group instead of whole training set, measured by the same Euclidean Distance. Suppose there are G groups. Thus, we locate the $k \times G$ closest feature vectors in the training dataset.
- **Step 2:** We define the group Euclidean Distance as the mean of the k Euclidean distances within the group. Then, we assign the testing feature vector with the group corresponding to the smallest group Euclidean Distance. This modified k-NN method can avoid the problem which arises when there is no majority of the k nearest samples in normal k-NN approach.

3.3.2.3 Gaussian Mixture Model (GMM)

Usually, the pattern classification problem can be reduced to an estimation problem of a probability density function (*pdf*), since the classification can be performed according to the Bayes decision rule if *a posteriori* probability of the input pattern is obtained [109], [110]. The Gaussian mixture model has been proposed as a general model for estimating an unknown probability density function [111]. The virtues of the model lie mainly in its good approximation properties and the variety of parameter estimation algorithms that exist in the literature. The model assumes that the unknown density can be written as a weighted finite sum of Gaussian kernels, with different mixing weights and different means and covariance matrices. A mixture of K Gaussians is:

$$p(\omega|x) = \sum_{k=1}^K \alpha_k G(x; \mu(k), \Sigma(k)) \quad (3.26)$$

$$G(x; \mu(k), \Sigma(k)) = (2\pi)^{-\frac{d}{2}} |\Sigma(k)|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x - \mu(k))^T (\Sigma(k))^{-1} (x - \mu(k))\right] \quad (3.27)$$

where α_k is the mixing parameter satisfying $\sum_{k=1}^K \alpha_k = 1$. $G(x; \mu(k), \Sigma(k))$ is the *pdf* corresponding to the k th Gaussian component. The Gaussian mixture model contains the following adjustable parameters: α_k with the mean $\mu(k)$ ($\mu(k) \in R^d$) and the standard deviation $\Sigma(k)$ ($\Sigma(k)$ is a $d \times d$ positive definite matrix) ($k = 1, \dots, K$) and they are parameterized by θ .

One of the parameter estimation algorithms often used for Gaussian mixture modelling is the expectation-maximization (EM) algorithm. The objective of the EM algorithm is to maximize the likelihood $p(X; \theta)$ of the data $X(X = x_1, x_2, \dots, x_N)$ drawn from an unknown distribution, given the model parameterized by θ . The likelihood $p(X; \theta)$ is defined as below:

$$p(X|\theta) = \prod_{p=1}^N p(x_p|\theta) = \prod_{p=1}^N \sum_{k=1}^K \alpha_k G(x_p; \mu(k), \Sigma(k)) \quad (3.28)$$

The GMM parameters are estimated using the standard iterative Expectation Maximization (EM) algorithm. The resulting updating rules for GMM is as follows:

$$\beta_{pk} = \frac{\alpha_k G(x_p; \mu_k, \Sigma_k)}{\sum_{i=1}^K \alpha_i G(x_p; \mu_i, \Sigma_i)} \quad (3.29)$$

$$\hat{\alpha}_k = \frac{1}{N} \sum_{p=1}^N \beta_{pk} \quad (3.30)$$

$$\hat{\mu}_k = \frac{\sum_{p=1}^N \beta_{pk} x_p}{\sum_{p=1}^N \beta_{pk}} \quad (3.31)$$

$$\hat{\Sigma}_k = \frac{\sum_{p=1}^N \beta_{pk} (x_p - \hat{\alpha}_k)(x_p - \hat{\alpha}_k)^T}{\sum_{p=1}^N \beta_{pk}} \quad (3.32)$$

where β_{pk} is an intermediate variable and $\hat{\theta}_k(\hat{\alpha}_k, \hat{\mu}_k, \hat{\Sigma})_k$ are the updated parameters in the next iteration. The iteration process continues until the parameters converge to particular value or the pre-defined maximum iteration number reaches. The details of the EM algorithm can be found in [112] and [113].

In our experiments, the priors are initialized to equal values summing to one, and the covariances are all the identity matrix (or equivalent). The centres are initialized randomly from a zero mean unit variance Gaussian. Based on the extracted features and their corresponding class label, the GMM is trained for audio classification. The number of Gaussians is defined as 4. The Bayes decision theory determines a specific class if a *posteriori* probability $p(\omega_i|x)$ of the vector belonging to the class i is larger than the probabilities $p(\omega_j|x), j \neq i$ of any other classes.

3.3.3 Neural Network Approach

The main characteristics of neural networks are that they have the ability to learn complex nonlinear input-output relationships using sequential training procedures and adapt themselves to the data. In spite of the seemingly different underlying principles, most of the well-known neural network models are implicitly equivalent or similar to classical statistical pattern recognition methods. Despite these similarities, neural networks offer several advantages, such as a unified approach for classification and flexible procedures for finding good, moderately nonlinear solutions.

A special type of neural network, called a probabilistic neural network (PNN), is utilized in this study. It was first introduced by Specht [114], [115]. A modified probabilistic neural network is found in [116]. The PNN can be treated

as a feed-forward network that implements a Gaussian mixture model [117]. The network uses one Gaussian kernel for each input sample, while the variance of each kernel is known constant and the mixing weights are equal to the reciprocal of the total number of inputs. The design of PNN for classification is straightforward and does not depend on training. A PNN is guaranteed to converge to a Bayesian classifier providing it is given enough training data and has good generalization ability. The network architecture of PNN is illustrated in Figure 3.18.

Figure 3.18: The PNN network architecture (cited from MathWorks [1]).

It is assumed that there are Q input/target vector pairs. Each target vector has K elements. One of these element is one and the rest is zero. Thus, each input vector is associated with one of K classes. The first layer input weights, $\mathbf{IW}^{1,1}$, are set to the transpose of the matrix formed from the Q training pairs, $\mathbf{P}'$. When an input is presented, the $\|\text{dist}\|$ box produces a vector whose elements indicate how close the input is to the vectors of the training set. These

elements are multiplied, element by element, by the bias and sent to the *radbas* transfer function. An input vector close to a training vector will be represented by a number close to one in the output vector $\mathbf{a}^1$. If an input is close to several training vectors of a single class, it will be represented by several elements of $\mathbf{a}^1$ that are close to one.

The second layer weights, $\mathbf{LW}^{1,2}$, are set to the matrix $\mathbf{T}$ of target vectors. Each vector has a one only in the row associated with that particular class of input, and zeros elsewhere. The multiplication $\mathbf{T}\mathbf{a}^1$ sums the elements of $\mathbf{a}^1$ due to each of the $\mathbf{K}$ input classes. Finally, the second layer transfer function, *compet*, produces a one corresponding to the largest element of $\mathbf{n}^2$, and zeros elsewhere. Thus, the network has classified the input vector into a specific one of $\mathbf{K}$ classes because that class had the maximum probability of being correct. In this way, we can identify the class label for the input audio feature vector.

3.3.4 Experimental Results

The audio classification experiments have been carried out based on the two audio databases introduced earlier. All files in the two database have two labels. One is a coarse label from the three major classes namely music, speech and sound, the other is a fine label of 16 classes for database I and 29 classes for database II. Both databases are split into two equal parts: one for training, and the other for testing (if the file number in a particular class is odd, then the training set has one more sample than the test set). There are four experiments conducted to examine the audio classification abilities of traditional statistical approach and neural network approach shown below:

- **Experiment 1:** Classifying database I into three major classes

- **Experiment 2:** Classifying database I into 16 classes
- **Experiment 3:** Classifying database II into three major classes
- **Experiment 4:** Classifying database II into 29 classes

In each experiment, the four classifiers described previously, NN, k-NN, GMM, and PNN together with the SFS feature selection scheme are used to perform the classification task. In the first and third experiments, the total number of mixing kernels and k for GMM and k-NN are both set to 4, while in the second and fourth experiments, the total number of mixing kernels and k are set to 2 due to fewer files in each class. In all the experiments, the maximum iteration of GMM is set to 32 and the smoothing coefficient of PNN is set to 1.

The classification accuracy versus feature dimension for the four classifiers in the experiments are shown in Figures 3.19, 3.20, 3.21, and 3.22. The overall and individual classification performance of these classifiers in each of the four experiments are given in Tables 3.6, 3.7, 3.8, and 3.9, respectively. The numbers of correctly classified files together with their corresponding accuracies in percentage given in the parenthesis are listed in the tables.

Figure 3.19: Performance of audio classification experiment 1.

Figure 3.20: Performance of audio classification experiment 2.

Figure 3.21: Performance of audio classification experiment 3.

Figure 3.22: Performance of audio classification experiment 4.

Table 3.6: The overall and individual classification accuracy of experiment 1.

Class name	Test	NN	k-NN	GMM	PNN
1.Speech	26	26(100)	26(100)	26(100)	23(88.5)
2.Music	150	150(100)	150(100)	149(99.3)	150(100)
3.Sound	31	31(100)	29(93.5)	29(93.5)	29(93.5)
Total	207	207(100)	205(99.0)	204(98.6)	202(97.6)

Table 3.7: The overall and individual classification accuracy of experiment 2.

Class name	Test	NN	k-NN	GMM	PNN
1.Speech	26	21(80.8)	22(84.6)	20(76.9)	18(69.2)
Female	18	14(77.8)	16(88.9)	15(83.3)	15(83.3)
Male	8	7(87.5)	6(87.5)	5(62.5)	3(37.5)
2.Music	150	141(94.0)	145(96.7)	133(88.7)	142(94.7)
Trombone	7	6(85.7)	7(100)	5(71.4)	5(71.4)
Cello	23	23(100)	23(100)	23(100)	23(100)
Oboe	16	14(87.5)	15(93.8)	13(81.3)	14(87.5)
Percussion	51	49(96.1)	49(96.1)	46(90.2)	49(96.1)
Tubular-bell	10	10(100)	10(100)	10(100)	9(90.0)
Violin-bowed	23	20(87.0)	22(95.7)	18(78.3)	23(100)
Violin-pizzicato	20	19(95.0)	19(95.0)	18(90.0)	19(95.0)
3.Sound	31	23(74.2)	25(80.6)	21(67.7)	25(80.6)
Animal	4	3(75)	4(100)	4(100)	4(100)
Bell	4	3(75)	2(50)	2(50.0)	3(75.0)
Crowds	2	2(100)	2(100)	2(100)	2(100)
Laughter	3	3(100)	3(100)	1(33.3)	3(100)
Machines	6	2(33.3)	4(66.7)	2(33.3)	3(50.0)
Telephone	8	8(100)	8(100)	8(100)	8(100)
Water	4	2(50.0)	2(50.0)	2(50.0)	2(50.0)
Total	207	185(89.4)	192(92.8)	174(84.1)	185(89.4)

Table 3.8: The overall and individual classification accuracy of experiment 3.

Class name	Test	NN	k-NN	GMM	PNN
1.Speech	66	44(66.7)	48(72.7)	57(86.4)	44(66.7)
2.Music	229	202(88.2)	202(88.2)	186(81.2)	186(81.2)
3.Sound	324	278(85.8)	300(92.6)	277(85.5)	304(93.8)
Total	619	524(84.7)	550(88.9)	520(84.0)	534(86.3)

Table 3.9: The overall and individual classification accuracy of experiment 4.

Class name	Test	NN	k-NN	GMM	PNN
1.Speech	66	35(53.0)	41(62.1)	38(57.6)	37(56.1)
Female	33	21(63.6)	26(78.8)	25(75.8)	26(78.8)
Male	27	13(48.1)	14(51.9)	13(48.1)	11(40.7)
Mixture	6	1(16.7)	1(16.7)	0(0)	0(0)
2.Music	229	165(72.1)	173(75.5)	152(66.4)	168(73.4)
2.1Instruments:	182	142(78.0)	153(84.1)	134(73.6)	147(80.8)
Brass	22	17(77.2)	18(81.8)	18(81.8)	15(68.2)
Keyboard	3	1(33.3)	1(33.3)	0(0)	1(33.3)
Percussion	62	47(75.8)	49(79.0)	42(67.7)	46(74.2)
String	70	61(87.1)	65(92.9)	55(90.2)	66(94.3)
Synthesis	7	1(14.3)	3(42.9)	3(42.9)	3(42.9)
Woodwind	18	15(83.3)	17(94.4)	16(88.9)	16(88.9)
2.2Melody	18	9(50.0)	6(33.3)	7(38.9)	7(38.9)
Melody	18	9(50.0)	6(33.3)	7(38.9)	7(38.9)
2.3Song	29	14(48.3)	14(48.3)	11(37.9)	14(48.3)
Cartoon Song	4	1(25.0)	0(0)	0(0)	0(0)
Normal song	25	13(52.0)	14(56.0)	11(44.0)	14(56.0)
3.Sound	324	181(55.9)	175(54.0)	156(48.1)	192(59.3)
3.1Animal	110	81(73.6)	81(73.6)	63(57.3)	87(79.1)
Birds	23	13(56.5)	11(47.8)	10(43.5)	7(30.4)
Mammals	87	68(78.2)	70(80.5)	53(60.9)	80(92.0)
3.2Effect	172	89(51.7)	83(48.3)	78(45.3)	94(54.7)
Alarm	8	3(37.5)	0(0)	2(25.0)	2(25.0)
Beat	21	10(47.6)	13(61.9)	6(28.6)	10(47.6)
Bells	25	10(40.0)	10(40.0)	13(52.0)	8(32.0)
Car	6	0(0)	0(0)	0(0)	0(0)
Clock	8	2(25.0)	0(0)	1(12.5)	2(25.0)
Explosion	22	15(68.2)	12(54.5)	13(59.1)	17(77.3)
Gun	12	8(66.7)	7(58.3)	3(25.0)	7(58.3)
Machines	55	31(56.4)	43(78.2)	30(54.5)	39(70.9)
Telephone	15	10(66.7)	8(53.3)	10(66.7)	9(60.0)
3.3Nature	20	4(20.0)	4(20.0)	4(20.0)	6(30.0)
Storm	5	1(20.0)	1(20.0)	1(20.0)	3(60.0)
Thunder	3	0(0)	0(0)	0(0)	0(0)
Water	10	3(30.0)	3(30.0)	3(30.0)	3(30.0)
Wind:	2	0(0)	0(0)	0(0)	0(0)
3.4People	22	7(31.8)	7(31.8)	11(50.0)	5(22.7)
Applause	4	2(50.0)	1(25.0)	3(75.0)	2(50.0)
Laughter	18	5(27.8)	6(33.3)	8(44.4)	3(16.7)
Total	619	381(61.6)	389(62.8)	346(55.9)	397(64.1)

From the figures, we can see that the classification performance increases rapidly with the increase of features at the beginning. After reaching a peak value, it remains constant or may even decrease after a particular number for a certain classifier. All classifiers reach their best performance at approximately 20 features. This shows that the SFS feature selection procedure is an efficient method to quickly find a small set of features to yield a satisfactory result among a large set. Thus, for simplicity and fair comparison, all the listed classification accuracies in the four tables are achieved by their corresponding classifiers using 20 features selected using the SFS method. Note that, the best 20 feature sets for different classifiers are different.

From the tables, it is observed that, for the overall classification accuracy, the four classifiers have different performance for different experiments. When the database is small and class number is few as in experiment 1, the NN classifier yields best performance, which can be as high as 100%. When the database grows larger and class number remains the same, there are more varieties in each class, as experiment 3 shows, and the k-NN achieved the best performance which is 88.9%. When we divide the database into more classes, the performance decreases, as shown in experiments 2 and 4. The k-NN again achieves the best performance for classifying the database I into 16 classes with 92.8% accuracy, while PNN obtains the best performance for classifying database II into 29 classes with 64.1% accuracy.

In particular, during the second experiment, our k-NN classifier with 28 features selected by the SFS method, yields 93.7% accuracy, as compared to that of 90.4% by nearest feature line (NFL) using 34 features, tested on the common audio database I [76].

3.4 Automatic Classification of Musical Instrument Sounds

One subdivision of audio classification, the music instrument classification, has gained special interests recently. In this section, we follow the traditional statistical approach and adopt a similar methodology as for general audio classification to carry out a study on classification of musical instruments. In the procedure, we extract 58 features for each record in the music database of 351 files and then apply the sequential forward selection method associated with statistical classifiers for the classification task. Three different classification techniques have been tested and an accuracy of up to 93% can be achieved by using 19 features to classify musical instruments into five major classes, namely string, brass, percussion, woodwind, and keyboard.

3.4.1 Introduction

The collection of musical instrument sounds is an obligatory part of comprehensive music digital libraries. Automatic musical instrument classification can be very helpful for indexing the database as well as for annotation and transcription. Kaminsky et al. [24] identified four instruments, namely guitar, piano, marimba and accordion, using an artificial neural network or nearest neighbor classifier. The results of this preliminary work achieved were encouraging although only temporal features were utilized. Martin and Kim applied a statistical pattern-recognition technique for the classification of musical instrument tones within a taxonomic hierarchy using perceptually salient acoustic features and resonance structure [25]. Miiva et al. [27] separated polyphonic music into

each monophonic one using comb filters and estimated musical instruments by frequency analysis. More recently, a system for musical instrument recognition was presented where a wide set of features were used to model the temporal and spectral characteristics of sounds [26].

As for the general audio classification, due to the increasing number of audio features extracted and available, we employ the SFS method with statistical classifier to pick up a small set of ‘good’ features and combine them to achieve higher classification accuracy for musical instruments classification. In the experiments, the nearest neighbor (NN) classifier, modified k-nearest neighbor (k-NN) classifier, and Gaussian mixture model (GMM) are utilized.

3.4.2 The Music Database

The musical instruments are commonly sorted into five families according to their vibration nature, which are string, brass, percussion, woodwind, and keyboard. Accordingly, our music instruments database has 351 samples collected from different instruments among these five categories. The lengths of the sound files range from a half second to around ten seconds. The average size of the sound files is around 17KB. A brief description of the sound files is given in Table 3.10.

3.4.3 Feature Extraction and Selection

We proceed with the similar feature extraction process, where every audio file is divided into frames of 256 samples, with 50% overlap at the two adjacent frames. Each frame is hamming-windowed and 58 features are extracted for

Table 3.10: The musical instrument collection.

Classes	Instruments
Brass	Fanfare, French horn, trombone, trumpet, tuba
Keyboard	Piano
Percussion	Bell, bongo, chime, conga, cymbal, drum, gong, maraca, tambourine, triangle, timbales, tomtom, tympani
String	Guitar, violin
Woodwind	Oboe, saxophone

each frame. Means and standard deviations of the frame-based features are computed as the final features for each audio file. The 58 features from three categories are showed in Table 3.11. The features 1-8 are temporal features, 9-32 are spectral features, and 33-58 are coefficient features. Again, each audio feature is normalized over all files in the database by subtracting its mean and dividing by its standard deviation. Then, the SFS method is used with the statistical method for classification.

Table 3.11: The descriptions of audio features for music instrument classification.

Feature number	Descriptions
1-2	Mean and standard deviation of volume root mean square(RMS)(μ_{RMS}, σ_{RMS})
3	Volume dynamic ratio(VDR)
4	Silence ratio(SR)
5-6	Mean and standard deviation of frame energy(E) (μ_E, σ_E)
7-8	Mean and standard deviation of zero crossing ratio(ZCR)(μ_{ZCR}, σ_{ZCR})
9-10	Mean and standard deviation of centroid(ω_c)($\mu_{\omega_c}, \sigma_{\omega_c}$)
11-12	Mean and standard deviation of bandwidth(B) (μ_B, σ_B)
13-20	Means and standard deviations of four sub-band energy ratios ($ER^{(1-4)}$) $\mu_{ER}^{(1-4)}, \sigma_{ER}^{(1-4)}$
21-22	Mean and standard deviation of pitch(P)(μ_P, σ_P)
23-24	Mean and standard deviation of salience of pitch(SP)(μ_{SP}, σ_{SP})
25-28	Means and standard deviations of first two formant frequencies(F1,F2)($\mu_{F1}, \mu_{F2}, \sigma_{F1}, \sigma_{F2}$)
29-32	Means and standard deviations of first two formant amplitudes(A1,A2)($\mu_{A1}, \mu_{A2}, \sigma_{A1}, \sigma_{A2}$)
33-58	Means and standard deviations of first 13 orders of $MFCC_{1-13}$ (Mel-Frequency Cepstral Coefficients)($\mu_{MFCC_{1-13}}, \sigma_{MFCC_{1-13}}$)

3.4.4 Experiments

The database is also split into two equal parts: one for training, and the other for testing. Three classifiers, NN, modified k-NN and GMM, are used to classify the musical instruments along with SFS method. In the modified k-NN, we first find the k ($k = 3$) nearest neighbors from each class instead of the whole training set, calculated and sorted their means, then assign the testing feature vector to the class corresponding to the smallest mean. The classification accuracies versus feature dimension of the three classifiers using combination of temporal, spectral, and coefficient features, are shown in Figure 3.23. From the figure, we can find the same situation as in general audio classification, i.e. that the performance increases rapidly with the increase of features at the beginning and then it remains more or less constant and even decreases after a particular number of features. Other experiments using single temporal, spectral, coefficient feature or any combination of them show a similar phenomenon. The classification accuracies of the best feature set and the corresponding feature numbers are listed in Table 3.12. The best feature sets for different classifiers are different. Among all the experiments, the modified k-NN classifier using 19 features, in which 6 are from temporal, 8 are from spectral, and 5 are from coefficients, achieves the highest accuracy of 93%.

3.4.5 Summary of Musical Instrument Classification

In our music instrument classification experiments, the NN, k-NN, and GMM classifiers with the SFS scheme are used to pick up the best feature set from temporal, spectral, and coefficient domain for classifying musical instruments into five families. The simple k-NN classifier using a small set of features can achieve

Table 3.12: The highest audio classification accuracies achieved based on different feature sets.

Features	NN	k-NN	GMM
Time	0.73(7)	0.73(6)	0.71(6)
Frequency	0.80(8)	0.82(10)	0.80(11)
Coefficient	0.86(17)	0.85(13)	0.80(17)
Time and Coefficient	0.90(14)	0.86(9)	0.84(14)
Frequency and coefficient	0.89(29)	0.91(15)	0.85(17)
Time and frequency	0.85(12)	0.87(7)	0.81(8)
Time, frequency, and coefficient	0.91(13)	0.93(19)	0.87(22)

a satisfactory result. Since the number of features is reduced, less computation time is required for classifying the music instrument sounds. In addition, this will be beneficial to real-time applications such as music retrieval from large databases.

3.5 Discussion

There are increasingly audio features extracted from time, frequency and coefficient domains. Among them, some are directly borrowed from speech processing. However, new features with better class separability are to be discovered and some existing feature extraction algorithms, such as timbre and pitch extraction, need further improvement to be robust.

For example, timbre is an important characteristic of audio especially for instruments, it is an subjective psychoacoustics feature and its property can be expressed as warm, cold, scratch (roughness), strong, and so on, by humans intuitively. In order to describe timbre accurately, multi-dimension description is crucial, such as onset-time, offset-time, duration and spectral envelope. Only

Figure 3.23: The audio classification accuracy versus feature dimension.

when all of these simple properties have been correctly calculated can we express timbre well.

Another example is pitch. The limitation of pitch determination by short time autocorrelation function is that, for environmental audio signal, the pitch period may not be constant and the autocorrelation algorithm is not able to accurately measure signals that vary widely and are not periodic in nature. It is sometimes subject to alias: an integer multiple of the actual pitch (octave) may be picked out and it is computationally intensive.

As for the feature selection scheme, the SFS method is computationally attractive; the drawback is that once a feature is retained as part of a best feature set, it can't be discarded. Hence, we can adopt alternative feature selection method such as “plus l take away r ($l > r$)” selection as mentioned in [110].

For audio classification, including one special case of music instrument clas-

sification, we have made an effort to conduct our experiments from various approaches, namely the Nearest Neighbor(NN) classifier, the modified k-Nearest Neighbor (k-NN) classifier, the Gaussian Mixture Model (GMM), and the Probabilistic Neural Network (PNN) classifier. The NN classifiers requires no training and it is very easy to implement. It is widely used as a benchmark for the classification task. However, the NN classifier is sensitive to feature selection and feature noise. To overcome the drawback of its sensitive to feature selection and feature noise, the modified k-Nearest Neighbor (k-NN) is proposed and implemented. The k-NN achieves the robustness by averaging k-nearest neighbors. Both the NN and k-NN classifiers are computationally expensive for performing classification if the feature dimension is high and the number of training samples is large. The PNN classifier can be designed in a straightforward manner and it does not depend on training. A PNN is guaranteed to converge to a Bayesian classifier providing it is given enough training data. However, the PNN requires a very large number of Gaussian kernels when the training set is large. When a larger number of Gaussian kernels are used, PNN is also computationally expensive to perform the classification. A complementary technique for PNN, the polynomial adaline, can be used instead [115]. The polynomial adaline has a equivalent decision boundary as the PNN, but it is computationally efficient for classification, though some sort of training is needed. Besides the non-parametric classifiers, the parametric classifier Gaussian Mixture Model is selected for its ability to model arbitrary probability density functions. The GMM requires intensive training to estimate the parameters by using the EM algorithm. The advantage of the GMM classifier is the fast classification performance once the training is completed.

For all the classifiers used in the experiments, each of them has its variations.

For the NN classifier, a distance measure other than Euclidean distance, such as the Mahalanobis (a normalized Euclidean) distance measure can be adopted instead. For the k-NN classifier, different k can be tested to find the best one with highest classification accuracy. For the GMM classifier, the total number of mixing kernels can be adjusted, and the performance varies with the random initials for the EM algorithm. For the PNN classifier, the smoothing coefficient can be chosen from 0 to 2. Therefore, higher classification accuracy can be achieved by tuning the parameters without changing the classifier structure.

In summary, the SFS feature selection method together with statistical classifiers or neural networks are efficient for the general audio classification and its subset of the music instruments classification. With modification of this approach, better performance may be achieved.

Chapter 4

Content-Based Audio Retrieval

The advent of the World Wide Web has increased the need for automated information retrieval for extremely large document collections as well as multimedia databases, such as digital libraries. Traditional text search engines, like Google and Yahoo, can provide a portal for web surfers to find their web pages of interest. Yet, commercial search engines for multimedia databases, especially for audio, are lacking. Although several search engines claim to provide some search ability on multimedia contents, most of these systems are based on the surrounding texts or manually added annotations of the multimedia data. However, users can benefit from the ability to directly search these media, which contain rich information but could not be precisely described by text. Hence, content-based indexing and retrieval technologies are the first crucial step towards building automatic multimedia search engines.

In this chapter, we focus on content-based audio retrieval. Conventional method of searching and retrieving text document can be frustrated by the inability to look inside the audio objects. Here, we access audio by its extracted

feature vector. A user may submit a query audio example and request relevant files to the query. Both the query and audio documents in the database are represented as feature vectors. A series of similarity measurements between the query and each sample document are computed, then a list of files based on the similarity are displayed to the user for listening and browsing.

4.1 Introduction

Content-based audio retrieval is less studied in the context of content-based multimedia retrieval. The process of a common audio retrieval system can be concisely divided into three parts: audio feature extraction, classifier mapping, and distance ranking. The first step towards these content-based audio database systems is to extract features from sound signals. Based on the extracted features, various classifiers can then be used for sound classification. Once an audio has its label, it can be indexed and annotated for browsing. Contrary to using keywords as queries in text retrieval, examples are used as queries in audio retrieval. Usually, the similarities between the audio samples in the database and the query example are calculated, then distance-ranked list is given as the retrieval result.

A full-fledged procedure of an integrated audio retrieval system is illustrated in Figure 4.1. Raw audio recordings are segmented based on abrupt changes within features. Then, audio segments are classified and indexed. They are stored in corresponding archives. The audio archives can be organized in a hierarchical way for the ease of the storage and retrieval of audio clips. When a user wants to find audio samples of interest in the archives, he/she may put

a set of features or a query sound into the system. The search engine will then find the best matched sounds and present them to the user. The user may also refine the query to get more audio material relevant to his/her interest.

Figure 4.1: The procedure of content-based audio retrieval.

4.2 Query-by-Example

Unlike text document retrieval, it is hard to retrieval audio database by key-words, because different users may have different opinions about the same audio and it is extremely difficult to extract the high level semantics to match by key-words. Therefore, query-by-example (QBE) is considered to be a promising approach since it provides a user with an intuitive way of query representation and the form of query expression is as same as the data to be retrieved. Sometimes the query example is generated by humming on site by the user. We use QBE to active the search engine in the following experiments.

4.3 Probabilistic Neural Network for Content-based Audio Retrieval

Obviously, the direct retrieval scheme is conducted by calculating some distances, such as Euclidean distances, between the query example and individual samples in the database. The retrieval list is then given based on the distance measurement. The search engine searches through the whole database. Many non-relevant sounds are retrieved and the speed is slow especially when the database is growing large. To increase the search speed and retrieve more relevant files to the query, a probabilistic neural network (PNN) classification combined with Euclidean distance measurement hierarchical scheme for retrieval is proposed. This is because the PNN classifier demonstrated better performance during the classification of the larger database as described in the previous chapter.

With this hierarchical retrieval strategy, query sound is firstly classified into one of three major classes, namely music, speech and sound by the calculated probabilities from PNN. Then the distance between query and sample audios in that class instead of the whole database are measured and an ascending distance list is given as the retrieval result. With this approach, many non-relevant documents are avoided before the search begins.

Key advantages of PNN are that training requires only a single pass and the decision surfaces are guaranteed to approach the Bayes optimal decision boundaries as the number of training samples grows. Furthermore, the shape of the decision surfaces can be made as complex as necessary, or as simple as desired, by choosing the appropriate value of the smoothing parameter. In prac-

tical applications, there exist fast learning algorithms. PNN is easy to use and extremely fast for moderate-sized database. For very large databases and for mature applications in which classification speed is more important than training speed, a complementary probabilistic neural network, polynomial adaline can be used as a substitute [115]. Based on these facts, PNN is chosen as the classifier for the first stage of retrieval. The following experiments are conducted using the direct retrieval method and the proposed hierarchical scheme. Note that the PNN classification is the same as that introduced in chapter 3 with the first 20 best features selected.

4.3.1 Performance Measurement

The performance of retrieval is measured by precision and recall [118]. Precision and recall are defined as follows:

$$Precision = \frac{Relevant\ Retrieved}{Total\ Retrieved} \quad (4.1)$$

$$Recall = \frac{Relevant\ Retrieved}{Total\ Relevant} \quad (4.2)$$

The precision provides an indication of the quality of the answer set, while the recall considers the ability of the system to retrieve total number of relevant documents. In an ideal situation, precision is always 1 at any recall point. Another measurement is the average precision (AP), which is used as a single indicator of retrieval performance, which refers to an average of precision at various points of recall.

$$AveragePrecision = \frac{1}{TR} \sum_{i=1}^{TR} Precision(i) \quad (4.3)$$

where TR is the number of total relevant files, $Precision(i)$ is the precision at i th relevant files retrieved. We assume that the files in a same class are relevant, otherwise, they are non-relevant. Hence, the performance can be measured automatically without hearing the sound.

4.3.2 Experimental Results

As for the test in audio classification, the two databases are split into two equal parts: one for training, and the other for testing. Each file in the testing set is used as a query sound and submitted to the search engine one by one.

One typical and the mean precision-recall (P-R) curve obtained by searching database I and II by the proposed method and direct method are shown in Figures 4.2, 4.3, 4.4, and 4.5. The mean P-R curve is drawn as follows. Firstly, we use the linear interpolation to find the precision values at individual step increasing recall points from 0.01 to 1 with 0.01 as step. Then, we have the mean value of the precision values at each recall points. Thus, we can draw the mean precision-recall curve, which is a better description of the overall retrieval performance.

As shown in Figures 4.2, 4.3, 4.4, and 4.5, the P-R curve of the proposed method is above the curve of the direct distance method. This is to say that when we recall the same number of documents, the proposed method can retrieve fewer total or non-relevant files. It also means if we retrieve the same number of total documents using both methods, among them, more relevant files can be retrieved by the proposed method. The average precision for database I retrieved by proposed scheme and the simple distance method are 0.57, and 0.53, while the average precision for database II are 0.31, and 0.25, respectively.

These results show that the proposed retrieval scheme yields better overall performance than the direct distance method in both recall-precision relation and average precision.

Most often, people only browse the files ranking in the top of the list. For the purpose of illustration, the top ten retrieved files for several queries with correct output of PNN classifier are listed in Tables 4.1 and 4.2. For example, shown in the second and third column of Table 4.1, the top ten searching results by the direct distance method for a male speech query are from the classes of male speech, percussion, percussion, female speech, female speech, machine, cello, percussion, violin-pizzicato, and percussion according to similarity. They are from six different audio classes and only one relevant sample retrieved. The top ten results by the proposed method for the same query are from the classes of male speech and female speech only. Although there are still 2 files retrieved from the same class of query by the proposed method, the rest of results are in the same coarse class “speech”.

Figure 4.2: One typical precision-recall (P-R) curve of audio retrieval using database I.

Figure 4.3: One typical precision-recall (P-R) curve of audio retrieval using database II.

Figure 4.4: The mean precision-recall (P-R) curve of audio retrieval using database I.

Figure 4.5: The mean precision-recall (P-R) curve of audio retrieval using database II.

Table 4.1: Experimental result of audio retrieval on database I.

Query	Male		Oboe	
Search method	Direct	Proposed PNN	Direct	Proposed PNN
No.1	Male	Male	Oboe	Oboe
No.2	Percussion	Female	Oboe	Oboe
No.3	Percussion	Female	Oboe	Oboe
No.4	Female	Female	Telephone	Oboe
No.5	Female	Female	Oboe	Trombone
No.6	Machines	Female	Trombone	Oboe
No.7	Cello	Female	Telephone	Oboe
No.8	Percussion	Female	Telephone	Oboe
No.9	Violin-pizzicato	Male	Oboe	Trombone
No.10	Percussion	Female	Oboe	Oboe
Relevant Number	1	2	6	8
Query	Animal		Laughter	
Search method	Direct	Proposed PNN	Direct	Proposed PNN
No.1	Animal	Animal	Laughter	Laughter
No.2	Telephone	Telephone	Animal	Animal
No.3	Telephone	Telephone	Tubular-bell	Laughter
No.4	Telephone	Telephone	Laughter	Machines
No.5	Violin-pizzicato	Animal	Percussion	Laughter
No.6	Trombone	Telephone	Percussion	Telephone
No.7	Animal	Laughter	Tubular-bell	Machines
No.8	Telephone	Telephone	Tubular-bell	Telephone
No.9	Trombone	Machines	Cello	Machines
No.10	Percussion	Animal	Machines	Telephone
Relevant Number	2	3	2	3

Table 4.2: Experimental result of audio retrieval on database II.

Query	Female		Percussion	
Search method	Direct	Proposed PNN	Direct	Proposed PNN
No.1	Machines	Female	Percussion	Percussion
No.2	Female	Female	Percussion	Percussion
No.3	Female	Female	Female	Melody
No.4	Thunder	Female	Clock	Percussion
No.5	Mammals	Male	Birds	Percussion
No.6	Female	Male	Beat	Percussion
No.7	Female	Female	Melody	Percussion
No.8	Male	Female	Mammals	Percussion
No.9	Mammals	Female	Water	Percussion
No.10	Mammals	Female	Birds	Percussion
Relevant Number	4	8	2	8
Query	Mammals		Machines	
Search method	Direct	Proposed PNN	Direct	Proposed PNN
No.1	Mammals	Mammals	Machines	Machines
No.2	Mammals	Mammals	Normal Song	Car
No.3	Female	Machines	Normal Song	Machines
No.4	Male	Birds	Car	Machines
No.5	Machines	Laughter	Machines	Explosion
No.6	Male	Mammals	Normal Song	Explosion
No.7	Birds	Mammals	Machines	Storm
No.8	Laughter	Birds	Explosion	Machines
No.9	Mammals	Beat	Melody	Machines
No.10	Male	Mammals	Explosion	Alarm
Relevant Number	3	5	3	5

4.4 Content-based Audio Retrieval With Relevance Feedback

In both the direct and proposed audio retrieval schemes introduced in the previous section, the user interaction is not considered. However, the user involvement, such as relevance feedback, can be beneficial to improve the initial retrieval performance. Basically, the purpose of relevance feedback is to move relevant files ranking to the top and non-relevant files ranking to the bottom. In principle, there are two ways to apply user's feedback strategy. One is to update the weights in the similarity measurement and the other is to refine the query. Here, we focus on the former approach.

Two relevance feedback algorithms for content-based audio retrieval have been proposed in this section. One is a modified version of a technique used for image retrieval with positive feedback; another is based on a constrained optimization concept. Experiments show that the latter approach can yield similar performance improvement to the former and it has the advantage of utilizing both negative and positive feedback in a unified approach. Note that in the experiments of audio retrieval with relevance feedback, the features numbered from 1 to 61 in Table 3.4 are used.

4.4.1 Relevance Feedback in the Related Research Field

Relevance feedback has been proved to be an effective method to increase performance in text retrieval [119], [120], [121], [122], image retrieval [123], [124], [125], [126], and video retrieval [127], [128]. Among the literature in the context of multimedia (image, video, and audio) retrieval, relevance feedback in

image retrieval has been most intensively studied. Rui et al. [129] proposed a relevance feedback based interactive retrieval approach, which aimed to narrow the gap between high-level concepts and low-level features and take advantage of the subjectivity of human perception of visual content. During the retrieval process, the user's high-level query and perception subjectivity were captured by dynamically updated weights based on the user's feedback. Muller et al. [130] compared a variety of strategies for positive and negative feedback in image retrieval. The utility of negative feedback was highlighted, especially over several feedback steps.

4.4.2 Relevance Feedback for Audio Retrieval

Among various feedback techniques proposed in the related fields, especially in image retrieval, some can be applied in audio without much effort, others are not feasible to be used for this purpose. Here, we introduce two relevance feedback techniques to carry out retrieval in audio domain. Both of the relevant feedback algorithms are proposed through the weights updating rule. The first method is a simplified feedback procedure originally proposed for image retrieval by Aksoy et al. in [131]. Aksoy et al. presented a weighted distance approach to relevance feedback in their work, where the weights are the ratios of the standard deviations of the feature values for both the whole database and the images selected as relevant by the user. The second method is proposed from a different perspective, which is based on constrained optimization.

An interface of such an audio retrieval system with relevance feedback we have developed is illustrated in Figure 4.6. In the system, a query example is used as an retrieval input. A similarity measurement, such as Euclidean

distance, between the query and sample audio files is computed. Then, a list of files according to the decreasing similarity or increasing distance is displayed to the user for listening and browsing. Upon the initial results, the user can mark files with relevant or non-relevant tags according to his or her judgment and preference. Then, the system updates the result based on the feedback in order to find more relevant files which appear on the top list progressively.

On the right side of the figure, the database list contains all the audio files in the database. The user can select any of them or record external resources through microphone. The waveform of the input will be plotted immediately. The features of the file in the database are already extracted, and can be used for retrieval directly. When an external resource is used, it must go through the feature extraction process before classification and retrieval. An initial result will be shown in the results list after the user press the “Query” button. Then, the user can press the “play” button to listen to the plotted audio and right-click mouse in the results list region to mark the highlighted file as relevant or non-relevant accordingly. After the “Feedback” button is pressed, a new retrieval list is generated and displayed.

4.4.3 Relevance Feedback Algorithm I

As introduced previously, the relevance feedback algorithm I is a simplified technique proposed by Aksoy et al. in [131]. The underlying concept of the algorithm can be often seen in other relevance feedback in image retrieval. The idea is that we should assign more weights to the feature, which has a diverse set of values in the whole database but a similar value for relevant images. Then, more relevant images will appear on the top retrieval list after feedback. Based

Figure 4.6: The GUI of the audio retrieval system.

on this observation, we simplify the algorithm and apply it to audio retrieval. During the feedback process, the weighted L_2 distance instead of Euclidean distance is adopted to calculate the similarity measurement. The weighted L_2 distance is defined as follows:

$$\rho(\mathbf{x}, \mathbf{y} : \mathbf{w}) = \left(\sum_{i=1}^N w_i (x_i - y_i)^2 \right)^{1/2} \quad (4.4)$$

where $\mathbf{x}$ and $\mathbf{y}$ are two given feature vectors, $\mathbf{w}$ is the weight vector and N is the dimension of the feature vectors.

The feedback retrieval algorithm is described as follows:

1. Initialize all the weights to $1/N$. That is,

$$w_i = 1/N, \quad i = 1, 2, \dots, N \quad (4.5)$$

2. Search the database, compare the distance between query example and each file in database by using $\mathbf{w}$ in (4.5) and obtain retrieval result list.
3. Get feedback from the user and form the relevance audio set R_{rel} (the original query example is always included in R_{rel}).
4. Calculate the standard deviation (denoted as std) of each feature in R_{rel} , $\sigma_i = std(F_{rel,i})$, where $F_{rel,i}$ is the i th feature components of the audios in R_{rel} .
5. Update weights according to the following rule:

$$w_i = \frac{(w_i + \Delta w_i)}{(\sum_{i=1}^N \Delta w_i + 1)} \quad (4.6)$$

where $\Delta w_i = \frac{1}{\sigma_i + \alpha}$. Note that the standard deviation of the i th feature components of the whole audio database is 1 according to our normalization scheme. The constant α is a positive number, which is experimentally determined. The sum of the new weights remains the same as the sum of old ones, which equals to 1.

6. Go back to step 2 and search using the updated weight vector $\mathbf{w}$.

4.4.4 Relevance Feedback Algorithm II

Suppose that we have obtained the relevant audios set R_{rel} , which includes the query example q and relevance feedback $f^j, j = 1, \dots, M$, where M is the number of relevant feedback files. If we can decrease the sum of the square weighted $L2$ distance $\sum_{j \in R_{rel}} \rho^2(f^j, q : w)$ between the files of relevance feedback and the query example, more relevant audios may emerge on the top of the retrieval list because of their similar feature characteristics. The weighted $L2$ distance is rewritten as follows:

$$\rho(\mathbf{f}^j, \mathbf{q} : \mathbf{w}) = \left(\sum_{i=1}^N w_i (f_i^j - q_i)^2 \right)^{1/2} \quad (4.7)$$

where the subscript i refers to the i th feature element, the superscript j refers to the j th file in the relevant set. Based on this observation, we consider minimizing the following objective function:

$$f(\mathbf{w}) = \mathbf{w}^T \mathbf{D} \mathbf{w} + \varepsilon \mathbf{w}^T \mathbf{w}, \text{ subject to } \mathbf{c}^T \mathbf{w} = 1 \quad (4.8)$$

where ε is a positive constant, and

$$\mathbf{D} = \text{diag}\{d_1, \dots, d_N\} = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_N \end{pmatrix}, \mathbf{c} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \quad (4.9)$$

Here, $d_i = \sum_{j \in R_{rel}} d_{ji}^2$ and d_{ji} is the distance between i th feature component of the j th relevance feedback and query example, i.e. $d_{ji} = f_i^j - q_i$. The term $\varepsilon \mathbf{w}^T \mathbf{w}$ is introduced to avoid very large variation of $\mathbf{w}$. This is a typical constrained optimization problem, which can be solved by the Lagrangian method. The

solution of the constrained optimization problem is given by

$$\mathbf{w} = \frac{\mathbf{R}^{-1}\mathbf{c}}{\mathbf{c}^T\mathbf{R}^{-1}\mathbf{c}} = \frac{1}{(r_1^{-1} + \dots + r_N^{-1})} \begin{pmatrix} r_1^{-1} \\ \vdots \\ r_N^{-1} \end{pmatrix} \quad (4.10)$$

which has an equivalent form as follows:

$$w_i = \frac{1}{(r_1^{-1} + \dots + r_N^{-1})} r_i^{-1} \quad (4.11)$$

where $\mathbf{R} = \mathbf{D} + \varepsilon\mathbf{I}$ and $r_i = d_i + \varepsilon$. In the case of negative feedback, the objective function can be adjusted as follow:

$$f(\mathbf{w}) = \mathbf{w}^T\mathbf{D}\mathbf{w} - \beta\mathbf{w}^T\mathbf{D}'\mathbf{w} + \varepsilon\mathbf{w}^T\mathbf{w} + \lambda(\mathbf{c}^T\mathbf{w} - 1) \quad (4.12)$$

where β is a positive number and is usually small to reduce the effect of negative feedback compared to positive feedback, $\mathbf{D}' = \text{diag}\{d'_1, \dots, d'_N\}$ and $d'_i = \sum_{j \in R_{irrel}} d'_{ji}$. R_{irrel} is defined as the non-relevant or negative feedback audio set. $d'_{ji} = f'_i{}^j - q_i$, and f'^j is the j th negative feedback file in the set R_{irrel} . In this case, the solution to Eq.(4.12) has the same form as in Eq.(4.10) and (4.11) with R being replaced by $\mathbf{R} = \mathbf{D} - \beta\mathbf{D}' + \varepsilon\mathbf{I}$. In our experiments, we set $\alpha = 0.5, \beta = 0.1, \varepsilon = 0.5$ empirically.

4.4.5 Query Refining and Other Issues

As we have mentioned earlier, another way to conduct feedback is to modify the query accordingly. The aim is to adjust the query for better representation of the user's information request. Intuitively, we use the mean of the relevant retrieved

files (including the query example) to form the new query in our experiments. The number of relevant files to select is another issue for consideration during feedback. In the experiments, we chose 1 to 3 files. This is because we notice that normally users are only willing to provide feedback on a few items and excessive feedback can't give much further performance improvement.

4.4.6 Experimental Results

In the retrieval system with feedback, different users may have different opinions and may choose different files as feedback or even determine the same file as relevant or non-relevant. Hopefully, since files in same class are already assumed to be relevant, we can mark those files from most similar to least similar to the query automatically. Therefore, the first 1-3 files are used as relevance feedback for weight updating. Thus, ambiguity of relevance judgment is avoided and experiments are conducted in a fully automatic way.

In most cases, however, users don't have the patience to listen to all the retrieved files. Normally, they are only interested in several files ranking at the top. Thus, AP is calculated again based on top T ($T=15$) retrieved files considered. We call it $AP(15)$, and defined as follows:

$$AP(15) = \frac{1}{TopR} \sum_{i=1}^{TopR} Precision(i) \quad (4.13)$$

where $TopR$ is the number of relevant files ranking at Top 15 retrieved files, $Precision(i)$ is the precision at i th relevant files retrieved. This $AP(15)$ may be a more accurate indicator for practical retrieval performance.

Experiments of audio retrieval with feedback are carried out on the two

databases in the same way as previously, where each audio file in the databases is submitted to the search engine as query example one by one. The mean APs and AP(15)s of the tests on the two databases are measured and listed in Tables 4.3 and 4.4. The retrieval performance without feedback is measured at the beginning. The original mean APs are 0.485 and 0.234 for the small and large database respectively, while mean AP(15)s are 0.852 and 0.799 when the top 15 files are considered only.

From Table 4.3, we can see that, for the database I, when the first 3 relevant files are chosen for relevance feedback, the AP performance can increase to 0.577 and 0.59 by algorithm I and II respectively. For database II, when the first 3 relevant files are chosen as feedback, the performance can increase to 0.305 and 0.32 by algorithm I and II respectively.

From Table 4.4, in the case of one feedback iteration with 3 relevance feedback, we can see that for database I, the mean AP(15) performance can increase to 0.943 and 0.946 by algorithm I and II respectively, while for database II, the mean AP(15) performance can increase to 0.889 and 0.918 by algorithm I and II respectively.

In order to show the overall performance improvement rather than a particular one, the AP difference of the small database after and before the first iteration of feedback with 3 relevant files selected using Algorithm II are shown in Figures 4.7 and 4.8. Figure 4.7 considers the whole retrieved files, while Figure 4.8 considers the top 15 retrieved files only. The bar above the horizontal zero line means that the AP after feedback is higher than before feedback and vice versa. We can clearly see that in most cases, the performance after feedback are better.

Table 4.3: The overall average-precision of the relevance feedback.

AP:0.485	Database I		AP:0.234	Database II	
	Algorithm I	Algorithm II		Algorithm I	Algorithm II
1 Files	0.517	0.52	1 Files	0.265	0.267
2 Files	0.551	0.558	2 Files	0.289	0.297
3 Files	0.577	0.59	3 Files	0.305	0.32

Table 4.4: The overall average-precision of relevance feedback based on Top T (T=15) files considered.

AP(15):0.852	Database I		AP(15):0.799	Database II	
	Algorithm I	Algorithm II		Algorithm I	Algorithm II
1 Files	0.886	0.892	1 Files	0.862	0.863
2 Files	0.92	0.924	2 Files	0.889	0.899
3 Files	0.943	0.946	3 Files	0.889	0.918

For comparison, the top 15 retrieved files of several queries using algorithm II with 3 files selected as relevance feedback and the direct search from Musciefish LLC are shown together in Table 4.5. The tests are conducted based on the common database I. For examples, the search results by Musciefish for a query of altotrombone have found 4 relevant files, while there are 11 relevant files in one feedback using our algorithm II. A search for cellobowed sounds is also increase from 3 to 9 after one iteration of feedback. There are 8 and 11 files in the same class of query retrieved from Musciefish for female speech and telephone ring respectively, while these increase to 12 and 13 correspondingly by our feedback algorithm II.

Figure 4.7: The difference in average-precision of retrieval performance with total files considered.

Figure 4.8: The difference in average-precision of retrieval performance with Top T files considered (T=15).

Table 4.5: The comparison with Musciefish LLC in audio retrieval.

Query	Altotrombone		Cello	
Method	Musciefish	Feedback	Musciefish	Feedback
No.1	Altotrombone	Altotrombone	Cellobowed	Cellobowed
No.2	Oboe	Altotrombone	Violinbowed	Cellobowed
No.3	Cellobowed	Altotrombone	Altotrombone	Cellobowed
No.4	Violinbowed	Altotrombone	Oboe	Cellobowed
No.5	Oboe	Altotrombone	Percussion	Cellobowed
No.6	Altotrombone	Altotrombone	Cellobowed	Cellobowed
No.7	Altotrombone	Altotrombone	Violinbowed	Violinbowed
No.8	Violinbowed	Altotrombone	Cellobowed	Cellobowed
No.9	Cellobowed	Altotrombone	Telephone	Violinbowed
No.10	Oboe	Violinbowed	Violinbowed	Violinbowed
No.11	Cellobowed	Violinbowed	Telephone	Violinbowed
No.12	Violinbowed	Altotrombone	Violinbowed	Cellobowed
No.13	Percussion	Oboe	Laughter	Violinbowed
No.14	Percussion	Altotrombone	Percussion	Oboe
No.15	Percussion	Violinbowed	Telephone	Violinbowed
Precision	3(0.2)	11(0.73)	3(0.2)	8(0.53)
Recall	0.23	0.85(13 total)	0.06	0.17(47 total)
Query	Female		Telephone	
Method	Musciefish	Feedback	Musciefish	Feedback
No.1	Female	Female	Telephone	Telephone
No.2	Female	Female	Telephone	Telephone
No.3	Water	Female	Telephone	Telephone
No.4	Female	Female	Telephone	Telephone
No.5	Male	Female	Telephone	Telephone
No.6	Male	Male	Telephone	Telephone
No.7	Percussion	Female	Telephone	Telephone
No.8	Female	Female	Oboe	Telephone
No.9	Female	Female	Oboe	Telephone
No.10	Percussion	Female	Telephone	Telephone
No.11	Male	Female	Altotrombone	Telephone
No.12	Male	Female	Telephone	Oboe
No.13	Female	Female	Telephone	Oboe
No.14	Female	Male	Telephone	Telephone
No.15	Female	Male	Altotrombone	Telephone
Precision	8(0.53)	12(0.8)	11(0.73)	13(0.87)
Recall	0.22	0.33(36 total)	0.65	0.76(17 total)

4.5 Discussions

Unlike query-by-keyword used in text retrieval [118], query-by-example is often adopted in content-based audio retrieval. However, when more audio classes can be identified during classification, the recognized classes can be taken as terms and the mixed query of keyword and example can be utilized with existing text retrieval strategies. As the database grows larger and more audio types can be identified, in-depth hierarchical PNN classifiers can be created together with coarse-to-fine searching ability. The hierarchical scheme can be more computationally efficient since it always searches in a smaller sub-database space.

In order to take advantage of user interaction, relevance feedback techniques are proposed for audio retrieval. The feedback is implemented by both updating weights in the similarity distance measurement and updating the query itself. Two relevance feedback algorithms for weight updating in audio retrieval are analyzed. The second feedback algorithm has the advantage of handling negative feedback under the same architecture as positive feedback. The algorithm can also be applied in other multimedia retrieval with feedback. Through relevance feedback, some intelligence or semantics can be added to the retrieval system; thus the gap between high-level semantics and low-level features can be narrowed.

Chapter 5

Audio Classification and Retrieval Using a Fuzzy Inference System

Since the invention of fuzzy sets and the maturing of fuzzy logic theory, fuzzy logic systems have been widely applied to various fields, such as fuzzy controllers, data mining, and so on. Potential applications using fuzzy logic are also being explored with the emergence of other technologies.

In this chapter, we introduce a fuzzy logic approach for hierarchical content-based audio classification and Boolean retrieval, which is intuitive due to the fuzzy nature of human perception of audio, especially audio clips of mixed types. The fuzzy nature of audio search lies in the facts that (1) both the query and target are approximations of the user's memory and desire and (2) exact matching is sometimes impossible or impractical. Therefore, a fuzzy logic system is a natural choice for audio classification and retrieval.

5.1 Introduction

The concept of fuzzy sets was first introduced by Zadeh in 1965 [132] and the analysis of complex systems and decision processes was presented in other paper in 1973 [133]. Since then, research based on both the theory and application of fuzzy logic have been carried out. There also exist some works in audio classification and retrieval using fuzzy logic. Beritelli et al. [134] has proposed a new method for multilevel speech classification based on fuzzy logic. Through simple fuzzy rules, their fuzzy voicing detector achieved a sophisticated speech classification, returning a range of continuous values between extreme classes of voiced/unvoiced. In classification of audio events in broadcast news [59], when fuzzy membership functions associated with features were introduced, the overall accuracy of hard threshold classifier was improved by 4.5% and reached 94.9%. All these related works have demonstrated the capability of fuzzy logic to enhance classification performance and thus provide impetus for us to conduct our research of audio classification and retrieval using fuzzy inference systems.

We adopt a fuzzy logic approach and build a hierarchical classifier named “fuzzy-tree” for content-based audio classification and retrieval. The fuzzy tree classifier is the core of the hierarchical content-based audio classification. At the beginning, audio features are extracted for all audio samples in the database. Proper features are then selected and used as input to a constructed fuzzy inference system (FIS). The outputs of the FIS are two types of hierarchical audio classes. The membership functions and rules are derived from the distributions of the audio features. Speech-plus-music and sound are discriminated by the FIS in the first hierarchy. Secondly, music and speech are separated. One particular sound, the telephone ring, has also been recognized in this level. In

the prototype system, the classification ability of up to fourth level has been explored. Hence we can use multiple FISs to form the ‘fuzzy tree’ for retrieval of different types of audio clips. With this approach, we can classify and retrieve generic audios using fewer features and less computation time, compared to other existing approaches.

During the searching process, we further propose a fuzzy expert system to deal with AND, OR and NOT Boolean queries. In many cases, users cannot provide sufficient examples at hand. However, additional queries can be generated through feedback during the searching process. The multiple query examples can be linked by Boolean operators and treated as a Boolean search problem. In the traditional textual documents retrieval, such a Boolean query is commonly used. With these observed, we propose a scheme to handle Boolean queries in the audio retrieval domain. We construct a framework to solve the three basic Boolean operators known as AND, OR and NOT, with concepts adopted from fuzzy logic. Because of the similarities between Boolean queries and fuzzy logic, the proposed fuzzy expert system can be used to translate the Boolean query for retrieval purposes. Note that, in this chapter, all experiments are conducted on the common database I.

5.2 Fuzzy Inference System

There are several important issues in building a Fuzzy Inference System (FIS), such as selecting the right features as inputs of the system, constructing proper membership functions and rules, and tuning parameters to achieve a better performance.

Figure 5.1: The hierarchical fuzzy tree for audio retrieval.

5.2.1 Selecting Features as Inputs

In order to select appropriate features as inputs to the FIS from the extracted ones, we use a simple nearest neighbor (NN) classifier together with the sequential forward selection (SFS) method to choose the appropriate features. The entire data set is divided into two equal parts for training and testing the NN classifier. Our objective is to use as few features as possible to achieve a reasonable performance. Experiments show that using one or two features selected by SFS is adequate to conduct the hierarchical audio classification at each level. These features are thus chosen as inputs to the FIS. Through experiments, a hierarchical ‘fuzzy-tree’ is constructed by combining up to 4-level FISs together, shown in Figure 5.1. The proper features for each FIS are listed in Table 5.1. Note that, *std* in Table 5.1 refers to ‘standard deviation’. For example, mean spectrogram and mean of the third MFCC coefficient are inputs for sound recognition and mean spectrogram and standard deviation of pitch salience ratio are used for discriminating speech and music.

Table 5.1: Fuzzy classifiers and their input features.

FIS Classifiers	Features
sound and others	mean Spectrogram(Spec)(μ_{Spec}) and mean of third MFCC coefficient(μ_{MFCC_3})
speech and music	mean Spectrogram(Spec)(μ_{Spec}) and std of salience of pitch(SP) (σ_{SP})
telephone ring and others	mean salience of pitch(SP)(σ_{SP})
female and male speech	mean pitch(P)(μ_P) and mean salience of pitch(SP)(μ_{SP})
percussion and others	mean salience of pitch(SP)(μ_{SP}) and mean of first MFCC coefficient(μ_{MFCC_1})
oboe and others	mean zero crossing ratio(ZCR)(μ_{ZCR}) and std of centroid(μ_{ω_c})

5.2.2 Membership Function and Rule Construction

In order to demonstrate the design procedure of fuzzy classifiers, we take the following two examples. One is to classify male and female speech in the third hierarchy and the other is to identify one particular sound, the telephone ring, from a sound collection in the second hierarchy. For the female and male speeches, their two feature-histograms (pitch and pitch salience) are shown in Figures 5.2 and 5.4, respectively. Each histogram is normalized by its peak value. After determining the inputs, the key to constructing the fuzzy classifier is to design the membership functions and extract rules. In practice, the membership functions of each input and output, as well as the rules, can be derived from simulating the feature distributions. We choose Gaussian membership functions in our experiments, whose characteristics are fully parameterized by their means and standard deviations. We calculate these parameters directly from the statistics of the features among the whole data source. We use ‘small’ and ‘large’ to

denote their memberships according to their class distributions. The resulting simplified Gaussian membership functions simulating the feature distributions are shown in Figures 5.3 and 5.5. Another two Gaussian membership functions are chosen for output, shown in Figure 5.6. One of its mean is zero and another is one. Both of these two membership functions have the same standard deviation that makes equal probabilities at the center of their distributions. An overview of the fuzzy classifier for discriminating female and male speech is given in Figure 5.7. For implementation, we adopt the most common methodology, known as Mamdani's fuzzy inference method [135], to build the fuzzy inference system. Mamdani-type inference is a type of fuzzy inference in which the fuzzy sets from the consequent of each rule are combined through the aggregation operator and the resulting fuzzy set is defuzzified to yield the output of the system. The details of the inference methods are listed in Table 5.2.

Table 5.2: The detailed methods used in fuzzy inference system.

The inference method	The implementation method
andMethod:	'prod'
orMethod:	'probor'
defuzzMethod:	'centroid'
impMethod:	'min'
aggMethod:	'max'

The rules in the FIS are listed as below:

- If (*Mean Pitch*) is *small* AND (*Mean Pitch Salience ratio*) is *small*
Then (*Type is male speech*)
- If (*Mean Pitch*) is *large* AND (*Mean Pitch Salience ratio*) is *large* Then
(*Type is female speech*)

Figure 5.2: The feature distribution of mean pitch for female and male.

Figure 5.3: The simulated feature distribution of mean pitch for female and male using Gaussian membership function.

Figure 5.4: The feature distribution of mean pitch salience ratio for female and male.

Figure 5.5: The simulated feature distribution of mean pitch salience ratio for female and male using Gaussian membership function.

Figure 5.6: The Gaussian membership function simulating the output for female and male.

System female&male: 2 inputs, 1 outputs, 2 rules

Figure 5.7: The FIS input-output diagram for female and male classifier.

Figure 5.8: The feature distribution of mean pitch salience ratio for telephone ring and other sounds.

Figure 5.9: The simulated feature distribution of pitch salience ratio for telephone ring and other sounds using Gaussian membership function.

Figure 5.10: The Gaussian membership function simulating the output for telephone ring and other sounds.

System telephone: 1 inputs, 1 outputs, 2 rules

Figure 5.11: The FIS input-output diagram for telephone ring classifier.

Another example is to identify one special sound, the telephone ring among the sound collection. Because the telephone ring sounds different from other sounds, it can be correctly classified with 100% accuracy using one feature, which is the mean pitch salience ratio. Its histogram is shown in Figure 5.8. The simplified Gaussian membership functions simulating the feature distributions is shown in Figure 5.9. The Gaussian membership functions for output is shown in Figure 5.10. The whole FIS for identification of the telephone ring is given in Figure 5.11. The rules for the FIS are as follows:

- If (*Mean Pitch salience ratio*) is *large* Then (*Type is telephone ring*)
- If (*Mean Pitch salience ratio*) is *small* Then (*Type is others*)

Similarly, the rules for the FIS to distinguish music and speech from sound are:

- If (*Mean Spectrogram*) is *small* AND (*mean of the third MFCC coefficient*) is *large* Then (*Type is music and speech*)
- If (*Mean Spectrogram*) is *large* AND (*mean of the third MFCC coefficient*) is *small* Then (*Type is sound*)

The rules for the FIS to classify music and speech are:

- If (*Mean Spectrogram*) is *small* OR (*Standard deviation of pitch salience ratio*) is *small* Then (*Type is music*)
- If (*Mean Spectrogram*) is *large* OR (*Standard deviation of pitch salience ratio*) is *large* Then (*Type is speech*)

The rules for the FIS to identify percussion from non-percussion musical instrumental sounds are:

- If (*Mean Pitch salience ratio*) is *small* AND (*mean of first MFCC coefficient*) is *small* Then (*Type is percussion*)
- If (*Mean Pitch salience ratio*) is *large* AND (*mean of first MFCC coefficient*) is *large* Then (*Type is non-percussion*)

The rules for the FIS to differentiate oboe from music instruments are:

- If (*Mean zero crossing ratio*) is *large* AND (*standard deviation of centroid*) is *small* Then (*Type is oboe*)
- If (*Mean zero crossing ratio*) is *small* AND (*standard deviation of centroid*) is *large* Then (*Type is others*)

Note that, we firstly use both AND and OR connector connecting inputs if the inputs are more than 1. Next, we choose to use the relation with the higher classification accuracy. During experiments, we have found that introducing more inputs doesn't improve performance greatly. In some cases, the performance can even decline. For example, the accuracy of oboe classifier with 1, 2, 3 inputs are 84%, 94% and 93% respectively. Therefore, we design our classifier with restriction to one or two inputs.

5.2.3 Tuning the FIS

Although the fuzzy inference systems are thus constructed completely, there exist ways to improve their performance, such as tuning parameters of those membership functions, choosing different types of membership function other than Gaussian, or using neural networks to train the membership functions for a closer approximation. Since those features selected by the sequential forward

selection method are sub-optimum inputs, we may also try other feature combinations as inputs for a better result.

5.3 FIS Classification and Its Application in Audio Retrieval

5.3.1 FIS Classification Results

In the integrated ‘fuzzy-tree’ classification system, shown previously in Figure 5.1, all classifications can be done hierarchically. For example, when an audio is submitted to the system for classification, the first level FIS can distinguish speech and music from sound. Then, if the result is music and speech, the second level FIS can further tell whether it is music or speech. If the result is sound, the second level FIS can determine whether it belongs to one particular sound, such as telephone ring. Other FIS classifiers can recognize female and male speech, identify percussion and oboe music instruments, etc. With more domain knowledge obtained, we may discover new features and rules which fit for identification of other sounds such as thunder, laughter, applause and so on, or further probe their semantic meanings. In addition, other music types can be recognized from instrument family by studying their vibration characteristics. Experiments have been done hierarchically to obtain the performance of all these fuzzy classifiers. At the first level of the fuzzy tree, each audio file is used as input to the fuzzy classifier. Starting from the second level of the fuzzy tree, the experiments are conducted based on corresponding subsets of the audio files. For example, 352 speech plus music examples are submitted to

music and speech classifier, 62 sounds are submitted to telephone ring detector. Further, 53 speech files, 299 and 197 music files are tested for female and male classifier, percussion detector and oboe detector respectively. The percussion is firstly distinguished from the rest of music instrument because of its inharmonic nature. All classification results are summarized in Table 5.3.

Table 5.3: The performance of Fuzzy Inference System (FIS) for audio classification.

FIS classifier	Classification Accuracies
Sound and Others	80%
Music-Speech	92%
Telephone ring	100%
Female-male	89%
Percussion	81%
Oboe	94%

5.3.2 Fuzzy-tree for Audio Retrieval

The fuzzy-tree architecture as shown previously can be helpful for retrieval. When a query input is presented, the direct search may retrieve audio clips in various classes. If we firstly classify the query into one particular node of the fuzzy tree, we can then search relevant files only in that subspace instead of the whole database. For example, various audio files can appear in the search results of a speech query. If we can firstly classify the query into one subset like the speech or music category, many non-relevant sounds can be discarded before the search begins. Thus, the precision will increase and the searching time will decrease. If the classification is wrong, we can search in other branches of the fuzzy tree with the user's feedback. When the database grows, new classes can be added to the tree. Only links between the new class and its immediate upper

level need to be updated, with the rest of the tree unchanged.

5.4 Boolean Query Using Fuzzy Logic

As for retrieval, the existing content-based audio retrieval systems usually adopt the query-by-example mechanism to search for desired audio files. However, only one single audio sample often cannot express the user's needs adequately. To overcome this problem, more audio files can be chosen as queries provided by the user or through feedback during the searching process. Correspondingly, we present a novel scheme to handle content-based audio retrieval with multi-queries connected by Boolean operators. We build a framework to solve the three basic Boolean operators known as AND, OR, and NOT, with concepts adopted from fuzzy logic. Experiments have shown that Boolean search can be very helpful in audio retrieval.

5.4.1 The Fuzzy Logic and Boolean Query

It is usually insufficient to use one query to express the user's retrieval needs. Therefore, the Boolean query can represent user's request more adequately by combining multiple query examples. A Boolean query has a syntax composed of query examples and Boolean operators. The most commonly used operators, given two basic queries q_1 and q_2 , are as follows:

- AND: the query (q_1 AND q_2) selects all documents which satisfy both q_1 and q_2 .
- OR: the query (q_1 OR q_2) selects all documents which satisfy either q_1

or q_2 .

- NOT: the query (q_1 AND (NOT q_2)) selects all documents which satisfy q_1 but not q_2 . In case (NOT q_2), all documents not satisfying q_2 should be delivered, which may retrieve a huge amount of files and is probably not what the user wants.

These three Boolean operations directly correspond with intersection, union and complement operations in a fuzzy logic system. Let U be a universal set, A and B be two fuzzy subsets of U , and $\bar{A}$ be the complement of A relative to U , and u be an element of U . These operations can be defined as:

- Intersection: $\mu_{(A) \cap (B)} = \min(\mu_A(u), \mu_B(u))$
- Union: $\mu_{(A) \cup (B)} = \max(\mu_A(u), \mu_B(u))$
- Complement: $\mu_{\bar{A}}(u) = 1 - \mu_A(u)$

where $\mu(\cdot)$ is the membership function.

5.4.2 Similarity Measurement and Membership Function

In order to utilize the fuzzy expert system for audio retrieval, we first calculate the Euclidean distance between the query and samples in the database and then define the similarity measurement as follows:

$$Dist(q, d_i) = \sqrt{\sum_{j=1}^N (q_j - d_{ij})^2} \quad (5.1)$$

$$Sim(q, d_i) = \frac{1}{Dist(q, d_i) + 1} \quad (5.2)$$

where q is the query audio feature vector, d_i is the feature vector of the i th file in the database. The q_j is the j th element of the feature vector q and d_{ij} is the j th element of the feature vector d_i . Since the distance $Dist(q, d_i)$ ranges from $[0, \infty)$, thus, the similarity $Sim(q, d_i)$ ranges from $(0, 1]$.

A membership function (MF) is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The function itself can be an arbitrary curve and the only condition a MF must really satisfy is that it must vary between 0 and 1. Here, we use the similarity measurement $Sim(q, d_i)$ as the MF to denote how “similar” the i th file d_i in the database to the query q . If they are identical, their distance is zero and hence the “similar” degree is one. If their distance approaches positive infinite, the “similar” degree will approach zero. For any file in the database, we can use the MF to find “similar” degree to a membership value between 0 and 1.

5.4.3 The Fuzzy Expert System for Boolean Query

In order to use fuzzy logic for Boolean retrieval, we need first to decompose the Boolean query into several single queries. For each single query, we retrieve a fuzzy set of files, and each of them is associated with the fuzzy set by a membership function as defined is in Eq. (5.2). Then, we perform the fuzzy set operations for the obtained fuzzy sets corresponding to all the single queries. The resulting final fuzzy set contains the files corresponding to the Boolean

query. The fuzzy operations “Intersection”, “Union”, and “Complement” are used for AND, OR, and NOT Boolean operations respectively.

In this way, the retrieval result for a Boolean query can be generated according to this combined “similar” membership function. The method of the decomposition is that we group the AND, OR and NOT logic together. During the calculation, the AND operation has the highest priority, and should be processed first. Then, the OR operation follows and is combined with the result of the AND operation. Finally, the NOT operation is merged with the result of the OR operation using the AND operation to form the final similarity for sorting.

For example, suppose a general Boolean query is proposed for searching audio files as follows:

“Find documents which sound similar to (((q1 and q2) or q3) and (not q4))”

Firstly, we decompose the Boolean query into 4 individual single queries sequentially from left to right:

- Query 1: Find documents which sound similar to q1
- Query 2: Find documents which sound similar to q2
- Query 3: Find documents which sound similar to q3
- Query 4: Find documents which sound similar to q4

Next, we derive the membership function for each query based on similarity measurement defined in Section 5.4.2. Then, we adopt the fuzzy operation “Intersection” to compute the combined MF for Query (q1 and q2) using

$\min(\text{Sim}(q1, d_i), \text{Sim}(q2, d_i))$. Further, we adopt the fuzzy operation “Union” to compute the combined MF for ((q1 and q2) or q3) using $\max(\min(\text{Sim}(q1, d_i), \text{Sim}(q2, d_i)), \text{Sim}(q3, d_i))$. The “Complement” fuzzy operation is adopted to calculate the MF for query (not q4) and merged with the resulting MF of ((q1 and q2) or q3) using the fuzzy operation “Intersection” to produce the final MF for query (((q1 and q2) or q3) and (not q4)), which is $\min(\max(\min(\text{Sim}(q1, d_i), \text{Sim}(q2, d_i)), \text{Sim}(q3, d_i)), (1 - \text{Sim}(q4, d_i)))$.

5.5 Boolean Query Experiments

We conducted Boolean query experiments on two databases. The first one is the common database I described earlier. The second database consists of samples from seven classes listed in Table 5.4. All the files of the second database are extracted clips of several seconds in length from the movie ‘Titanic’ with a total number of 116. During the experiments, single query and Boolean queries are submitted to search for interested audio documents. The performance is evaluated and compared between single queries and Boolean queries. Note that, the audio class “mixture of all” means that the audio clip in this class contains all of the speech, music and sound contents simultaneously.

5.5.1 Experiments on AND Boolean Queries from the Same Class

The AND Boolean operator is the most frequently used in Boolean search, which normally links two examples from the same class to give a more precise representation of the user’s query. This experiment is conducted on the first

Table 5.4: The ‘Titanic’ audio database for Boolean query.

Audio class	Number of files
Music	10
Speech	32
Sound	12
Speech-music mixed	16
Speech-sound mixed	18
Sound-music mixed	14
Mixture of all	14
Total	116

database. The result of one example of such an AND Boolean query against the results of the two individual queries are shown in Table 5.5 and Figure 5.12. Because normally users only browse files listed at the top, which is often the first screen of the search result, we only list the 15 top ranked files in Table 5.5. In Table 5.5, q1 and q2 are both audio files from class Violinbowed. From Table 5.5, we can see that by using a single query, there are 8 or 9 relevant files retrieved from the Violinbowed class. By using the AND query formed by the two examples, there are 12 relevant retrieved. The average precision of the three queries are 0.32, 0.35 and 0.39 respectively. The recall and precision curve in Figure 5.12 reveals that the AND query performs generally better than the two individual queries.

5.5.2 Experiments on AND Boolean Queries from Different Classes

Likewise, the AND Boolean operator can also link two queries from different classes. The result of one example of such an AND Boolean query against the results of the two individual queries are shown in Table 5.6. This experiment

Table 5.5: The retrieval result of AND Boolean queries from same class.

Rank	q1	q2	q1 and q2
1	Violinbowed	Violinbowed	Violinbowed
2	Violinbowed	Violinbowed	Violinbowed
3	Violinbowed	Violinbowed	Cellobowed
4	Cellobowed	Cellobowed	Violinbowed
5	Altrotrombone	Violinbowed	Violinbowed
6	Violinbowed	Violinbowed	Violinbowed
7	Violinbowed	Violinbowed	Altrotrombone
8	Cellobowed	Violinbowed	Violinbowed
9	Cellobowed	Oboe	Violinbowed
10	Oboe	Cellobowed	Violinbowed
11	Violinbowed	Cellobowed	Violinbowed
12	Violinbowed	Violinbowed	Violinbowed
13	Violinpizz	Violinbowed	Violinbowed
14	Altrotrombone	Cellobowed	Violinbowed
15	Violinbowed	Oboe	Oboe
Relevant	8	9	12

is conducted on the second database. Here, we list the first 10 ranked files. In Table 5.6, q1 and q2 are audio files from class speech and class sound respectively. It is shown that if the two examples linked by the AND operator are from different audio classes, the retrieved samples of the AND Boolean query contain both characteristics of the two classes. This is because in the AND Boolean operation, the smaller MF value will be the resulting MF value. Thus only files similar to both of the two query examples will appear at the top.

5.5.3 Experiments on OR Boolean Queries

In the third experiment, we test the result of OR Boolean query linking female and male speech based on the first database again. The result in Table 5.7 shows that samples from either query class could appear in the OR Boolean

Figure 5.12: Precision-recall curve of AND Boolean query.

query. This is due to the fact that in OR query, the files similar to any of the two query examples may rank at the top.

5.5.4 Experiments on Mixed Boolean Queries

A mixed Boolean query which contains both AND and OR operators from different classes is shown in Table 5.8. This experiment is conducted on the second database again. The q1, q2, and q3 are from speech, music, and sound, respectively. From Table 5.8, we can see that samples which satisfy AND query are found first, then are merged with the OR Boolean query to obtain the final results. Note that, in the NOT query, such as q1 NOT q2, the MF for query (NOT q2) at sample q2 is zero and zero with any “intersection” operation is always zero. So, the undesired sample q2 can be moved to the bottom in the retrieved list by the NOT Boolean operation.

Table 5.6: The retrieval result of AND Boolean queries from different classes.

Rank	q1	q2	q1 and q2
1	Speech	Sound	Speech-sound
2	Speech	Sound	Speech-sound
3	Speech-sound	Sound	Speech-sound
4	Speech	Sound-music	Speech-sound
5	Speech	Sound	Sound-music
6	Speech	Sound-music	Speech-sound
7	Speech	Speech-music	Speech-sound
8	Speech	Sound-music	Speech-sound
9	Speech	Sound	Speech-sound
10	Speech	Music	Mixture

5.6 Discussions

In this chapter, we propose a fuzzy inference system for audio classification and retrieval, which is one step further toward a multimedia search engine for the Internet. The benefits of the fuzzy classifier lie in the facts that no further training is needed once the fuzzy inference system is constructed. Thus, classification can be performed very quickly. In addition, when the database grows, new classes can be added to the fuzzy tree. Only links between the new class and its immediate upper level are required to be updated, with the rest of the tree unchanged. With this architecture, fast online web applications can be built. Future work along this direction is to use neural networks to train the parameters to obtain more accurate membership functions, and to explore new features and rules to classify various audios with the so-called ‘fuzzy tree’ for hierarchical retrieval.

Furthermore, we put forward a general method based on a fuzzy expert system to handle Boolean query in audio retrieval. The Boolean query can be

Table 5.7: The retrieval result of OR Boolean queries.

Rank	q1	q2	q1 or q2
1	Female	Male	Female
2	Female	Male	Male
3	Male	Female	Female
4	Male	Female	Male
5	Female	Female	Male
6	Female	Machines	Female
7	Male	Animal	Female
8	Female	Female	Female
9	Female	Female	Machines
10	Female	Female	Male

used in both direct Boolean search and user's feedback. In this way, we can not only improve the retrieval performance but also enhance searching ability. The Boolean search algorithm can be used in image and video retrieval, as well as for the user feedback occasions.

Table 5.8: The retrieval result of AND, OR mixed Boolean queries.

Rank	q1	q2	q3	q1 and q2	q1 and q2 or q3
1	Speech	Music	Sound	speech	Sound
2	Speech	Music	Sound	Speech-music	Sound
3	Speech	Sound-Music	Speech-music	Speech-music	Speech-music
4	Speech	Mixture	Sound-music	Speech-music	Sound-music
5	Speech	Music	Sound	Music	Sound
6	Speech	Sound	Sound	Sound	Sound
7	Speech	Sound	Sound-music	Speech-sound	Speech
8	Speech	Speech-music	Sound-music	Mixture	Sound-music
9	Speech	Music	Sound	Speech-sound	Speech-music
10	Speech	Sound	Sound	Speech-Sound	Sound-music

Chapter 6

Audio-visual

Anchorperson-based News Video

Story Segmentation

Audio information can be utilized in a number of ways to assist its associated video signal processing. In this chapter, we present a novel method of audio-visual anchorperson-based story segmentation for news video. The process can be divided into three stages. The first stage is called audio segmentation, where the news video is split into small segments based on the ratio of one audio feature called total spectral flux. In the second stage, each of these small segments is classified into speech or non-speech segment based on additional audio features. Adjacent non-speech segments are then merged into longer clips. The third stage is called anchorperson detection, where anchor people are detected in the key frames near the speech segment boundaries. The speech segments between two anchorpersons are combined together with the leading anchorper-

son segment and formed into a story. Thus, news video can be segmented into stories based on anchorperson with simple computation. Experiments are carried out on several one half hour long noon news programs. The news program is broadcasted daily by China Central TeleVision (CCTV) from 12:00PM to 12:30PM. Several video files of the news program have been used in our experiments. Similar encouraging results are obtained with all the video files. The experimental results presented in this thesis are based on the news broadcasted on November 19th, 2002 as an example.

6.1 Audio-visual Based News Video Story Segmentation

Temporal audio segmentation is often a primitive process that segments long audio files into uniform clips. A segment boundary is set when there is an abrupt change detected in the analyzed contents or extracted features. Lu et al. [136] presented a study of audio content analysis for classification and segmentation, in which an audio stream was segmented according to audio type or speaker identity. A robust approach that was capable of classifying and segmenting an audio stream into speech, music, environment sound, and silence was then proposed. Pfeiffer illustrated the determination of temporal audio segmentations at different semantic levels using perceptual pause concept in [137].

As most video is broadcasting with its associated audio, it is a natural way to utilize this additional information to assist video processing. Audio as a support to scene change detection and characterization of video sequences was discussed in [138], where a general framework that used the results of audio

classification was then proposed for organizing video information. In many cases, the audio and video information are combined together and used for video segmentation. Nam et al. [139] demonstrated such a new approach for video sequence segmentation into individual shots, while Sundaram et al. [140] modelled a video scene as a semantically consistent chunk of audio-visual data for segmentation. News video as one typical video sequence is widely available in TV programs and worthy of study. The issue of story segmentation, including commercials labelled as separate stories, was addressed and a fully automatic system that could extract story boundaries using available audio, video and closed-captioning cues was clarified in [141]. In this chapter, we propose a different method for audio-visual based news video story segmentation.

6.2 News Video Segmentation Procedure

6.2.1 Overview of the Process

Normally, a news program can be described by a sequence of news reports and commercials interlaced sequentially. News segments can often be further divided into anchorperson speeches and field reports. The process of our proposed audio-visual based story segmentation for news video can be separated into three steps shown in Figure 6.1, which are audio segmentation, speech/non-speech discrimination, and anchorperson detection. In this way, news video can be segmented into stories with leading anchorperson and commercials, where the logo video of the TV channel is also regarded as a special commercial by simple computation.

There are several methods adopted to make the computation fast. Firstly,

Figure 6.1: Overview of the story segmentation process.

we perform audio segmentation using total spectral flux (TSF) based on a one-second-window instead of spectral flux (SF) based on one-frame-window. Secondly, after segmentation, we select the first 3 seconds videos of each segment to represent the whole segment. Thirdly, we extract two key frames one second before and after each speech segment boundaries for anchorperson detection. All of these can greatly reduce the computational overhead, yet the original information is preserved.

6.2.2 Audio Segmentation Based on Spectral Flux

Audio segmentation is first conducted on the audio signal associated with the news video. This is a coarse segmentation, where the video is split into small segments based on the ratio of Total Spectral Flux (TSF). The Spectral Flux (SF) is used to determine changes of spectral energy distribution of two successive windows and originally defined as the 2-norm of the difference between the magnitude of the Short Time Fourier Transform (STFT) spectrum evaluated

at two successive sound frames, where the STFT is normalized in energy [61]. Here, we define the Total Spectral Flux (TSF) as the summation of Spectral Flux of all the frames within one second. For each frame, Spectral Flux is computed using the 1-norm of difference between the magnitude of the Discrete Fourier Transform (DFT) spectrum evaluated at two successive sound frames. The SF at m th frame of i th one-second-window is computed as follows:

$$SF(i, m) = \sum_{j=1}^K |F_i(m, j) - F_i(m-1, j)| \quad (6.1)$$

where $F_i(m, j)$ is the absolute value of j th subband of the DFT spectrum at m th frame of i th one-second-window, K is the number of DFT points. Then, the TSF is computed as below:

$$TSF(i) = \sum_{m=1}^{N-1} SF(i, m) \quad (6.2)$$

where N is the number of frames in one second window. The number of audio frames per second is same as the number of video frames, which are both 25. The $TSF(i)$ is the total spectral flux of the i th second window. The i th TSF ratio is defined as the ratio between the current TSF and the average of its previous L $TSFs$ as follows:

$$Ratio(i) = \frac{TSF(i)}{\frac{1}{L} \sum_{j=L-5}^{i-1} TSF(j)} \quad (6.3)$$

The first L TSF ratios are set to 1. If $Ratio(i)$ is either higher than $T1$ or lower than $T2$, a segment boundary is determined, where $T1$ and $T2$ are fixed thresholds. In our experiment, L is chosen to be 5, as five-second length is a moderate length for capturing the TSF dynamics, though other lengths are

also possible. We have observed the statistics of $Ratio(i)$ over the segment boundaries. Seen from Figure 6.2, we have empirically set the two threshold parameters T1 and T2 to be 2 and 0.2 respectively. T1 is used to detect TSF changes from slow to fast while T2 is used to detect TSF changes from fast to slow. Since TSF ratio is normalized by surrounding TSFs, it can adapt to audio dynamics. The TSF ratio of one part of audio associated with news video is shown in Figure 6.2. The vertical lines represent the segment boundaries.

Figure 6.2: The Total Spectral Flux (TSF) ratio of part of audio streams.

6.2.3 Speech/non-speech Discrimination

After the first step coarse level audio segmentation, each of these segments is represented by its first 3 seconds to reduce computational load. If the segment is shorter than 3 seconds, whole segment is selected. Experiments have shown that three seconds is long enough to describe the homogeneous segments for human perception. Then, speech/non-speech discrimination is conducted on those representative segments using a nearest neighbor classifier based on ten additional frame-level audio features. We have randomly extracted 10 clips of speech, 20 clips of commercials (including 10 logos) from the original audio associated with

news video as training samples. The length of these samples ranges several seconds. Ten features are extracted for each clip, namely volume dynamic ratio (VDR), silence ratio, means and standard deviations of root mean square (RMS, μ_{RMS}, σ_{RMS}), zero crossing ratio (ZCR, μ_{ZCR}, σ_{ZCR}), pitch frequency (μ_P, σ_P) and salience of pitch (μ_{SP}, σ_{SP}). Each audio feature is normalized over all files of the training samples by subtracting its mean and dividing by its standard deviation. The audio features extracted from the representative segments are also normalized before classification. After discrimination, adjacent non-speech segments are merged to form one longer commercial.

6.2.4 Anchorperson Detection

For each speech segment, we select two key frames located one second before and after the segment boundary, to detect whether an anchorperson appears in these frames. This is because the frames at boundaries may contain transitional or mixed frames that may cause errors of judgment. Anchorperson detection is then conducted on these key frames. Field reports followed by the anchorperson are merged together into one full news item. Thus, news video can be segmented into stories with simple computation according to anchor people.

In the video key frames, the 24-bit RGB color histograms are counted in the R,G,B channels separately. Each color has eight bits indicating one of 256 color intensities. Twenty-five bins with bin centers from 10 to 250 with step-increase of 10 are used for histogram calculation. The RGB color histograms are defined as $r_i, g_i, b_i, i = 1, 2, \dots, 25$. They are further normalized by their maximum

color bin values in each color channel as below:

$$R_i = \frac{r_i}{\max(r_i)} \quad (6.4)$$

$$B_i = \frac{b_i}{\max(b_i)} \quad (6.5)$$

$$G_i = \frac{g_i}{\max(g_i)} \quad (6.6)$$

Then, these R,G,B color histograms are linked together and 75 bins combined color histograms ($CCH(i), i = 1, 2, \dots, 75$) are thus formed and used as features for anchorperson detection.

Two anchor people appear in our news video program. One is a female and the other is a male. They can appear in different frames or the same frame simultaneously. Some typical anchorperson video frames are shown in Figure 6.3. The CCHs of those anchorperson frames containing female, male and both are shown in Figures 6.4, 6.5, and 6.6. Some typical non-anchorperson video frames such as logo, field report and commercial are shown in Figure 6.7. Their corresponding CCHs are illustrated in Figures 6.8, 6.9, and 6.10.

(a) Female anchorperson frame

(b) Male anchorperson frame

(c) Both anchorpersons frame

Figure 6.3: The anchorperson sample frames.

Figure 6.4: The combined color histograms of female anchorperson frame.

Figure 6.5: The combined color histograms of male anchorperson frame.

Figure 6.6: The combined color histograms of both female and male anchorperson frame.

(a) Title frame

(b) Field report frame

(c) Advertisement frame

Figure 6.7: The non-anchorperson sample frames.

Figure 6.8: The combined color histograms of logo frame.

Figure 6.9: The combined color histograms of field report frame.

Figure 6.10: The combined color histograms of advertisement frame.

For anchor person detection, in each of these three kinds of anchorperson frames, we pick up two frames as samples. This sample selection can also proceed automatically according to the facts that each anchorperson frame appears in the whole news video at different times and occupy a certain interval (usually longer than 3 seconds) in each appearance continuously. Thus, we can find the sample anchorperson frames as follows:

1. Pick up the 1st frame in every 3 seconds as testing frame in the news video
2. Find the locations of adjacent testing frames whose Euclidean distances of their CCH are smaller than threshold T_3 , which is set to 0.75
3. Calculate the Euclidean distances of the CCH between the first location and the rest of other found locations
4. If there exist many distances (such as more than 3) that are smaller than T_3 , the testing frame at the first location would most likely to be an anchorperson frame thus a sample frame is identified
5. Continue to check the frames at second location to the last found location using the same judgement or stop when there are 2 samples from each kind of anchorperson frames are picked up

Once we have samples, the Euclidean distances of CCH between the key frame and sample frames are calculated. If any of the distances are below threshold T_3 , an anchorperson is detected in this key frame. The segment, which contains this key frame, is believed to be an anchorperson segment.

6.3 Experiments

6.3.1 Experiments Setup

We have carried out our experiments on several one half hour CCTV noon news programs. This program broadcasts daily with an organized structure. Similar experimental results are obtained and the news program broadcasted on November 19th, 2002 is chosen as an example for segmentation. We first manually set the segment boundary of the news program with the resolution of one second. This manually produced reference segmentation boundaries may have inaccurate points because of human interaction. Such inaccuracies should be tolerable. Taking this into consideration, a hit is registered whenever a calculated boundary can be found within a two seconds long interval with the reference boundary point as center. A missed segment is a reference segment that has no counterpart in the calculated segmentation and a false segment is an excessive segment in the calculated segmentation. Hence, the overall segmentation performance of our first stage audio segmentation are measured by hit ratio, missed ratio and false ratio given below:

$$Hit = \frac{Hseg}{Tref} \quad (6.7)$$

$$Missed = \frac{Tref - Hseg}{Tref} \quad (6.8)$$

$$False = \frac{Tcal - Hseg}{Tcal} \quad (6.9)$$

where $Hseg$ is the number of hits between the reference and the calculated segmentation, $Tref$ is the number of total reference boundary points, and $Tcal$

is the number of total calculated boundary points. Note that the sum of *Hit* rate and *Missed* rate equals to one.

6.3.2 Experimental Results

There are 17 news stories with leading anchorperson(s) in the half hour news video shown in Table 6.1. The 3-step news video story segmentation results are given in Table 6.2. In the first step called audio segmentation, there are 114 segment boundaries detected by human perception. These boundaries occur where there are pauses or sudden change of audio spectra. There are 116 boundaries detected in the calculated segments with 72 hits. The accuracy of the second step named speech/non-speech discrimination is 99%, only one non-speech segment is classified as a speech segment, which is part of a commercial with a mixture of speech and background music. After the third step called anchorperson detection, the identified adjacent anchorperson segments are merged, where there are pauses detected during anchorperson speech across a segment boundary. The anchorperson detection can also correct the errors made in classification. For example, if a segment is classified as a speech after non-speech segments without the leading anchorperson being detected, this segment can be reclassified to non-speech. Although the hit ratio is 63%, after the final processing, all of the story segments are correctly detected.

6.4 Discussions

In this chapter, we present a three-step approach for news video segmentation into story items. First, the news video is divided into segments based on the

Table 6.1: The manual segmentation results of the news video.

Time(s)	Contents	Time	Contents
Start-8	Logo	961-1001	News 10
8-53	Commercials	1001-1009	Logo
53-97	Logo	1009-1073	News 11
97-179	News 1	1073-1165	News 12
179-344	News 2	1165-1228	News 13
344-400	News 3	1228-1294	News 14
400-480	News 4	1294-1429	News 15
480-654	News 5	1429-1487	News 16
654-660	Logo	1487-1601	News 17
660-757	News 6	1601-1609	Commercials
757-824	News 7	1609-1760	Logo
824-879	News 8	1760-1767	News summary
879-961	News 9	1767-1786	Music

Table 6.2: The automatic story segmentation results of the news video.

Hseg	72	Hit ratio	0.63
Tref	114	Missed ratio	0.37
Tcal	116	False ratio	0.38
The speech/non-speech discrimination accuracy			99%
Story segmentation accuracy			100%

total spectral flux ratio of its associated audio. Then, the segments are classified as speech/non-speech segments. Next, the speech segments are scanned for anchorpersons to create “story segments”. It can be used to set the story boundary of the news program automatically. Experiment results are shown for one typical half hour news video. Similar results are also observed for other news videos of the same broadcast program. The preliminary performance give satisfying results. The speech recognition and speaker identifications can also benefit from this segmentation. With simple modification, the process can also handle streaming video as well. The development of the algorithm will be able

to organize video information according to semantic content of the media.

Chapter 7

Conclusions and Recommendations

7.1 Conclusions

With the objective to facilitate the automatic classification and retrieval of audio documents by their contents, and ease the manual labor involved through the process, we have discussed, in this thesis, the techniques involved in audio search engine development and carried out experiments for performance improvement from statistical and soft computing approaches.

After a comprehensive literature review on content-based audio signal processing, we focus our discussion on a sequence of technical issues regarding to content-based audio classification and retrieval.

Firstly, we extract and normalize audio features from time, frequency, and coefficient domains to represent the audio files. Then, we adopt the sequential feature selection method together with corresponding classifiers for audio clas-

sification. Statistical nearest neighbor (NN), k-nearest neighbor (k-NN), and Gaussian mixture model (GMM) classifiers and probabilistic neural networks (PNN) have been employed in our audio classification experiments. Among them, our modified k-nearest neighbor (K-NN) classifier using 28 features yields 93.7% accuracy in classifying audio into 16 classes, as compared to that of 90.4% by nearest feature line (NFL) using 34 features, tested on a common database. In addition, the statistical NN, k-NN and GMM are also adopted in the music instrumental classification with satisfactory results based on a small set of features selected.

In audio retrieval, we combine the PNN classifier with the Euclidean distance measurement for hierarchical retrieval using query-by-example. It has been shown that the performance of the proposed probabilistic neural network (PNN) hierarchical retrieval scheme can achieve better overall performance than a simple distance method in both the recall-precision relation and average precision. In order to improve performance and take advantage of user interactions, we have proposed and compared two relevance feedback algorithms. The proposed method based on the constrained optimization has the capability of handling both positive and negative feedback in a unified framework.

As there is a similarity in nature between fuzzy logic and the user's ambiguous perception of audio, we find that the fuzzy inference system can be utilized for the task of audio classification and retrieval. A hierarchical Fuzzy Inference System (FIS) has been proposed to classify audios from coarse classes to detailed classes. For each fuzzy classifier of FIS, as few as 2 features are utilized as inputs. The effectiveness and efficiency of using fuzzy logic are thus demonstrated by constructing a fuzzy-tree for hierarchical audio classification. The

so-called hierarchical “fuzzy tree” can be employed for top-down audio search as well. The audio Boolean retrieval for multiple queries connected by AND, OR, NOT and mixed Boolean logic is also investigated to further enhance the searching ability.

Finally, an audio-aided content based video processing technique, especially for the news video story segmentation, is proposed and demonstrated. The coarse level audio segmentation with very low computational cost can help to make the video segmentation easier. Classification of speech and non-speech audios has been applied to the story segmentation of news video. News video can be segmented to an individual ‘story item’ based on anchorperson detection. Therefore the story boundaries in news programs can be detected automatically. Experimental results show that very high accuracy of audio-aided news story segmentation can be achieved.

7.2 Recommendations for Further Research

In order to fully realize the vision of audio classification, retrieval and its assistance in affiliated multimedia processing, there are several topics recommended for further research.

The first one is to explore new audio features which are domain dependent for a particular application such as classification of designated musical instruments or find ways to accurately represent high level semantic features such as Timbre. Algorithms to extract new features for better audio understanding, especially psychoacoustics features, are expected. Some perceptual features such as pitch and timbre are particularly important and thus reliable and robust ex-

traction of such features must be thoroughly investigated. New ways to combine various features to form a better feature vector with feature correlation taken into consideration can be further explored. Only based on the right features can we achieve a satisfactory classification result or make significant accuracy improvements of the classifier without much effort. Thus, the tree-structure hierarchical audio classification and retrieving can be made possible.

The second one is to learn how to improve the audio classification accuracy with the advancement of pattern recognition techniques and enhance the retrieve performance and capabilities through various approaches. It is always desirable to find efficient ways to further improve the classification accuracy. For examples, we can utilize the latest machine learning techniques such as the Support Vector Machine (SVM) [142], Relevance Vector Machine (RVM) [143] for audio classification purpose. As for the enhancement of retrieval ability, we can further improve the relevance feedback techniques, Boolean query processing and so on. With the ability to identify more individual audio classes as ‘audio keywords’, the text retrieval strategy can be applied, such as searching in combination with traditional keyword query and query-by-example. Furthermore, with the user profile provided, we can build a tailor-made retrieval system to fit for individual customized search engine for personalized usage.

The third topic is to find new applications for content-based audio signal processing such as audio information filtering, integrated audio-visual interaction for multimedia processing, or audio ontology utilization in the fast growing Internet, digital libraries, MPEG-7 and MPEG-21 scenarios.

MPEG-7 is an ISO standard that facilitates searching for media content much as current text-based search engines for retrieval of HTML content. The

goal of this standard is to develop a rich set of standardized tools to enable both humans and machines to generate and understand audiovisual descriptions which can be used to enable fast and efficient retrieval from digital archives as well as filtering of streamed audiovisual broadcasts on the Internet. Quackenbush and Lindsay gave an overview of the MPEG-7 audio standard, in terms of the applications it might support, its structure, the process by which it was developed, and its specific descriptors and description schemes [144]. The XML Schema specifications of the audio descriptors were described in detail in [145].

MPEG-21, the newest standard in MPEG family, will fill the gaps and allow existing components to be used together, thereby increasing interoperability. MPEG-21 defines a multimedia framework to enable transparent use of multimedia resources across a wide range of networks and devices used by different communities. For a detailed examination and description of the requirements for the MPEG-21 multimedia framework, readers can refer to the MPEG-21 technical report, “Vision, Technologies and Strategy” [146]. The current version of the standard can be found on the MPEG home page [147].

The fundamental concept in MPEG-21 framework is the ‘digital item’, a structured digital object with a standard representation, identification and metadata. The digital item includes three parts, resource (individual asset), metadata (data about or pertaining to the item), and structure (relationships among the parts of the item), shown in Figure 7.1. For example, a digital item called “audio music album” can have its mp3 songs (possibly encoded in different bit-rates) as its resource; text lyrics, artist biography and intellectual property information as its metadata; and links to web page to purchase the album as its structure.

Figure 7.1: The ‘digital item’ structure.

Based on increasing common audio concepts and semantics with high-level meanings discovered and agreed upon, it is possible to create an audio ontology to provide a universally acceptable and machine-processable audio retrieval system, which can be integrated in the semantic web.

Using the available audio ontology and the concept of “digital item”, we can design a preliminary prototype system based on the so-called “audio item” for semantic audio retrieval. Here, we illustrate a framework for audio retrieval using the concept of “audio item” with ontology and all the available content-based audio processing techniques, shown in Figure 7.2. The audio item is built by segmentation, classification (speech, music and sound discrimination), and different further treatment depending on the classification result. For example, the speech recognition and word spotting (identification of keywords) can be conducted for speech, while music classification such as instrument recognition

can be carried out for music, and sound spotting (identification of predefined sound classes) can be performed for sound with necessary user annotation when applicable. Based on all these processings, an audio item is constructed including (1) raw audio file and extracted features as its resource; (2) proper ID (including URL and Title), the start and end time obtained from segmentation, descriptions obtained from the corresponding classification procedure, manually added copyright information as its metadata; and (3) the audio ontology as its structure. Note that the steps in dashline blocks need user interactions. During the audio item retrieval process, the search engine can utilize the power of both text search ability based on its structure and metadata and content-based query-by-example search ability based on its resource. For example, the search engine can go through text search based on query-by-ontology to see whether there is a match. Then, it can perform content-based similarity distance calculation in that category instead of the whole database to retrieve audio files to meet with user’s requirements.

Figure 7.2: The framework for audio retrieval using MPEG-21 concept “audio item” with ontology.

Future content-based audio retrieval systems should be adaptable to provide

access to any content or semantic concept such as “happy music”, not just a few simple concepts predefined by the system. This way, more intelligence can be added towards building an intelligent audio retrieval systems with adaption to rapidly increasing web contents. The ultimate goal of content-based audio processing is to manage audio files in a similar way to text documents, where user can not only hear them at ease but also “read” them inside out.

Author's Publications

1. Mingchun Liu and Chunru Wan and Lipo Wang, "Intelligent Content-Based Audio Classification and Retrieval for Web Applications", in *Book, Edited by Yan-Qing Zhang, Abraham Kandel, T.Y. Lin and Y.Y. Yao, "Computational Web Intelligence: Intelligent Technology for Web Applications", ISBN 981-238-827-3, World Scientific, pp: 257-281, 2004.*
2. Mingchun Liu and Chunru Wan and Lipo Wang, "A Fuzzy Logic Approach for Content-Based Audio Classification and Boolean Retrieval", *an invited chapter to the book: "Fuzzy Logic and the Internet", Loia, Vincenzo; Nikraves, Masoud; Zadeh, Lotfi A. (Eds.), ISBN: 3-540-20180-7, Physica-Verlag, Springer, pp: 135-156, 2004.*
3. Mingchun Liu and Chunru Wan and Lipo Wang, "Content-Based Audio Classification and Retrieval Using A Fuzzy Logic System: Towards Multimedia Search Engines", *Journal of Soft Computing, Springer, Vol. 6 (2002), Issue 5, pp: 357-364.*
4. Chunru Wan and Mingchun Liu, "Content-Based Audio Retrieval With Relevance Feedback", *Submitted to Pattern Recognition Letters, Springer.*
5. Mingchun Liu and Chunru Wan, "Weight Updating for Relevance Feedback in Audio Retrieval", In *Proceedings of the 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.
6. Mingchun Liu and Chunru Wan, "Boolean Search for Content-based Audio Retrieval Using Fuzzy Logic", In *Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'02)*, 2002.
7. Mingchun Liu and Chunru Wan, "Feature Selection for Automatic Classification of Musical Instrument Sounds", In *Proceedings of the ACM/IEEE Joint Conference on Digital Library'01 (JCDL'01), pp:247-248, Roanoke, VA, USA, 2001.*

8. Mingchun Liu and Chunru Wan, "A Study on Content-Based Classification and Retrieval of Audio Database", In *Proceeding of International Database Engineering and Application Symposium(IDEAS-01)*, pp:339-345, Grenoble, France, 2001.
9. Mingchun Liu and Chunru Wan, "Feature Extraction and Selection for Automatic Audio Classification", In *Proceedings of the 3rd International Conference on Information, Communications and Signal Processing (ICICS2001)*, Singapore, 2001.
10. Chunru Wan and Mingchun Liu and Lipo Wang, "Content-based Sound Retrieval for Web Application", In *Proceedings of the First Asian-Pacific Conference on Web Intelligent (WI2001)*, Maebashi, Japan, 2001.

Bibliography

- [1] Neural networks toolbox for use with matlab, user's guide, version 4, the mathworks.
- [2] <http://trec.nist.gov/>.
- [3] <http://www.nist.gov/speech/tests/sdr/sdr2000/sdr2000.htm>.
- [4] Garofolo et al. The trec spoken document retrieval track : A success story. In *Proceedings of the Text REtrieval Conference TREC-9 Spoken Document Retrieval Track*, 2000.
- [5] S.E. Johnson, P. Jourlin, G.L. Moore, K.S. Jones, and P.C Woodland. The cambridge university spoken document retrieval system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 49–52, 1999.
- [6] M. Viswanathan, H.S.M. Beigi, S. Dharanipragada, and A Tritschler. Retrieval from spoken documents using content and speaker information. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, pages 567–572, 1999.
- [7] Bo-Ren Bai, Lee-Feng Chien, and Lin-Shan Lee. Very-large-vocabulary mandarin voice message file retrieval using speech queries. In *Proceedings of the Fourth International Conference on Spoken Language*, volume 3, pages 1950–1953, 1996.
- [8] Mikko Kurimo. Thematic indexing of spoken documents by using self-organizing maps. *Speech Communication*, 38(1), 2002.
- [9] J. Makhoul, F. Kubala, T. Leek, Daben Liu, Long Nguyen, R. Schwartz, and A. Srivastava. Speech and language technologies for audio indexing and retrieval. *Proceedings of the IEEE*, 88(8):1338–1353, 2000.

- [10] Yuen-Hsien Tseng. Content-based retrieval for music collections. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 176–182, Berkeley, CA, USA, 1999.
- [11] Chih-Chin Liu, Jia-Lien Hsu, and Chen A.L.P. An approximate string matching algorithm for content-based music data retrieval. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 451–456, 1999.
- [12] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith. Query by humming: musical information retrieval in an audio database. In *Proceedings of the third ACM international conference on Multimedia*, pages 231–236, San Francisco, CA USA, 1995.
- [13] N. Kosugi, Y. Nishihara, S. Kon'ya, M. Yamamuro, and K. Kushima. Music retrieval by humming-using similarity retrieval over high dimensional feature vector space. In *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 404–407, 1999.
- [14] A. L.P. Chen, M. Chang, and J. Chen. Query by music segments: An efficient approach for song retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2000.
- [15] A.L.P. Chen and J.C.C. Chen. Query by rhythm an approach for song retrieval in music databases. In *Proceedings of the IEEE International Workshop on Research Issues in Data Engineering*, pages 139–146, 1998.
- [16] Ta-Chun Chou, A.L.P. Chen, and Chih-Chin Liu. Music databases: indexing techniques and implementation. In *Proceedings of the International Workshop on Multimedia Database Management Systems*, pages 46–53, 1996.
- [17] Yip Chi Lap and B. Kao. A study on n-gram indexing of musical features. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 2, pages 869–872, 2000.
- [18] Cheng Yang. Macs: music audio characteristic sequence indexing for similarity retrieval. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 123–126, 2001.

- [19] E.D. Scheirer. Towards music understanding without separation: segmenting music with correlogram comodulation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 99–102, 1999.
- [20] K Jensen. Pitch independent prototyping of musical sounds. In *Proceedings of the IEEE 3rd Workshop on Multimedia Signal Processing*, pages 215–220, 1999.
- [21] K. D. Martin, E. D. Scheirer, and B. L Vercoc. Music content analysis through models of audition. In *Proceedings of the ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications*, 1998.
- [22] C. Raphael. Automatic segmentation of acoustic musical signals using hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):360–370, 1999.
- [23] D. Pye. Content-based methods for the management of digital music. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 2437–2440, 2000.
- [24] I. Kaminsky and A. Materka. Automatic source identification of monophonic musical instrument sounds. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 1, pages 189–194, 1995.
- [25] K. D. Martin and Y. E. Kim. Musical instrument identification: a pattern-recognition approach. In *Proceedings of the 136th Meeting of the Acoustical Society of America*, Norfolk, VA, 1998.
- [26] A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 63–71, 2000.
- [27] T. Miiva and Y. Tadokoro. Musical pitch estimation and discrimination of musical instruments using comb filters for transcription. In *Proceedings of the 42nd Midwest Symposium on Circuits and Systems*, volume 1, pages 105–108, 1999.
- [28] Adam; Beth Logan; Daniel P.W. Ellis; Brian Whitman Berenzweig. A large-scale evaluation of acoustic and subjective music similarity measures. In *Proceedings of the 4th International Symposium on Music Information Retrieval (ISMIR2003)*, Baltimore, Maryland (USA), 2003.

- [29] Jean-Julien; Franois Pachet Aucouturier. Music similarity measures: What's the use? In *Proceedings of the 3rd International Symposium on Music Information Retrieval(ISMIR2002)*, Paris (France), 2002.
- [30] Roger Dannenberg. Music information retrieval as music understanding. In *Proceedings of the 4th International Symposium on Music Information Retrieval(ISMIR2001)*, Bloomington, Indiana (USA), 2001.
- [31] Geoffroy; Amaury La Burthe; Xavier Rodet Peeters. Toward automatic music audio summary generation from signal analysis. In *Proceedings of the 3rd International Symposium on Music Information Retrieval(ISMIR2002)*, Paris (France), 2002.
- [32] Cheng Yang. Macsis: A scalable acoustic index for content-based music retrieval. In *Proceedings of the 3rd International Symposium on Music Information Retrieval(ISMIR2002)*, Paris (France), 2002.
- [33] Hui; H. V. Jagadish Jin. Indexing hidden markov models for music retrieval. In *Proceedings of the 3rd International Symposium on Music Information Retrieval(ISMIR2002)*, Paris (France), 2002.
- [34] Jon Dunn. Beyond variations: Creating a digital music library. In *Proceedings of the 4th International Symposium on Music Information Retrieval(ISMIR2000)*, Plymouth, Massachusetts (USA), 2000.
- [35] David Bainbridge. The role of music ir in the new zealand digital music library project. In *Proceedings of the 4th International Symposium on Music Information Retrieval(ISMIR2000)*, Plymouth, Massachusetts (USA), 2000.
- [36] Stephen J. Whither Downie. Mir research: Thoughts about the future. In *Proceedings of the 4th International Symposium on Music Information Retrieval(ISMIR2001)*, Bloomington, Indiana (USA), 2001.
- [37] Alexandra Uitdenbogerd. Music ir: Past, present, and future. In *Proceedings of the 4th International Symposium on Music Information Retrieval(ISMIR2000)*, Plymouth, Massachusetts (USA), 2000.
- [38] David M.Howard and James Angus. *Acoustics and Psychoacoustics*. Focal Press, 1996.
- [39] Thomas D.Rossing. *The Science of Sound*. Addison-Wesley Publishing Company, 1983.

-
- [40] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.
 - [41] Silvia.Pfeiffer, Stephan.Fischer, and Wolfgang.Effelsberg. Automatic audio content analysis. In *Proceedings of the ACM Multimedia*, pages 21–30, 1996.
 - [42] T. Zhang and C.-C. Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, 9(4):441–457, 2001.
 - [43] G. Tzanetakis and P. Cook. A framework for audio analysis based on classification and temporal segmentation. In *Proceeding the 25th EUROMICRO Conference*, volume 2, pages 61–67, 1999.
 - [44] George Tzanetakis and Perry Cook. Sound Analysis Using MPEG Compressed Audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000.
 - [45] Albert S Bregman. *Auditory scene analysis : the perceptual organization of sound*. MIT Press, 1990.
 - [46] Ozgur Izmirli. Using a spectral flatness based feature for audio segmentation and retrieval. In *Proceedings of the International Symposium on Music Information Retrieval*, 2000.
 - [47] W Reichl and Wu Chou. A unified approach of incorporating general features in decision tree based acoustic modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 573–576, 1999.
 - [48] B.L. Vercoe, W.G. Gardner, and E.D. Scheirer. Structured audio: creation, transmission, and rendering of parametric sound representations. *Proceedings of the IEEE*, 86(5):922–940, 1998.
 - [49] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 452–455, 2000.
 - [50] G. Tzanetakis and P. Cook. Multifeature audio segmentation for browsing and annotation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*,, pages 103–106, 1999.

- [51] K. Melih and R. Gonzalez. Audio source type segmentation using a perceptually based representation. In *Proceedings of the Fifth International Symposium on Signal Processing and Its Applications*, volume 1, pages 51–54, 1999.
- [52] K. Melih and R. Gonzalez. Source segmentation for structured audio. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 2, pages 811–814, 2000.
- [53] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1423–1426, 2000.
- [54] Tong Zhang and C.-C. Jay Kuo. Heuristic approach for generic audio data segmentation and annotation. In *Proceedings of the ACM Multimedia Conference*, pages 67–76, Orlando, 1999.
- [55] M. De Santo, G. Percannella, C. Sansone, and M. Vento. A Neural Multi-Expert Classification System for MPEG Audio Segmentation. In *Proceedings of the International Conference on Advances in Pattern Recognition*, pages 50–59, 2001.
- [56] C. De Stefano, A. Della Cioppa, and A. Marcelli. An investigation on MPEG audio segmentation by evolutionary algorithms. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 952–956, 2001.
- [57] Guojun Lu and Templar. Hankinson. A technique towards automatic audio classification and retrieval. In *Proceedings of the International Conference on Signal Processing*, 1998.
- [58] Guojun Lu and T. Hankinson. An investigation of automatic audio classification and segmentation. In *16th World Computer Congress, Proceedings of the 5th International Conference on Signal Processing (WCC-ICSP)*, pages 776–781, 2000.
- [59] Z. Liu and Q. Huang. Classification of audio events in broadcast news. In *Proceedings of the IEEE Signal Processing Society Workshop on Multimedia Processing*, pages 364–369, 1998.

- [60] J. Foote. A similarity measure for automatic audio classification. In *Proceedings of the AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, Stanford, California, 1997.
- [61] E Scheirer and M Slaney. Construction and evaluation of a robust multi-feature speech/music discriminator. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 1331–1334, 1997.
- [62] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 11, pages 993–996, 1996.
- [63] Y. Qi and B.R. Hunt. Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. *IEEE Transactions on Speech and Audio Processing*, 1(2):250–255, 1993.
- [64] T.Lambrou, P.Kudumakis, M.Sandler, R.Speller, and A.Linney. Classification of audio signals using statistical features on time and wavelet transform domains. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, USA, 1998.
- [65] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, and Tom McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22:533–544, 2001.
- [66] Nitin Sawhney. Situational awareness from environmental sounds. Technical report, MIT Media Lab, 1997.
- [67] Tong Zhang and C.-C. Jay Kuo. Hierarchical system for content-based audio classification and retrieval. In *Proceedings of the SPIE's Conference on Multimedia Storage and Archiving Systems III*, volume 3527, pages 398–409, Boston, 1998.
- [68] Tong Zhang and C.-C. Jay Kuo. Classification and retrieval of sound effects in audiovisual data management. In *Proceedings of the 33rd Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, 1999.
- [69] M.J. Paradie and S.H. Nawab. The classification of ringing sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2435 –2438, 1990.

- [70] Tong. Zhang and C.-C. Jay. Kuo. Content-based classification and retrieval of audio. In *Proceedings of the SPIE's 43rd Annual Meeting - Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII*, volume 3461, pages 432–443, San Diego, 1998.
- [71] P.J. Moreno and R. Rifkin. Using the fisher kernel method for web audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 2417–2420, 2000.
- [72] Nakajima.Y., Yang Lu, Sugano.M, Yoneyama. A, Yamagihara. H, and Kurematsu.A. A Fast Audio Classification from MPEG Coded Data. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3005–3008, 1999.
- [73] J. Foote. An overview of audio information retrieval. *ACM Multimedia Systems*, 7:2–10, 1999.
- [74] D. Keislar, T. Blum, J. Wheaton, and E. Wold. Audio databases with content-based retrieval. In *Proceedings of the International Computer Music Conference*, pages 199–202, San Francisco, 1995.
- [75] E. Wold, T. Blum, D. Keislar, and et al. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, pages 27–36, 1996.
- [76] S. Z. Li. Content-based audio classification and retrieval using the nearest feature line method. *IEEE Transactions on Speech and Audio Processing*, 8(5):619–625, 2000.
- [77] Zhu Liu and Qian Huang. Content-based indexing and retrieval-by-example in audio. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 2, pages 877–880, 2000.
- [78] Jonathan T. Foote. Content-based retrieval of music and audio. In *Proceedings of the SPIE*, volume 3229, pages 138–147, 1997.
- [79] G Smith, H Murase, and K Kashino. Quick audio retrieval using active search. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 6, pages 3777–3780, 1998.
- [80] K. Kashino, G. Smith, and H. Murase. Time-series active search for quick retrieval of audio and video. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 2993–2996, 1999.

- [81] K. Kashino, T. Kurozumi, and H. Murase. Feature fluctuation absorption for a quick audio retrieval from long recordings. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 3, pages 98–101, 2000.
- [82] S.E. Johnson and P.C. Woodland. A method for direct audio search with applications to indexing and retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1427–1430, 2000.
- [83] P. Piamsa-Nga, N.A. Alexandridis, S. Srakaew, G.C. Blankenship, and S.R. Jr, Subramanya. In-clip search algorithm for content-based audio retrieval. In *Proceedings of the Third International Conference on Computational Intelligence and Multimedia Applications*, pages 263–267, 1999.
- [84] Tong Zhang and C.-C. Jay Kuo. Hierarchical classification of audio data for archiving and retrieving. In *Proceedings of the IEEE International Conference On Acoustics, Speech, and Signal Processing*, volume 6, pages 3001–3004, Phoenix, 1999.
- [85] Melih. K and R Gonzalez. Audio retrieval using perceptually based structures. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, pages 338–347, 1998.
- [86] Kathy Melih and Ruben Gonzalez. Identifying perceptually congruent structures for audio retrieval. In *Proceedings of the 5th International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services*, pages 125–136, 1998.
- [87] Z. Liu, J. Huang, and Y. Wang. Classification TV programs based on audio information using hidden Markov model. In *Proceedings of the IEEE Second Workshop on Multimedia Signal Processing*, pages 27–32, 1998.
- [88] J Huang, Z Liu, Y Wang, Y Chen, and E. K Wong. Integration of multi-modal features for video scene classification based on hmm. In *Proceedings of the IEEE 3rd Workshop on Multimedia Signal Processing*, pages 53–58, 1999.
- [89] J.S. Boreczky and L.D. Wilcox. A hidden markov model framework for video segmentation using audio and image features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 6, pages 3741–3744, 1998.

- [90] Jincheng Huang, Zhu Liu, and Wang Yao. Integration of audio and visual information for content-based video segmentation. In *Proceedings of the International Conference on Image Processing*, volume 3, pages 526–529, 1998.
- [91] Silvia Pfeiffer, Rainer Lienhart, and Wolfgang Effelsberg. Scene determination based on video and audio features. In *Proceedings of the IEEE Conference on Multimedia Computing and Systems*, pages 685–690, 1998.
- [92] Hao Jiang, Tong Lin, and Hong-Jiang Zhang. Video segmentation with the assistance of audio content analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 3, pages 1507–1510, 2000.
- [93] Tong Zhang and C.-C. Jay Kuo. Audio-guided audiovisual data segmentation and indexing. In *Proceedings of the Conference on Storage and Retrieval for Image and Video Database VII in IST/SPIE's Symposium on Electronic Imaging*, volume 3656, pages 316–327, San Jose, California, 1999.
- [94] David Pye, Nicholas J. Hollinghurst, Timothy J. Mills, and Kenneth R. Wood. Audio-visual segmentation for content-based retrieval. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, 1998.
- [95] S. Tsekeridou and I. Pitas. Audio-visual content analysis for content-based video indexing. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 667–672, 1999.
- [96] S. Tsekeridou and I. Pitas. Content-based video parsing and indexing based on audio-visual interaction. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(4):522–535, 2001.
- [97] T. Muramoto and M. Sugiyama. Visual and audio segmentation for video streams. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 3, pages 1547–1550, 2000.
- [98] Wei Qi, Lie Gu, Hao Jiang, Xiang-Rong Chen, and Hong-Jiang Zhang. Integrating visual, audio and text analysis for news video. In *Proceedings of the International Conference on Image Processing*, volume 3, pages 520–523, 2000.

-
- [99] Berners-Lee T., Hendler J., and Lassila O. The semantic web. *Scientific American*, pages 34–43, 2001.
 - [100] Gruber T.R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.
 - [101] Natalya Fridman Noy and Deborah L. McGuinness. Ontology development 101: A guide to creating your first ontology. Technical Report KSL-01-05 and SMI-2001-0880, March 2001.
 - [102] Weng Zu Mao and D. Bell. Integrating visual ontologies and wavelets for image content retrieval. In *Proceedings of the Ninth International Workshop on Database and Expert Systems Applications*, pages 379–384, 1998.
 - [103] L. Khan and D. McLeod. Audio structuring and personalized retrieval using ontologies. In *Proceeding of IEEE Advances in Digital Libraries*, pages 116–126, 2000.
 - [104] George Tzanetakis and Perry Cook. Experiments in computer-assisted annotation of audio. In *Proceedings of the Int. Conf. Auditory Display (ICAD00)*, 2000.
 - [105] Kyu Phil Han, Young Sik Park, Seong Gyu Jeon, Gwang Choon Lee, and Yeong Ho Ha. Genre classification system of tv sound signals based on a spectrogram analysis. *IEEE Trans. on Consumer Electronics*, 44(1):33–42, 1998.
 - [106] N. Adam and A. Gangopadhyay. Content-based retrieval in digital libraries. *IEEE Computer*, 31:93–95, 1998.
 - [107] Z. Liu, J. Huang, Y. Wang, and et al. Audio feature extraction and analysis for scene classification. In *Proceedings of the IEEE 1st Multimedia Workshop*, 1997.
 - [108] Yao Wang, Zhu Liu, and Jin-Cheng Huang. Multimedia content analysis—using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6):12–36, 2000.
 - [109] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. Wiley, 2000.

- [110] A.K. Jain, P.W. Duin, and Jianchang Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.
- [111] G.L. Goodman and D.W. McMichael. A Gaussian mixture model classifier using supervised and unsupervised learning. In *Proceedings of the Fourth International Symposium on Signal Processing and its Applications*, pages 565–566, 1996.
- [112] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society Series B*, 39:1–38, 1977.
- [113] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
- [114] D. F. Specht. Probabilistic neural networks. *Neural Networks*, 3:109–118, 1990.
- [115] D. F. Specht. Probabilistic neural networks and the polynomial adaline as complementary techniques for classification. *IEEE Transactions on Neural Networks*, 1:111–121, 1990.
- [116] A. Zaknich. Introduction to the modified probabilistic neural network for general signal processing applications. *IEEE Transactions on Signal Processing*, 46(7):1980–1990, 1998.
- [117] N Vlassis and A Likas. Kurtosis-based dynamic approach to Gaussian mixture modeling. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 29:393–399, 1999.
- [118] David A. Grossman and Ophir Frieder. *Information Retrieval: Algorithms and Heuristics*. Kluwer Academic Publishers, 1998.
- [119] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, 1971.
- [120] J. Zhang, Y. Zhao, Z. Yang, and J. Wang. Adaptive parameter tuning for relevance feedback of information retrieval. In *Proceedings of the 4th World Congress on Intelligent Control and Automation*, volume 3, pages 2132–2135, 2002.

- [121] P. Amo, F.L. Ferreras, F. Cruz, and M. Rosa. Smoothing functions for automatic relevance feedback in information retrieval. In *Proceedings of the 11th International Workshop on Database and Expert Systems Applications*, pages 115–119, 2000.
- [122] C.H. Chang and C.C. Hsu. Enabling concept-based relevance feedback for information retrieval on the www. *IEEE Transactions on Knowledge and Data Engineering*, 11(4):595–609, 1999.
- [123] Tat-Seng Chua, Chun-Xin Chu, and M. Kankanhalli. Relevance feedback techniques for image retrieval using multiple attributes. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 890–894, 1999.
- [124] S.D. MacArthur, C.E. Brodley, and Chi-Ren Shyu. Feedback decision trees in content-based image retrieval. In *Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 68–72, 2000.
- [125] R. Schettini, G. Ciocca, and I. Gagliardi. Content-based color image retrieval with relevance feedback. In *Proceedings of the International Conference on Image Processing*, volume 3, pages 75–79, 1999.
- [126] F. Jing, M. Li, H.J. Zhang, and B. Zhang. Learning region weighting from relevance feedback in image retrieval. In *Proceedings of the IEEE ICASSP'02*, volume 4, pages 4088–4091, 2002.
- [127] A.D. Doulamis, Y.S. Avrithis, N.D. Doulamis, and S.D. Kollias. Interactive content-based retrieval in video databases using fuzzy classification and relevance feedback. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, volume 2, pages 954–958, 1999.
- [128] R. Wang, M.R. Naphade, and T.S. Huang. Video retrieval and relevance feedback in the context of a post-integration model. In *Proceedings of the IEEE Fourth Workshop on Multimedia Signal Processing*, pages 33–38, 2001.
- [129] Yong Rui, T.S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 644–655, 1998.

- [130] H. Muller, W. Muller, S. Marchand-Maillet, T. Pun, and D.M. Squire. Strategies for positive and negative relevance feedback in image retrieval. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 1, pages 1043–1046, 2000.
- [131] S. Aksoy, R.M. Haralick, F.A. Cheikh, and M. Gabbouj. A weighted distance approach to relevance feedback. In *Proceedings of the 15th international conference on pattern recognition*, volume 4, pages 812–815, 2000.
- [132] L.A.Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.
- [133] L.A.Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. on Systems, Man and Cybernetics*, 3:28–44, 1973.
- [134] F. Beritelli, S. Casale, and M. Russo. Multilevel speech classification based on fuzzy logic. In *Proceedings of the IEEE Workshop on Speech Coding for Telecommunications*, pages 97–98, 1995.
- [135] E. H. Mamdani. Advances in linguistic synthesis of fuzzy controllers. *Internatinal Journal of Man, Machine Studies*, 8:669–678, 1976.
- [136] L. Lu, H-J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7):504–516, 2002.
- [137] S. Pfeiffer. Pause concepts for audio segmentation at different semantic levels. In *Proceedings of the ACM Multimedia*, pages 187–193, 2001.
- [138] C. Saraceno and R. Leonardi. Audio as a support to scene change detection and characterization of video sequences. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 2597–2600, 1997.
- [139] J. Nam and A.H. Tewfik. Combined audio and visual streams analysis for video sequence segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 2665–2668, 1997.
- [140] H. Sundaram and Shih-Fu Chang. Video scene segmentation using video and audio features. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 2, pages 1145–1148, 2000.

- [141] A.G. Hauptmann and M.J. Witbrock. Story segmentation and detection of commercials in broadcast news video. In *Proceedings of the IEEE International Forum on Research and Technology Advances in Digital Libraries*, pages 168–179, 1998.
- [142] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*, 2nd ed. Springer, 2000.
- [143] C. M. Bishop and M. E. Tipping. Variational relevance vector machines. In *Proceedings of the 15th International Workshop on Statistical Modelling*, page 117, Bilbao, Spain, 2000.
- [144] S. Quackenbush and A. Lindsay. Overview of MPEG-7 audio. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):725–729, 2001.
- [145] ISO/IEC 15938-4 FCD Information Technology - Multimedia Content Description Interface - Part 4: Audio, March 2001. Singapore.
- [146] Vision, Technologies and Strategy, MPEG Document: ISO/IEC JTC1/SC29/WG11. Technical Report N3939, 2001.
- [147] <http://www.cselt.it/mpeg>.