USC-SIPI REPORT #410

Contextual Modeling of Audio Signals Toward Information Retrieval

by

Samuel Kim

December 2010

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

CONTEXTUAL MODELING OF AUDIO SIGNALS

TOWARD INFORMATION RETRIEVAL

by

Samuel Kim

A Dissertation Presented to the
FACULTY OF THE THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

December 2010

Dedication

Soli Deo Gloria

This dissertation is dedicated to my ladies,

Kate Youngin Kim

and

Evelyn Eugene Kim.

Acknowledgements

I like to thank my advisor Dr. Shrikanth Narayanan for his wonderful guidance which involves not only inspiring motivation but also encouraging patience. I also like to thank the committee members, Dr. Panayiotis Georgiou, Dr. C.-C. Jay Kuo, Dr. Cyrus Shahabi, and Dr. Antonio Ortega, for their comments and suggestions.

Wrapping up the quarter-century long school education, there is no way to write an acknowledging statement that can express my gratitude enough. If any, the acknowledgement would be much longer than the dissertation itself. Those who are in USC Signal Analysis and Interpretation Lab (SAIL) members, USC Good Shepherd (GS) members, Lamp Presbyterian Church members, Digital Signal Processing Lab (DSP) members at Yonsei University, etc., I really thank all of you from the bottom of my heart.

Special thanks to Kate for her loving support. As a life-time partner, we will keep asking ourselves what our only comfort in life and death really is.

Table of Contents

List Of Tables

List Of Figures

Abstract

The main focus of this dissertation is on audio modeling and indexing toward audio information retrieval. In this regard, various novel methodologies are proposed in the direction of capturing *audio context* within a wide spectrum of audio contents; from well-structured music to unstructured environmental sound. This dissertation consists of two major parts depending on the types of audio contents: music information retrieval and general audio information retrieval.

In the first part, an efficient context-based music information retrieval method using music fingerprint is introduced. The music fingerprint is proposed to encapsulate musical context of a given music audio in a compact representation obtained directly from the music audio signal; it provides an efficient handle for music information retrieval in terms of both accuracy and computing requirements. The musically meaningful aspects considered in deriving this representation include harmonic structures and their temporal dynamic information (a.k.a. chord progression). Empirical results on various music information retrieval tasks, such as opus identification, composer identification and semantic description annotation show that the proposed music fingerprint is competitive to the state-of-the-art systems in terms of accuracy and computing power requirements.

In the second part, a new contextual modeling algorithm for general audio information retrieval is introduced. Assuming that hidden acoustic topics exist and they represent the context of an audio clip, we proposed a latent acoustic topic model that learns a probability distribution over a set of hidden topics of a given audio clip in an unsupervised manner. We use the latent Dirichlet allocation (LDA) method to implement the latent acoustic topic model and introduce the notion of *acoustic words* to support modeling

within this framework. The proposed audio information retrieval system also aims to provide users with flexibility in formulating their retrieval queries using naïve text as well as pre-determined categories or audio examples. To mitigate interoperability issues between the annotation and retrieval processes inherent in text descriptions, we propose an intermediate audio description layer (iADL) spanned by onomatopoeic and semantic labels in conjunction with context-based text transformation methods that map naïve descriptions onto the proposed iADL.

Chapter 1

Introduction

The main focus of this dissertation is to model context in audio signals with applications to information retrieval. It studies the relationship among an audio signal, embedded information, and verbal descriptions within an audio information retrieval framework. This is motivated by the need for efficient data mining and search schemes for audio contents in browsing, retrieval, summarization, and annotation. The goal is to propose novel methodologies for context-based audio information retrieval by investigating various aspects of audio signals, embedded information, and verbal descriptions.

1.1 Motivation

A huge amount of multimedia data has been archived, and it is exponentially growing through various routes such as broadcasting services, WEB 2.0 applications, and online stores. Advances in storage and data sharing technologies have been accelerating this growth. This multimedia data explosion has created an *information overload phenomena* [60]. In the studies of *attention economy* which deals with information management strategies, researchers showed that users are easily overwhelmed or frustrated by too much information [60, 18]. Furthermore, the information overload may cause *information pollution* problems which consume users' attention to unsolicited or undesirable information [18].

Information management strategies for multimedia data, therefore, become crucial as the amount of multimedia data grows [52]. In fact, information retrieval from the multimedia data has been studied from various aspects: text, image, video, audio, and their combinations. In this work, we focus on the aural aspect of multimedia data which provides audible information to users. It can be found in many multimedia formats; in some formats, such as movies or animations, audio signals play important roles for users to experience the multimedia data. On the other hand, in some multimedia formats, such as music or sound clips, audio signals are the only media that users can experience. Audio information retrieval, therefore, provides an essential, complementary in some cases, way to access the information embedded in the multimedia data.

1.2 Open Challenges

The main question in designing an audio information retrieval system is *how to link audio signals to descriptions of embedded information in audio signals and desired information of users*. If we rewrite the question in machine learning terminologies, it can be rephrased as *how to "model" audio "features" according to their "categories"*. It accompanies three classical machine-learning concerns: reliability, efficiency, and scalability. Reliability should be considered for the system to be toll quality, and efficiency should be considered for dealing with a huge database. To deal with a fast growing amount of data, such as today's multimedia data, the scalability factor is inevitable.

One simple idea in extracting audio features is to use manually-generated descriptions such as filenames and metadata embedded in audio files. Although it may satisfy the reliability criterion by providing high accuracy within a given data set, it is not scalable because manually labeling individual audio files in an exponentially growing database is not tractable. Therefore, many systems are being developed in the direction of a content-based approach which aims to extract desired information directly from audio signals without prior-tagged labels.

Content-based audio information retrieval, however, also includes various open challenges. An audio signal typically represents a heterogeneous mixture of several sound sources; each sound source carries its own array of information and the mixture preserves individual characteristics. Furthermore, these heterogeneous aspects are time-variant; sound sources that compound an audio signal are distributed over time unevenly. In this section, we point out a couple of challenges related to the heterogeneity.

The challenges from the heterogeneous characteristics are often related to extracting reliable features from audio signals using various signal processing algorithms. Since not all of the sources are of interest or relevant, it is necessary to consider only desirable sound sources among multiple sound sources. This raises technical questions that have not been completely solved yet. Actually, one of the critical stumbling blocks is that the way of combining the sound sources is usually not known or difficult to estimate. In the speech signal processing society, for example, researchers have being trying to mitigate ambient noises to extract robust speech features [11, 62]. In pursuing the goal, many methodologies are being developed based on a set of assumptions about how speech signals and ambient noises are combined; common assumptions may include that they are statistically independent or uncorrelated. The results show that, however, the algorithms have not reached the complete solution so far, which indicates that the way of combining speech signals and ambient noises is not fully understood yet. The time-variant heterogeneous characteristics also prohibit extracting reliable features from desirable sound sources; it is because the distribution of sound sources can be changed as time changes. To deal with the time-variant heterogeneous characteristics, salient region extraction [38] and foreground/background classification [22] have been studied for extracting desired sound sources.

Another challenges from the heterogeneous characteristics are related to the context-dependent characteristics of audio signals; similar audio contents indicates different meanings according to surrounding sounds. Suppose an engine sound is present in an audio clip. Even if an audio information retrieval system correctly recognize the engine sound,

some degree of uncertainty still exist in terms of where the sound comes from; it can be recorded in a factory, in constructing site, or in a car. Considering surrounding sound would help to disambiguate the uncertainty. If there is a sound of baby crying along with an engine sound, it is likely that the sound is recorded in a car.

On the other hand, the variability in describing embedded information is also a challenging factor in designing audio information retrieval systems. The variability in descriptions is inherited from various factors; one of the factors is the heterogeneity of audio signals, and another is the way of describing audio signals. From the heterogeneity point of view, individual sound sources preserve their own array of information in a mixed audio signal. Even a signal from a single sound source contains acoustic variability so that it carries a variety of information. In speech signals, for example, there exists an array of information in a single utterance: lexical content, speaker identity, emotion, health condition, etc. This abundant information tends to be modeled separately with task-dependent suboptimal representations for target applications [31]. Consequently, models trained for a specific audio information retrieval system are usually not usable for other systems that retrieve other types of information. This property is contrary to the scalability criterion because a system should retrain the models if a user requires a new information for which the system has not been trained yet. In addition to the problems in dealing with audio signals, there exist ambiguities in descriptions themselves due to inherent characteristics of text descriptions; people can use different words to describe the same object and the same word can indicate different meanings (polysemic words). There are many studies to tackle this problem in text information retrieval applications and natural language processing [6, 7, 37].

1.3 Proposed Approaches

The main focus of this dissertation is on context-based audio information retrieval, which investigates how an audio signal is generated by analyzing neighboring sounds. As we

approach the problem, the multifaceted nature of the audio contents should be considered in developing a context-based audio information retrieval system. The multifaceted nature of the audio contents represents that each category of sound has their own strategy to build the sound. Fig. 1.1 represents the spectrum of audio contents with respect to how the audio contents are generated along with a few examples. The right side of the spectrum represents the audio contents that are generated in a "structured" way, while the left side of the spectrum represents the audio contents that are generated in an "unstructured" way. The terms "structured" and "unstructured" are used to denote whether there exist evident rules in producing the sound.

Music audio signals, for example, are located in the most right side of the spectrum among the given examples. Music audio signals are typically mixtures of a variety of musical instruments and voices, and there exist a set of rules that govern individual sound sources. The rules are usually written in music scores which explicitly denote how to make the sound, i.e. play, in terms of pitch, timing, and intensity. The degree of freedom that the music audio signals have is, if any, limited. Speech signals, the second most right example, are also highly structured in terms of a person having to articulate his or her organs to make the sound, i.e. speak, for generating and transmitting linguistic information. It is well structured in the sense that the linguistic information can be

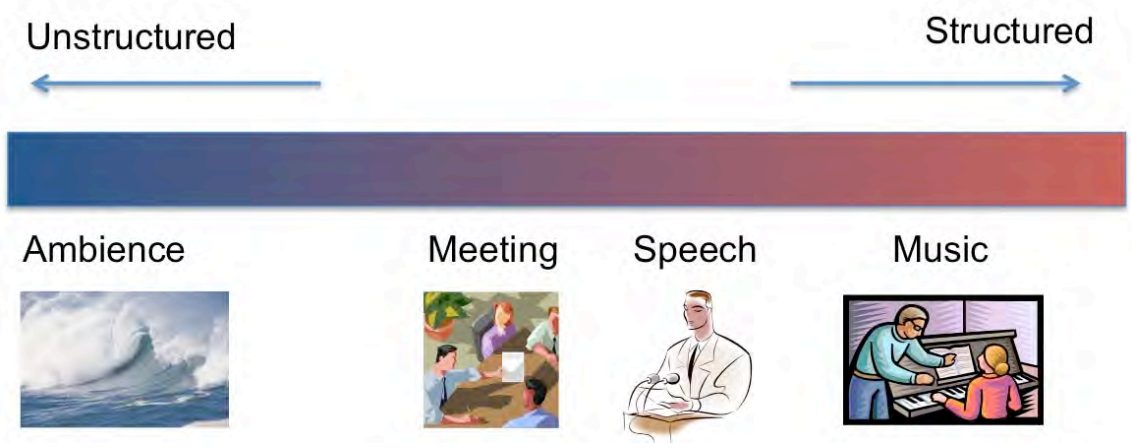


Figure 1.1: Spectrum of audio contents in terms of how the audio contents are generated.

written in a text form. However, the degree of freedom is relatively higher than that of the music audio signals. Let us consider two different speech signals which contain the same linguistic information. Even though they have the same linguistic information, they differ according to various factors such as speaker characteristics, emotions, and health conditions. This variability increases if there is more than one speaker talking, such as in a meeting environment (the second left example). The most left side example is ambient audio signals which can be classified as unstructured sounds. They are usually mixtures of various environmental sound sources, but the sound generating rules to combine those sound sources are rarely evident.

In this dissertation, we aim to develop a context-based audio information retrieval system that models the context of a given audio signal according to the characteristics of the given audio signal.

1.4 Objectives

In this dissertation, we propose to explore various aspects of the wide spectrum of audio contents in the audio information retrieval framework. This spectrum was introduced in the previous section as we analyzed audio contents with respect to how audio contents are generated. In this wide spectrum, our exploration will be conducted in the way of achieving the following specific objectives.

The first objective is to introduce "context-based" approaches for audio information retrieval which shed a light on the problems from the perspective of how the sounds are produced rather than "content-based" approaches which focus on what the sounds are like. These two types of approaches may seem similar, but they differ in the sense that a context-based approach seeks any constructing rules in the contents. In pursuing the goal, we aim to devise the methodologies to extract constructing rules from two extreme cases of audio contents: well-structured audio signals and general audio signals which include unstructured audio signals. We will emphasize the efficiency criterion in dealing with well-structured audio signals, since constructing rules may be evident and relatively

easy to extract. In general audio signals, we emphasize extracting non-evident, possibly hidden rules to produce the sound signals.

The second objective is to use the extracted context as features to index the corresponding sound. In pursuing the objective, we apply the proposed context-based approaches to different categories of descriptions to show if the constructing rules are valid representations for various types of information. We will also investigate whether the methodologies satisfy the efficiency and scalability criteria as well as the reliability criterion in various empirical tasks.

To evaluate and apply these ideas, we use two different application scenarios: music information retrieval and general audio information retrieval; in each application scenario, we propose novel methodologies to model constructing rules and use the rules for audio information retrieval. The following subsections will provide brief descriptions of the proposed algorithms.

1.4.1 Music Information Retrieval

Music audio signals are relatively well-structured according to the spectrum of audio contents. The constructing rule, which governs various sound sources such as instruments and voices, is well studied in *music theory*. In this study, we propose to investigate the role of the musical rules based on the music theory in music information retrieval systems. Specifically, we focus on extracting musically meaningful representations such as harmony structures and their dynamics (a.k.a chord progression). The rationale behind using these musical attributes is that they capture unique features of a given music audio signal.

Although it may not be necessary for all the music information retrieval applications to have musically meaningful features, some applications such as those considered in this paper rely on, or benefit from, such representations and models. In this dissertation, we propose an efficient methodology to extract constructing rules, such as harmony structures and their dynamics, of music audio signals and use it for context-based music information retrieval. We call the proposed representation *music fingerprint* [40, 44] in the sense that

it is a signal description that can be embedded in audio files compactly in a limited memory space. The music fingerprint consists of a set of real values which are extracted directly from acoustic signals.

The proposed context-based music fingerprint offers several advantages. It requires less computing power so it can be implemented in energy-sensitive mobile devices such as portable mp3 players. The proposed music fingerprint has also been found to be attractive in terms of accuracy through experiments on various information retrieval tasks. Furthermore, since it is directly extracted from signals, the music fingerprint can be generated readily if the audio file does not have or is missing the music fingerprint. Therefore, it can be used as an alternative and complementary access to music information that text metadata provides.

In this dissertation, the advantages of the proposed music fingerprint are demonstrated on a variety of tasks whose focuses are on different types of information; namely *opus identification*, *composer identification*, and *semantic annotation tasks*. The reason we choose these specific tasks is that the target information for which the individual tasks seek is usually written as text metadata and they cover a variety of information that is used to describe music audio signals.

The details of the proposed methodologies along with experimental results are described in Chapter 2.

1.4.2 General Audio Information Retrieval

Dealing with general audio signals is particularly difficult due to the heterogeneity of the signals; they include unstructured audio signals whose constructing rules are, if any, not known or difficult to estimate. The embedded information is also ambiguous due to the variety of information that the heterogenous mixture includes.

To extract possibly hidden semantic rules from general audio signals, we propose a *latent acoustic topic model* influenced by the topic model that was originally proposed for text information retrieval [35, 17]. In text information retrieval, it models each document

as a distribution over a fixed number of unobservable hidden topics. Each topic, in turn, can be modeled as a distribution over a fixed number of observable words. One of the motivations for modeling hidden topics in a document is to handle the ambiguities of interpretations of words; although individual words have their own meanings, the interpretations of the words vary according to the topics of the document.

The latent acoustic topic model is motivated by drawing analogies between text and sound. We hypothesize that short segments of audio signals play similar roles as words in text and that there are latent topics in audio signals which would be determined by the context of audio signals. In other words, each audio clip is viewed to consist of latent *acoustic topics* that generate *acoustic words*. We use the latent Dirichlet allocation (LDA) method [15] to model the latent acoustic topics.

We also attempt to provide users with flexibility for their queries, so that people can use naive descriptions or audio examples for their queries. To this end, we introduce a novel method to disambiguate the descriptions of embedded or desired information. Assuming that onomatopoeia and semantic labels provide essential descriptions about sounds, we introduce a method that transforms naive text queries or audio examples to onomatopoeia and semantic labels for audio information retrieval. The reason we choose the onomatopoeic words and semantic labels is that they were shown to carry abundant information about sounds in human-to-human audio information retrieval tasks [78]. Furthermore, as it is seen in Fig. 3.2, these labels are particularly interesting because they are highly related to psychoacoustic activities which connect physical properties and human experience of sounds; onomatopoeia labels can be considered from the perspective of the *sensation* process, and semantic labels from *cognition* process [61].

The details of the proposed methodologies along with experimental results are described in Chapter 3.

1.5 Contribution Summary

The proposed context-based audio information retrieval presents new ideas, techniques, and future directions. This section summarizes the contributions of this dissertation; specifically in music information retrieval and general audio information retrieval. In both applications, we contributed

1. by providing novel structure analysis methodologies for finding context of audio contents
2. by utilizing the constructing rules as features for information retrieval tasks.

The following points describe the specific contributions of this work.

- Music Information Retrieval
  - Finding audio musical context
    - Modeling musically meaningful attributes: harmony structure, usage of pitch classes, and their dynamics
    - Extracting dynamic information: delta chroma feature and chromatic delta feature
    - Music Fingerprint: unique and compact representation for a given music piece
  - Using the musical context for information retrieval
    - Music fingerprint
    - Accurate, efficient, and scalable handle for music information retrieval
    - Applicable in portable devises such as MP3 players
    - Alternative and complementary to text metadata

- General Audio Information Retrieval
  - Finding general audio context
    - Latent acoustic topic model: discovering latent structure of audio signal using latent Dirichlet allocation (LDA)
    - Acoustic words: representing an audio signal as a text signal to make a text-like audio signal
    - N-gram approach: capturing partial dynamics of acoustic words considering adjacent acoustic words
  - Using the rules for information retrieval
    - Intermediate audio description layer (iADL): 2-dimensional domain using onomatopoeia and semantic labels
    - Providing users with query flexibility: audio examples and naive text queires
    - Query transformation: transformation of naive text queries onto intermediate audio description layer

These contributions initiate the following future directions.

- Multimodal approaches
  - Fusion with video features; general multimedia data analysis
  - Analysis of multimedia data in Web 2.0 applications, such as SNS and Youtube data
- Analysis of community-contributed multimedia data
  - How people respond to multimedia data
  - Analysis of social tags: sentiments and emotional tags
  - Ratings, recommend/unrecommend

1.6 Structure of the Dissertation

This dissertation is organized as follows. We will focus on information retrieval from music audio signals and general audio signals in Chapter 2 and Chapter 3, respectively. In each chapter, the detailed descriptions of the proposed context-based information retrieval from the corresponding audio signals will be provided along with experimental settings and results. In Chapter 4, the conclusions and future work will be presented.

Chapter 2

Music Information Retrieval

2.1 Introduction

A variety of music information retrieval (MIR) systems have been proposed and implemented targeting specific end use and interface requirements. The need for such music information retrieval systems has accompanied recent advances in storage and networking capabilities that have accelerated the multimedia data explosion, and have fueled this need for intuitive and efficient data mining and search schemes. A singular evidence for this is provided by the growth of an annual evaluation, music information retrieval evaluation exchange (MIREX), exemplifying the growing body of interests of various MIR applications [25].

The use of text metadata embedded within audio files, providing manually-generated language based descriptions of the corresponding audio file, is one of the simple solutions for MIR applications to enable convenient user interactions. This could include an array of information including, but not limited to, composer, genre, lyrics, mood reflected, and opinions expressed by listeners. With a large volume of data, however, it is not tractable to manually label the whole database to generate metadata that adequately capture the rich music information. Also, there are other forms of interactions that involve direct example based query than text based query. Therefore, many systems are also being developed in the direction of content-based information retrieval which aims to extract useful information directly from audio signals. Building a reliable content-based music

information retrieval system, however, is also very challenging due to several factors. Firstly, dealing with the audio signal is much more difficult than processing text data. Since the music audio signal typically represents a mixture of several instruments or voices, from a signal processing point of view, it is necessary to handle multiple pitches (polyphony) and multiple timbres that raise technical questions that have not been completely solved yet. Secondly, computing power and memory requirements are usually greater than for information retrieval using text metadata. It is not a desirable option for low computing power devices, such as portable mp3 players and mobile phones. Finally, extracting musically meaningful information including rhythm, harmonic structures, and chord progression is a challenging issue and includes several open problems. Although it may not be necessary for all the applications to have musically meaningful features, some applications (such as those considered in this paper) rely on, or benefit from, such representations and models. There are, however, numerous, often disparate, questions and challenges in this line of work, and one approach researchers have adopted is to focus on the specific aspect of musicality relevant to their retrieval task at hand.

In this work, our goal is to propose an efficient handle for context-based music information retrieval using musically meaningful context. Specifically, the proposed method is based on extracting two types of musical context: vertical and temporal context. Vertical context represents the way of different pitch classes are played simultaneously, while temporal context represents the movement of pitch classes along with time. They can be denoted as harmonic structure and harmonic progression (a.k.a. chord progression or simultaneity succession) respectively in musical terms. The advantages of the proposed method is shown in an *opus identification* framework. The opus identication task is to classify given music pieces with their opus numbers; it is very similar to the cover song identification task in MIREX evaluation except it deals with Classical music. The same opus can be played in various ways: different tempo and keys and/or with different instruments. The advantages of the proposed music fingerprint are also demonstrated on a variety of tasks, namely *composer identification*, and *semantic annotation tasks*. The

reason we choose these tasks is that the target information for which the individual tasks seek is usually written as text metadata.

The main contributions of this work are two-fold: a novel feature extraction procedure that incorporates temporal context and a new modeling scheme for musical context that capture unique information of music pieces. Assuming that conventional chroma feature vectors model the vertical context of given music audio signal segments by representing energy distributions over Western pitch classes, we propose a couple of methods to describe sequential dynamics in the chroma feature vectors to capture the temporal context of a given music piece. To model the musical context with a compact representation, we use second-order statistics of the features instead of highly profiled machine learning algorithms. Although using a sophisticated machine learning approach, such as with an HMM or GMM, may yield accurate performance in some applications, considerable computing power and storage space are demanded. The proposed contextual modeling method requires less computing power so it can be implemented in energy-sensitive mobile devices such as portable mp3 players. Furthermore, since it is directly extracted from signals, it can be generated readily if the audio file does not have or is missing the model.

We will describe the related work and a summary of our contributions later in this section. The description of the proposed context modeling approach will be provided in the following sections: extracting feature vectors in Section 2.2, building music fingerprint in Section 2.3, and a similarity measure for use with the proposed built music fingerprint in Section 2.4. The experimental setups and results for the target MIR applications are described in Section 2.5 and Section 2.6, respectively.

2.1.1 Related Work

Fig. 2.1 shows the basic system diagram of a generic content-based music information retrieval system which takes an acoustic music signal as an input and yields retrieved information. It typically consists of three major components: feature extraction, modeling, and similarity measurement. Each processing step faces many theoretical and practical

issues, and a variety of different approaches are being pursued to tackle those issues. In this section, we introduce related previous work categorized in terms of major contributions. It should be noted that tackling a certain issue arising from one component, however, cannot be performed independently since individual components are highly related to one another. The modeling procedure should consider the type of feature vector, while the similarity measure needs to be chosen according to the modeling scheme.

2.1.1.1 Feature Extraction

The focus here is to seek, from a given audio music signal, features that describe desirable attributes for the target application. Robustness issues (e.g. against timbre variation and noisy environment) need to be consider because dealing with signals that are a mixture of multiple pitches and multiple timbres is challenging.

Mandel *et al.* utilized a timbre-related feature to measure the similarity in terms of the artist of music [55]. They computed the overall distribution of mel frequency cepstral coefficients (MFCC) for each piece of music audio, which is one of the most

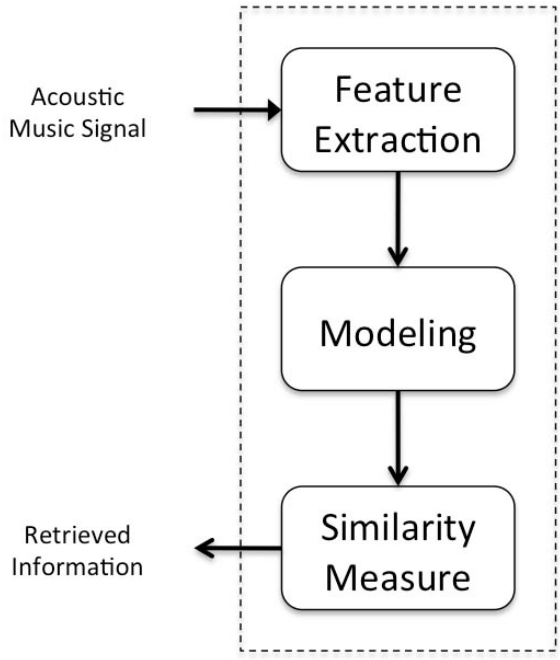


Figure 2.1: System diagram of general music information retrieval system.

well-known features in audio and speech processing [63]. Even if it models the timbre-related attributes of the signal regarding human auditory characteristics, the aspects of musical attributes such as pitch class information can be barely described through MFCC. Instead, many researchers have utilized a *chroma feature vector* which describes an energy distribution on a chromatic scale [76, 30, 49, 47]. It is based on Shepard's helix model (see Fig. 2.2) which factorizes the perception of frequency into *tone height* and *chroma* as follows [68].

$$f = 2^{h+c} \qquad h \in \mathbb{Z}, \quad c \in [0, 1) \tag{2.1}$$

where h, c, and f represent tone height, chroma, and frequency, respectively. We can compute the *chromagram* by first performing a short-time power spectrum analysis,

$$x_c(t) = \sum_h s\left(t, 2^{h+c}\right) \tag{2.2}$$

where $s\left(t, 2^{h+c}\right)$ represents a short time power spectrum at time t. Appropriately quantizing the chroma into twelve levels yields a twelve dimensional vector $\mathbf{x}(t)$ that can closely match the Western chromatic pitch classes (A to $G\#$). These quantized quantities are usually called a chroma feature vector, and each element of the vector represents the energy of the corresponding pitch class at the time instance t. In practice, since the power spectral analysis is performed on a short-time segment the discrete short-time segment index number n is used instead of continuous time t. Therefore, $\mathbf{x}(n)$ represents the chroma feature vector at the segment n.

Some systems quantize the chroma feature into the letter representations of harmony structures, such as C or D_m [48, 77, 10, 56, 57]. Bello used a string matching algorithm to analyze the effects of possible quantization errors in chord representations, such as shift, gaps, swaps, and beats [10]. Based on the results, it was argued that the similarity measurement based on the chord representations works reasonably robustly in music information retrieval systems even with some chord estimation errors.

2.1.1.2 Modeling

The goal of modeling is to capture the characteristics of a music piece with a set of extracted features. Various machine learning algorithms have been adopted to devise an appropriate scheme to capture the characteristics of a music piece. These include hidden markov models (HMMs), support vector machines (SVMs), language models (LMs; N-gram symbol sequence models), and string matching algorithms.

Kim *et al.* used an HMM to model the feature vectors derived from chroma features [47]. Assuming that similar songs will have similar state sequences in the HMM framework, they estimated the similarity between two songs by comparing the histograms of the maximum likelihood state in the HMM models toward identifying the cover songs. N-gram language models along with the quantized chroma feature vectors have been utilized in query-by-example tasks [77]. In their work, Unal *et al.* extracted letter representations

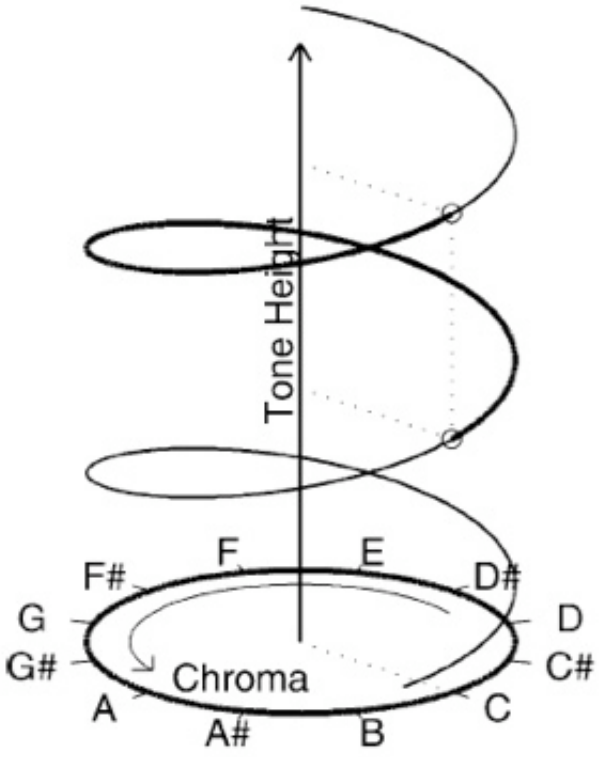


Figure 2.2: Speard's helix model [68] and Western music scale on piano keyboard.

derived by the spiral array model proposed in [20, 23]. They applied an n-gram model and computed a perplexity measure for query-by-example retrieval tasks. Turnbull *et al.* performed automatic semantic annotation and retrieval for music and sound effects [75]. They applied the *mixture hierarchies estimation* method, a supervised multiclass labeling (SML) model, to solve multiclass classification problems. The results showed that their algorithm is useful for content-based annotation and retrieval when each music piece or sound effect has multiple semantic descriptions.

Some systems do not create parameterized models. Instead, raw feature vectors are used directly to extract desired information. In [48], Lee utilized dynamic time warping (DTW) with a sequence of automatically transcribed chord representation proposed in [49]. Assuming that similar songs usually preserve the harmonic content even if they vary in other aspects, he used the chord representation quantized from the chroma feature vectors. An HMM framework was used to compute the cost function for the dynamic time warping (DTW) algorithm in his work. Ellis *et al.* proposed a cross-correlation scheme with the chroma feature vectors [30, 29]. Instead of using dynamic programming schemes, they used a beat synchronous segmentation method in [21] to generate tempo-invariant features. A cover song identification system using the algorithm won the first place at the MIREX 2006. Serrà and Gómez utilized a dynamic programming local alignment (DPLA) toward the cover song identification [67], and they won the first place at the MIREX 2007 [65] and 2008 [66].

2.1.1.3 Similarity measure

The choice of the similarity measure highly depends on the modeling scheme and the intended application. At the same time, however, some of the global-level variations such as key differences can be taken care of in this stage. For example, in a system that utilizes the chord representation for cover song identification task, Lee *et al.* used the first detected chord representation to adjust the possible key differences [48]. Ellis *et al.* measured the similarity with every possible combination of key transitions [29].

In summary, the aforementioned efforts provide a concrete context for the present work in terms of both musically inspired features and models, as well as the engineering methods to handle them.

2.1.2 Contributions of this work

The main contributions of this work include both a novel feature extraction procedure that incorporates dynamic information and a new modeling scheme for extracting context-based music fingerprint. We utilize the chroma feature vector, rather than timbre-related features, to provide musically meaningful and interpretable information. The proposed feature representations to capture the dynamic information in chroma feature vectors are expected to model syntax information of music to some extent.

Although using a sophisticated machine learning approach, such as with an HMM or GMM, may yield accurate performance in some applications, considerable computing power and storage space are demanded. Since the purpose of this work is to devise an efficient and low complexity method which can be implemented even on energy-scarce portable devices, we did not focus on highly profiled machine learning algorithms. In this work, instead, we use second order statistics of features to model music audio to provide an efficient method in terms of both computational power and storage space, as well as a reliable discrimination measurement. In particular, it aims to provide musically intuitive and meaningful descriptions of a given music audio, such as harmony structure and its dynamics, that can enable us to analyze the roles of musical semantic and syntax as music signatures.

To demonstrate the advantages of the proposed context-based music fingerprint, we perform three different experiments: *opus identification*, *composer identification* and *semantic annotation tasks*. These tasks are chosen because this information is often saved as text metadata and the purpose of this work is to devise context-based music fingerprint that can provide an alternative or supplemental access to the information that text metadata provides.

2.2 Feature Extraction: Vertical and Temporal Context

2.2.1 Vertical context feature

We use the chroma feature vector that many researchers use to describe an energy distribution on a Western chromatic scale [76, 30, 49, 47]. This conventional chroma feature vector can be interpreted as representing *vertical context* in the sense that it describes the way of different pitch classes are played simultaneously in terms of energy distribution within a given music audio segment. It is based on Shepard's helix model which factorizes the perception of frequency into *tone height* and *chroma* as describe earlier.

The chroma feature vector describes the energy distribution over the Western pitch classes. Since it is an energy-related quantity, it is often normalized to mitigate sound level differences. Normalization is usually designed to have a unit length vector, i.e.,

$$x_c[n] \leftarrow \frac{x_c[n]}{\|\mathbf{x}[n]\|} \tag{2.3}$$

2.2.2 Temporal context feature

In this section, we introduce two types of dynamic modeling schemes to capture temporal context based on the chroma feature vectors: *delta chroma feature* and *chromatic delta feature*. By considering only one adjacent feature vector in both methods, we expect to obtain the dynamic information between adjacent time segments. The reason we choose to consider only one adjacent segment is not only because of its simplicity but also due to the evidence provided by psychological experiments. In experiments designed to analyze how human beings perceive musical tension in a long chord sequence, Bigand argued that musical events are perceived in local chord structures [13]. In this work, we utilize a beat-synchronous flexible length segmentation based on the beat detection algorithm presented in [28]. With the beat-synchronous segmentation, the features hence model the dynamic information between two adjacent beats.

2.2.2.1 Delta chroma feature

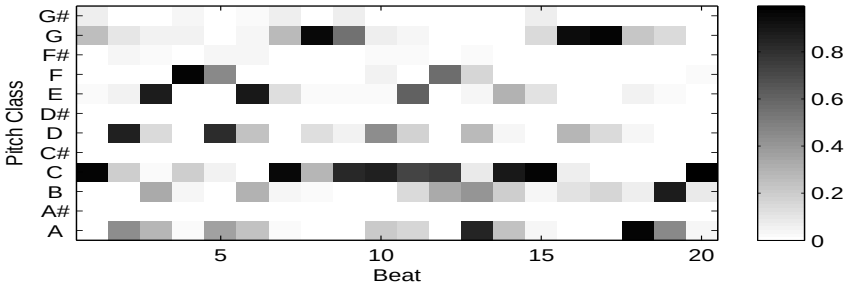
The delta chroma feature can be computed as follows:

$$\Delta \mathbf{x}(n) = \mathbf{x}(n+1) - \mathbf{x}(n) \tag{2.4}$$

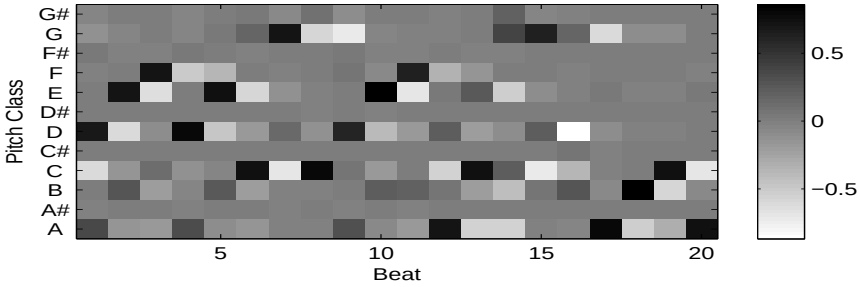
It models the temporal dynamic change in each pitch class (an example is shown in Fig. 2.3(b)). This is akin to the dynamic feature used for automatic speech recognition, except that we presently use only one adjacent frame feature vector rather than applying a several-tap long FIR filter. The reason for this choice of just one adjacent segment is not only because of its simplicity but also due to the evidence provided by psychological experiments. In experiments designed to analyze how human beings perceive musical tension in a long chord sequence, Bigand argued that musical events are perceived in local chord structures [13].

2.2.2.2 Chromatic delta feature

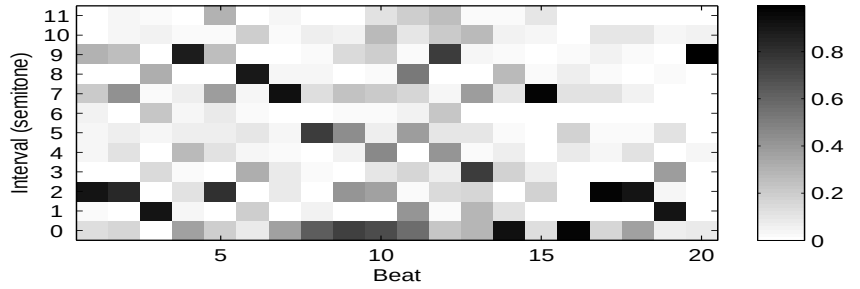
We also introduce a chromatic delta feature motivated by psychophysical observations. It is a well-known fact that humans perceive or produce relative pitch changes with greater ease than absolute pitch values, and this characteristic has been utilized in several music information retrieval systems (e.g. query-by-humming systems [76]). This argument can be partially supported by results in neuroscience. In [81], Warren et. al. used functional magnetic resonance imaging (fMRI) to show the psychophysical effects of pitch changes in the human brain by manipulating the pitch values of the signal that subjects listen to. The results showed specific brain regions of activation attributed to pitch changes: the pitch was represented in the anterior to primary auditory cortex, while the pitch height change was represented in the posterior to primary auditory cortex. These observations inspired us to explore the usefulness of dynamic chroma information as a signal modeling feature.



(a) Chroma feature vectors



(b) Delta chroma feature vectors



(c) chromatic delta chroma feature vectors

(d) Pianoroll

Figure 2.3: Examples of (a) chroma feature vectors, (b) delta chroma vectors, and (c) chromatic delta chroma feature vectors extracted from a synthesized audio signal (BWV772) along with (d) the pianoroll figure of the MIDI data from which the signal is synthesized.

We define chroma change as a relative interval between the pitch classes that are played sequentially in terms of semitone. For example, if the pitch class D is played after C is played, the relative chroma change interval is $+2$ semitones. A scalar value would represent the chroma change in case of a monophonic melody signal. In most cases, however, the music audio signal is polyphonic, representing a mixture of multiple pitches from various instruments. It leads to multiple chroma changes at the same time. For example, if the pitch classes D and F are played simultaneously after C is played, the relative chroma change interval can be both $+2$ semitones and $+5$ semitones. To deal with the simultaneous multiple chroma changes, a vector representation is required. Hence, we propose a new vector representation to describe the degree of chroma changes on all possible intervals.

Note that the magnitude of the delta chroma feature in (2.4), i.e. $\|\Delta x[n]\|$, represents the Euclidean distance between two adjacent chroma feature vectors. It can be also interpreted as the likelihood of not sustaining the same pitch class (zero interval chroma change); the smaller the value it represents, the more likely the pitch classes move toward the zero interval chroma change (no change). In other words, if the value is close to zero, it is likely for the pitch classes to be retained as they are.

We can get similar quantities considering any chroma change interval i by circularly rotating the latter chroma feature vector, i.e.,

$$\left\|\Delta \mathbf{x}^{i}\left[n\right]\right\| = \left\|\mathbf{x}^{i}\left[n+1\right] - \mathbf{x}\left[n\right]\right\| \quad ; \quad 0 \leq i \leq 11, \tag{2.5}$$

where $\mathbf{x}^{i}$ represents the rotated vector $\mathbf{x}$ whose elements are circularly moved by i semitones. The value represents the unlikelihood of moving toward the i chroma change interval. Similar to the zero chroma change interval case shown above, the smaller the value it represents, the more likely are the pitch classes to move toward the i chroma change interval. For simplicity, we define the range of i as in the above equation (2.5). One should note that i is modulus of 12 so that a -2 interval can be interpreted as $+10$ interval and vice versa.

Based on the above quantities, we can define a new vector representation which describes the likelihood of moving toward individual chroma change intervals. Since the above quantities are unlikelihoods, we need a reciprocal function that transforms unlikelihood to likelihood values. In this work, we simply put a negative sign and add the maximum value among the elements to make a vector whose elements are non-negative. Therefore, the proposed dynamic chroma feature vector can be written as

$$\nabla \mathbf{x}\left[n\right] = \left\{\nabla x_0\left[n\right], \nabla x_1\left[n\right], \cdots, \nabla x_{11}\left[n\right]\right\}^T \quad , \tag{2.6}$$

where

$$\nabla x_i\left[n\right] = -\left\|\Delta \mathbf{x}^i\left[n\right]\right\| + X_{\max} \tag{2.7}$$

and

$$X_{\max} = \max_j \left\|\Delta \mathbf{x}^j\left[n\right]\right\| \quad . \tag{2.8}$$

As seen in Fig 2.3, the chromatic delta chroma feature shows the relative chroma change interval between the adjacent time segments while the delta chroma feature shows the temporal dynamic information of each pitch class.

2.3 Music Fingerprint

In the design of handle for music information retrieval system, it is desirable to have a small memory and low complexity as well as high accuracy in capturing the unique characteristics of a given music piece. The work by Jensen *et al.* is particularly relevant in this regard [36]. They applied a filter bank on the trajectory of individual chroma feature vector elements, which yields a small matrix that is efficient in terms of computing power and memory requirement. In this work, we propose to use the covariance matrix of the chroma feature vectors as a representative feature of music piece, i.e.

$$\mathbf{\Phi} = E\left[\left(\mathbf{x} - E\left[\mathbf{x}\right]\right)\left(\mathbf{x} - E\left[\mathbf{x}\right]\right)^T\right] \tag{2.9}$$

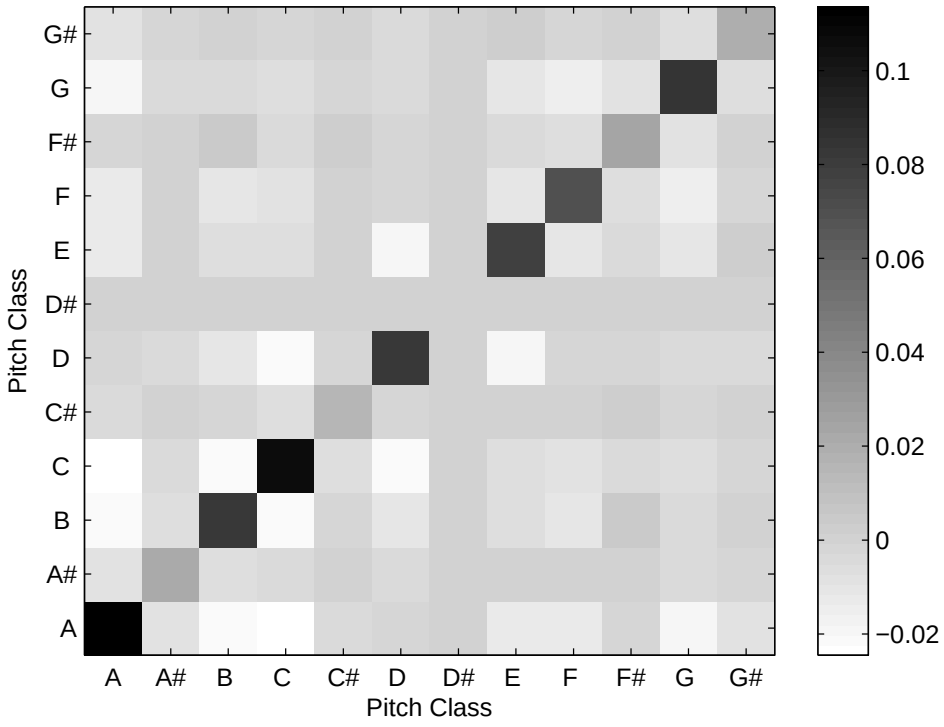


Figure 2.4: An example of music fingerprint using chroma feature vectors (BWV 772).

where T represents the matrix transpose. Although this metric is identical to the covariance matrix in a single multivariate Gaussian approach, the usages differ in the sense that it is used as a template rather than generating probability distribution (detailed description on similarity measure using the proposed metric is provided in Section. 2.4).

In our previous work [40, 39], it is also called a *music fingerprint* in the sense that the metric is a two-dimensional image-like matrix that captures muscial idiosyncrasy. For consistency, hereafter, we call the covariance matrix a music fingerprint.

2.3.1 Vertical Context Model

We start with a fingerprint that is defined by the chroma feature vectors. We hypothesize that this covariance matrix of chroma feature vectors contains useful information about a music piece, such as overall usage of pitch classes and the harmonic structure. This is

based on the analogy between the covariance matrix and harmony: just as the covariance matrix of vectors reveals how individual elements in the vector are related one another, the term harmony is used to describe the event of more than one pitch class being played simultaneously.

- Usage of pitch classes: We hypothesize that the diagonal elements of the covariance matrix represent the degree of activity for each pitch class. The greater the value, the more dominantly or frequently the corresponding pitch class is used in the music piece. For example, in Fig. 2.4, the diagonal elements for *C, D, E, F, G, A* and *B* are relatively greater than others. It implies that those pitch classes are used more dominantly than others in the given music piece.

- Harmonic structure: Each column of the covariance matrix denotes how the various individual pitch classes are related to a given pitch class. Since the relationships between pitch classes represent harmonic information, each column of the music fingerprint can be interpreted as capturing the harmonic structure for a given pitch class. Pitch classes with positive values are likely to be played with a given pitch class simultaneously, while pitch classes with negative values are likely not to be played with a given pitch class simultaneously. Pitch classes with values close to zero do not have specific trends. For example, in Fig. 2.4, *D* is rarely harmonized with *C* or *E*, while it does not have any distinctive tendency of co-presence with other pitch classes.

Fig. 2.5 and 2.6 support our hypotheses by comparing the MIDI data with the synthesized audio signal. The dotted lines represent the results from MIDI data, and the solid lines represent the synthesized audio signals. Even though the dotted and solid lines are not identical, they provide a rough idea about the usage of pitch classes and the underlying harmonic structures. We argue that the difference between the two results is primarily due to signal processing challenges associated with polyphonic audio signals. Some of these challenges are also illustrated in Fig. 2.3. Comparing the pianoroll of the

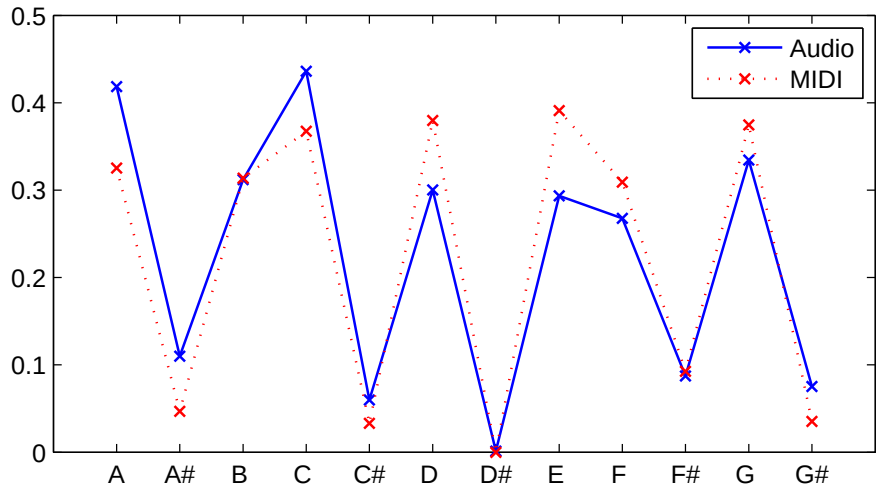


Figure 2.5: An example of usage of notes information; a comparison between music fingerprint and MIDI data (BWV 772).

MIDI data and the chroma feature vectors obtained from the synthesized audio signal, one can easily observe that the chroma feature vectors obtained from the synthesized audio signal show effects of noise. For instance, there are several non-zero quantities in the chroma vectors from the audio signal when the corresponding pitch class is not played (e.g. G, A, B, and C in the first 4 notes). This might be caused by the characteristics of overtones, which impose considerable amount of energy on the perfect 5-th (7 chromatic interval) pitch class. Noise due to release-time differences can be also observed. Residual energy beyond MIDI events and vanishing energy during MIDI events are also evident. This might be caused by the fact that the release time is dependent on the specific type of musical instruments. The chroma feature vectors from polyphonic and multiple-instrument audio signals would be even more complicated than the given example which is nearly monophonic with just one musical instrument.

Even though seeking more robust alternatives to the chroma feature vectors is beyond the scope of this present work, we attempt to minimize the aforementioned effects by using the covariance matrix. Note that it can be easily shown that other second order statistics methods, such as the correlation matrix, would introduce additional assumptions about

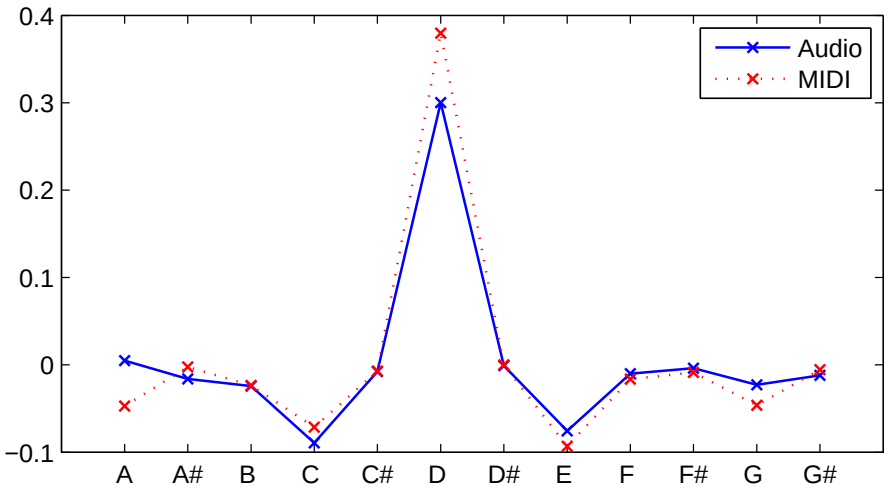


Figure 2.6: An example of harmonic structure information; a comparison between music fingerprint and MIDI data (BWV 772).

the noise by using MIDI-synthesized audio signals. Suppose the chroma feature vector from the audio signal is corrupted by additive noise, i.e.,

$$\mathbf{x}[n] = \overline{\mathbf{x}}[n] + \varepsilon[n] \quad , \tag{2.10}$$

where $\overline{\mathbf{x}}$ and ε represent the chroma feature vector from MIDI data and the noise vector which can be observed in Fig. 2.3, respectively. Then, the music fingerprint can be represented as

$$\Phi = \overline{\Phi} + \Delta \tag{2.11}$$

where $\overline{\Phi}$ and Δ represent the music fingerprint from MIDI data and the noise matrix, respectively.

In the proposed covariance matrix framework, the noise matrix can be written as

$$\Delta = 2\left\{E\left[\varepsilon\overline{\mathbf{x}}^T\right] - E\left[\varepsilon\right]E\left[\overline{\mathbf{x}}\right]^T\right\} + \left\{E\left[\varepsilon\varepsilon^T\right] - E\left[\varepsilon\right]E\left[\varepsilon\right]^T\right\}. \tag{2.12}$$

If other second-order statistics are utilized to model the harmony structure, those can be also easily derived: for example, the correlation matrix can be written as

$$\Delta = 2\left\{E\left[\varepsilon\overline{\mathbf{x}}^T\right]\right\} + \left\{E\left[\varepsilon\varepsilon^T\right]\right\}, \tag{2.13}$$

and the mean matrix is

$$\Delta = 2\left\{E\left[\varepsilon\right]E\left[\overline{\mathbf{x}}\right]^T\right\} + \left\{E\left[\varepsilon\right]E\left[\varepsilon\right]^T\right\}. \tag{2.14}$$

Compared with (2.12), the method using the correlation matrix assumes $E[\varepsilon]E[\overline{\mathbf{x}}]^T = \mathbf{0}$ and $E[\varepsilon]E[\varepsilon]^T = \mathbf{0}$ which are equivalent to zero-mean signal processing noise assumption, which is not necessarily true in practice. The assumption embedded in the method using the mean matrix is even stronger. It assumes $E[\varepsilon\overline{\mathbf{x}}^T] = \mathbf{0}$ which is equivalent to saying $\overline{\mathbf{x}}$ and ε are orthogonal. It also assumes that the signal processing noise is uncorrelated with itself. As shown earlier in Fig. 2.3, however, the assumption is not valid in the given chroma feature extraction algorithm. The noise appears highly correlated with the corresponding pitch class (e.g. considerable amount of energy on the perfect 5-th pitch class of the played pitch class). See [39] for empirical results.

2.3.2 Temporal Context Model

The covariance matrix of the chroma features only captures static information. Instead, we construct super-vectors which consist of the chroma feature vectors and dynamic feature vectors, comprising either the delta chroma features or chromatic delta features proposed in Section 2.2, to specify the music fingerprint. Computing the covariance matrix of the super-vectors generates the proposed music fingerprint:

$$\mathbf{\Phi}_\Delta = E\left[\left(\mathbf{x}_\Delta - E\left[\mathbf{x}_\Delta\right]\right)\left(\mathbf{x}_\Delta - E\left[\mathbf{x}_\Delta\right]\right)^T\right] \tag{2.15}$$

where

$$\mathbf{x}_{\Delta} = \begin{bmatrix} \mathbf{x} \\ \Delta\mathbf{x} \end{bmatrix}. \tag{2.16}$$

or

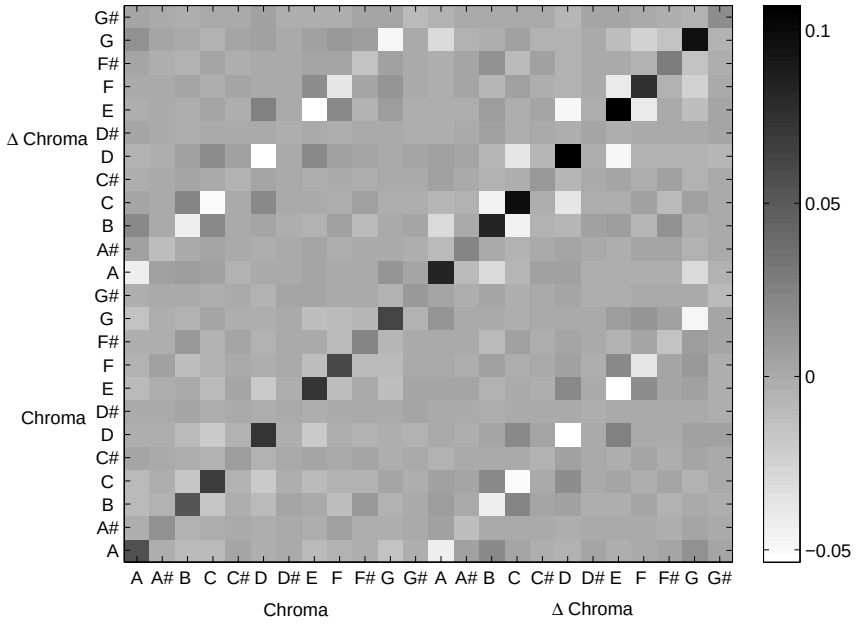
$$\mathbf{\Phi}_{\nabla} = E\left[\left(\mathbf{x}_{\nabla} - E\left[\mathbf{x}_{\nabla}\right]\right)\left(\mathbf{x}_{\nabla} - E\left[\mathbf{x}_{\nabla}\right]\right)^{T}\right] \tag{2.17}$$

where

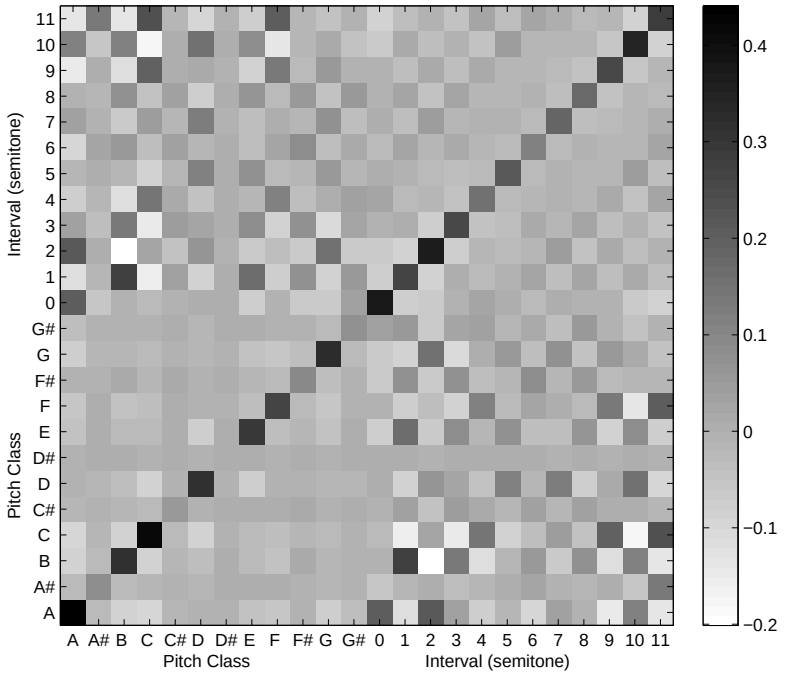
$$\mathbf{x}_{\nabla} = \begin{bmatrix} \mathbf{x} \\ \nabla\mathbf{x} \end{bmatrix}. \tag{2.18}$$

Fig. 2.7 shows examples of music fingerprints using super-vectors, which model the dynamic properties of chroma feature vectors. Since the length of the vector is doubled, the size of the fingerprint is four times larger. We analyze each quadrant separately to interpret it in terms of musical attributes. Note that the third quadrant of each fingerprint, which is in the bottom-left of the plane, is identical to the fingerprint that uses only chroma feature vectors.

The first quadrant of the fingerprint represents the covariance matrix of the delta chroma feature vectors. Since the delta chroma feature vectors represent the relative energy difference between two consecutive segments, these quantities are more related to temporal changes in intensity rather than the intensity itself. Diagonal elements of the quadrant denote the trajectory dynamics of individual pitch classes, and they can be also interpreted as the intensities of on-set and release events. They are somewhat related, but not identical, to the diagonal elements of the third quadrant. One can also predict the dominant pitch classes of the music by observing these quantities. For example, in the first quadrant in Fig. 2.7(a), the diagonal elements for *C,D,E,F,G,A* and *B* are relatively larger than others. It indicates that those pitch classes are used more frequently than others, which is similar to the previous conclusion in Fig. 2.4. Elements in each column represent the tendency of the temporal movements with respect to a given pitch class. Positive values indicate that the corresponding pitch classes tend to move collaboratively with the given pitch class, either on-set or release, while negative values represent that



(a) Chroma feature vectors



(b) Delta chroma feature vectors

Figure 2.7: Examples of music fingerprints using delta chroma feature vectors and chromatic delta feature vectors extracted from music audio (BWV772)

they collude inversely. If the value is close to zero, there is no significant movement along with the given pitch class. For example, C and E are more likely not to be played simultaneously with D than any other pitch classes. It is also highly related with the harmonic structure modeled in the third quadrant since it reveals which pitch classes are harmonized with a given pitch class. From this analysis, we argue that the first quadrant also models usage of pitch classes and harmonic structure based on temporal dynamics, so it may provide complementary information.

When we consider the second or fourth quadrant (note that those two quadrants are symmetric), we can observe the cross-covariance matrix between chroma feature vectors and delta chroma feature vectors. Each vector describes the movements of the pitch classes that follow with respect to a given current pitch class. If the value is positive, there is a tendency of on-set of the pitch class after the given pitch class. If it is negative, there is a tendency of release of the pitch class after the given pitch class. If the value is close to zero, there is no significant movement on the pitch class after the given pitch class. For example, B and D are more likely to be on-set than any other pitch classes after C is played, while C itself is likely to be released. This reveals the global pitch progression between two adjacent segments.

If we use a super-vector of the chroma feature vector and the chromatic delta feature vector, i.e., $\mathbf{x} = \left[\mathbf{x}^T \nabla \mathbf{x}^T\right]^T$, the content of the music fingerprint is somewhat different compared to the one with delta chroma feature vectors (see Fig. 2.7(b) for the examples). Firstly, the axes of the music fingerprint with the proposed delta chroma feature vectors consist of the pitch classes and the relative intervals rather than pitch classes. In the first quadrant of the figure, the diagonal elements represent the intensities of chroma changes. Each vector in the first quadrant describes how the chroma changes happen simultaneously with the corresponding chroma change.

In the second quadrant of the figure (it is symmetric to the fourth quadrant), each vector illustrates which direction the chroma change happens after the corresponding pitch class is played. The greater the value is, the stronger the tendency exists for

the pitch classes that are simultaneously played with the corresponding pitch class to move toward the corresponding interval. It is remarkable that the temporal dynamic information is modeled as a group of the pitch classes that are simultaneously played with the corresponding pitch class. For example, after the pitch class A is played, there exists a tendency for the pitch classes that are played with the pitch class A to be retained as they are, or move toward 2 semitones above.

2.4 Similarity Measure

We use a simple template matching to measure the similarity of the two candidate music fingerprints. The similarity between music i and j is computed as follows.

$$s_{ij} = \sum_{k=1}^{D} \sum_{l=1}^{D} \phi_{kl}^{(i)} \phi_{kl}^{(j)} \quad , \tag{2.19}$$

where ϕ_{kl} represents the k-th row and l-th row element of the music fingerprint $\mathbf{\Phi}$. A greater value represents higher similarity between two pieces of music.

2.4.1 Key compensation

It should be noted that even in playing the same music, it is possible to transpose the key of the music. To compensate for possible key transposition, we circularly shift one of the fingerprints in the diagonal direction by one semi-tone step to get the maximum similarity value.

$$s_{ij} = \max_{m} \sum_{k=1}^{D} \sum_{l=1}^{D} \phi_{kl}^{(i)} \phi_{kl}^{m(j)} \qquad ; \quad 0 \leq m \leq 11, \tag{2.20}$$

where

$$\phi_{kl}^{m} = \phi_{mod((k+m)/12)mod((l+m)/12)} \tag{2.21}$$

and $mod(\cdot)$ represents the modulus of the division. In the case of using delta chroma features, the shifting process should done separately in each quadrant. Especially with the chromatic delta feature vector, special care should be paid to deal with the possible

transposition. Since the chroma change interval is a relative value and independent of the key, it should not be moved during the key compensation process. Therefore, the first quadrant should be retained as it is and the second (or fourth) quadrant should circularly move to the right (or upper) direction to compensate for possible key difference.

2.4.2 Fingerprint Normalization

Since the music fingerprint contains energy-related quantities, it is crucial to normalize the covariance matrix appropriately to balance overall the loudness level, i.e,

$$s_{ij} = \max_{m} \sum_{k=1}^{D} \sum_{l=1}^{D} N\left(\phi_{kl}^{(i)}\right) N\left(\phi_{kl}^{m(j)}\right) \quad , \tag{2.22}$$

where $N(\cdot)$ represents the chosen normalization algorithm. In this paper, we use a column-wise normalization which lays emphasis on the harmonic structure of each associated pitch class by dividing the squared sum (energy) in the column of the fingerprint, i.e.,

$$N(\phi_{kl}) = \frac{\phi_{kl}}{\sqrt{\sum_{m} (\phi_{ml})^2}} \ . \tag{2.23}$$

The column-wise normalization (CN) scheme prevents neglecting information from less dominant pitch classes [40].

2.4.3 Score Fusion Method

As we described in the previous section, each quadrant of the proposed music fingerprint compacts different musical attributes. Since the roles of individual aspects can be different depending on the target application, we introduce a weighted sum of similarities from individual quadrants.

$$s_{ij} = \sum_{q=1}^{4} \lambda_q \cdot s_{ij}(q) \quad , \tag{2.24}$$

where $s_{ij}(q)$ denotes the similarity computed as in (2.20) for quadrant q, and λ_q represents the weighting coefficient with a constraint $\sum_q \lambda_q = 1$.

Although there are four quadrants, we only consider three quadrants because the second and the fourth are symmetric. It simplifies the problem that the degree of freedom is two, i.e.

$$s_{ij} = \lambda_1 \cdot s_{ij}(1) + \lambda_2 \cdot s_{ij}(2) + (1 - \lambda_1 - \lambda_2) \cdot s_{ij}(3) \quad , \tag{2.25}$$

where $s_{ij}(1)$, $s_{ij}(2)$, and $s_{ij}(3)$ represent the similarities using the first, second (or fourth), and third quadrant respectively. The weighting coefficients are determined empirically in each experimental setup.

2.5 Experimental Setup

In this work, we perform three different experiments: *opus identification*, *composer identification*, and *semantic annotation tasks*. The reason we choose these tasks is that the target information for which the individual tasks seek is usually written in text metadata. Since the purpose of this work is to devise context-based music fingerprint that can provide alternative access to the information that such text metadata provides, these empirical tasks were considered reasonable case studies.

To evaluate these tasks, we utilize three different databases: MIDI-synthesized Classical music, real recordings of Classical music, and Western popular songs, each used according to the designated application.

2.5.1 Databases

2.5.1.1 Database I

In Database I, there are approximately 2000 recordings by 11 classical music composers; Bach, Beethoven, Brahms, Chopin, Debussy, Handel, Haydn, Mozart, Schubert, Tchaikovsky, and Vivaldi (Approx. 1000 pieces and 2 variations of each piece). They were originally recorded in the MIDI format [1], and the audio signal for each was generated using Timidity++ toolkit [2] to have 16kHz sampling rate. The length of the pieces varies from 1 minute to 10 minutes, and the pieces whose length exceeds 10 minutes were truncated

to 10 minutes for simplicity. Each piece of music has two different versions with possible changes of tempo, orchestration, and key. In this database, however, we have observed that expressiveness in MIDI data is very restricted or quantized (e.g. velocity values).

2.5.1.2 Database II

Besides the MIDI synthesized audio data, we also have real recordings of classical music. We have collected the works of J.S. Bach, specifically Inventions and Sinfonias. There are 30 pieces of music whose opus numbers are from BWV 772 to BWV 801. In the collection, there are 6 different recordings of the Inventions and Sinfonias yielding 180 audio files total. They differ in various aspects; player, tempo, instruments, and even adding or omitting some notes according to players' expressive intention. They were originally encoded in mp3 format (320kbps, 44kHz sampling rate, stereo) and converted to wave format (44kHz sampling rate, mono).

2.5.1.3 Database III - CAL500

We also use Western popular songs collected by Turnbull *et al.* which consists of 500 different songs [75]. It is called the Computer Audition Lab 500 (CAL500) database, and it includes 1708 subjective annotations evaluated by 66 subjects. The annotations are collected by asking the subjects to label songs with acoustically relevant words. The words can be categorized into 6 categories; emotion, genre, instrument, solo, usage, and vocal. After pruning the words that are represented by fewer than five songs, the total number of the words is 174. See [75] for more details.

2.5.2 MIR applications

2.5.2.1 Opus Identification task

For the opus identification task, we utilize Database I and Database II which include various versions of the same opus. The opus identification task is very similar to the cover song identification task in MIREX evaluation except it only deals with Classical

music. We use one of the versions as a test set and the other remaining versions as training sets. We make a decision by using the maximum similarity score among the training data set, and consider the result correct when it is the same opus with the query among the test set.

2.5.2.2 Composer Identification task

For the composer identification task, we utilize Database I which includes 11 different composers. We perform a two-fold experiment so that each experiment does not have the same opus in the dataset to prevent artificially high performance by choosing opuses which are obviously written by the same composer. In each dataset, we make a decision based on the composer of the most similar piece with one-leave-out method.

2.5.2.3 Semantic Description Annotation task

There are many types of descriptions about music other than simple opus number or composer name. We utilize Database III to evaluate semantic description annotation task performance. We make a decision based on likelihood of annotation words of k nearest neighbors, i.e.,

$$l(w) = \sum_{r=1}^{k} s(r)\, \psi_w(r) \tag{2.26}$$

where $s(r)$ and $\psi_w(r)$ represent similarity between the corresponding song and the r-th closest song and ground truth annotation of r-th closest song considering word w, respectively. After the likelihood of word w is computed, only A words that score the highest likelihoods will be chosen as the semantic description of the corresponding song. We perform a ten-fold cross validation and compute *per-word* precision and recall to evaluate the performance.

2.6 Results and Discussion

2.6.1 Opus Identification Task

Table 2.1 shows the performance of the opus identification task using music fingerprint built with vertical context and their temporal context derivatives in terms of accuracy. In both data sets, using dynamic feature vectors along with chroma feature vectors improves accuracy rate. Although dynamic feature vectors themselves are not as efficient as chroma feature vectors, they play an important role as complementary feature vectors in identifying the opus of a music audio signal. The contributions of different types of dynamic feature vectors may vary with respect to data sets; the accuracy can be maximized by using the chromatic delta feature in Database I, while using the delta chroma feature vectors can maximize the accuracy in Database II.

To provide comparison with conventional algorithms, we use a cover song identification system proposed by Ellis *et al.* [29] as a baseline system. We chose their system because they won first place in the MIREX 2006 evaluation and they published their algorithm in their web page [27]. Since our target application is the opus identification whose scenario is somewhat different from cover song identification tasks, a direct comparison between the proposed scheme and the conventional algorithm may be difficult. Nevertheless, the comparison provides a glimpse of the advantages of the proposed algorithm; the new method is competitive with the conventional method in terms of accuracy rate and significantly outperforms in terms of searching time[1]. Tables 2.1 and 2.2 show that the proposed algorithm outperforms the conventional system in terms of searching time without suffering too much in accuracy (the best accuracy rates are comparable).

In experiments with Database I, the proposed algorithm significantly outperforms the baseline system. The reason is that the proposed music fingerprint captures and compares the global characteristics of given music signals while the conventional system utilizes local similarity to compare two different music pieces. Computing the local similarities can be

[1]The searching time is estimated using Matlab on a Windows machine with Pentium IV 3.06GHz and 2GB RAM. It excludes the time for the chroma feature extraction procedure.

Table 2.1: Performance of the opus identification task using music fingerprint in terms of accuracy. A plus sign indicates that the dynamic feature vectors are used to make super-vectors with chroma feature vectors.

Accuray (%)	Database I	Database II
Chroma	75.0	90.6
Delta	65.8	57.2
+ Delta	78.0	92.8
Chromatic Delta	52.6	57.2
+ Chromatic Delta	78.5	91.7
Baseline [30]	65.0	93.9

Table 2.2: Performance of the opus identification task using music fingerprint in terms of searching time in seconds.

Searching time (sec.)	Database I	Database II
Proposed method	$\sim$ 600	$\sim$ 40
Baseline [30]	$\sim$ 42,000	$\sim$ 900

powerful when only small portions of two different music pieces are similar (many cases can be found in popular songs - cover song identification). Especially with a large data set such as Database I, two music pieces whose opus numbers are different can be falsely identified to be the same when they have high local similarity values.

2.6.2 Composer Identification Task

Table 2.3 shows the performance of composer identification using music fingerprint in terms of accuracy. For this task, we utilize *Database I* which includes approximately 1,000 pieces by 11 different composers. The results show that the proposed music fingerprint is able to model the signature of a Classical music composer well. With only the chroma

feature vectors, the identification rate is significantly higher than chance level. This is reasonable because the way of building the harmony, i.e., *harmony structure*, is often governed by common practice period and reflects inherent characteristics of composers in Western music.

From the results, it can be easily seen that the dynamic information - delta and chromatic delta - can not perform better than chroma feature vectors by themselves. However, the performance improves when the dynamic feature vectors are used as complementary feature vectors. This observation reveals that the characteristics of composers are embedded not only in harmony structure but also in how the harmony structure changes. It is also notable that greater improvement is achieved by using chromatic delta feature vectors as complementary feature vectors rather than simple delta feature vectors.

With regard to the recent results of MIREX evaluations [4, 5], although it is difficult to compare the performance directly due to the different data sets, the proposed method seems to provide a similar range of performance against these state-of-the-art systems.

Table 2.3: Performance of the composer identification task using music fingerprint in terms of accuracy. A plus sign indicates that the dynamic feature vectors are used to make super-vectors with chroma feature vectors.

Accuray (%)	Database I
Chroma	45.4
Delta	39.3
+ Delta	46.8
Chromatic Delta	37.3
+ Chromatic Delta	49.7
MIREX 2007 [4]	19.7 $\sim$ 53.7
MIREX 2008 [5]	34.1 $\sim$ 53.3

2.6.3 Semantic Description Annotation Task

Table 2.4 and Fig. 2.8 show the performance of the semantic description annotation task using music fingerprint in terms of *per-word* precision and recall. The bottom of the table shows the baseline and the performance of a conventional system that was introduced in [75]. Note that we use the *per-word* precision and recall instead of *per-song*, which means the metric focuses on predicting all the words in the vocabulary. The motivation and advantage of this measure are well described in [75], and we use the same measurement for consistency.

Based on the 10 fold cross-validation results, the proposed method is found to outperform the random baseline. Furthermore it was found that using dynamic feature vectors as complementary feature vectors, especially the chromatic delta feature, can improve performance. The results indicate that the proposed music fingerprint is competitive to the conventional model based learning algorithms. The proposed method outperforms the

Table 2.4: Performance of the semantic description annotation task using music fingerprint in terms of *per-word* precision and recall.

	Precision	Recall
Chroma	0.191 (0.007)	0.132 (0.009)
Delta	0.186 (0.011)	0.126 (0.005)
+ Delta	0.195 (0.017)	0.135 (0.013)
Chromatic Delta	0.193 (0.011)	0.134 (0.008)
+ Chromatic Delta	0.201 (0.017)	0.142 (0.016)
ModelAvg [75]	0.189 (0.007)	0.108 (0.009)
MixHier [75]	0.265 (0.007)	0.158 (0.006)
Random [75]	0.144 (0.004)	0.064 (0.002)

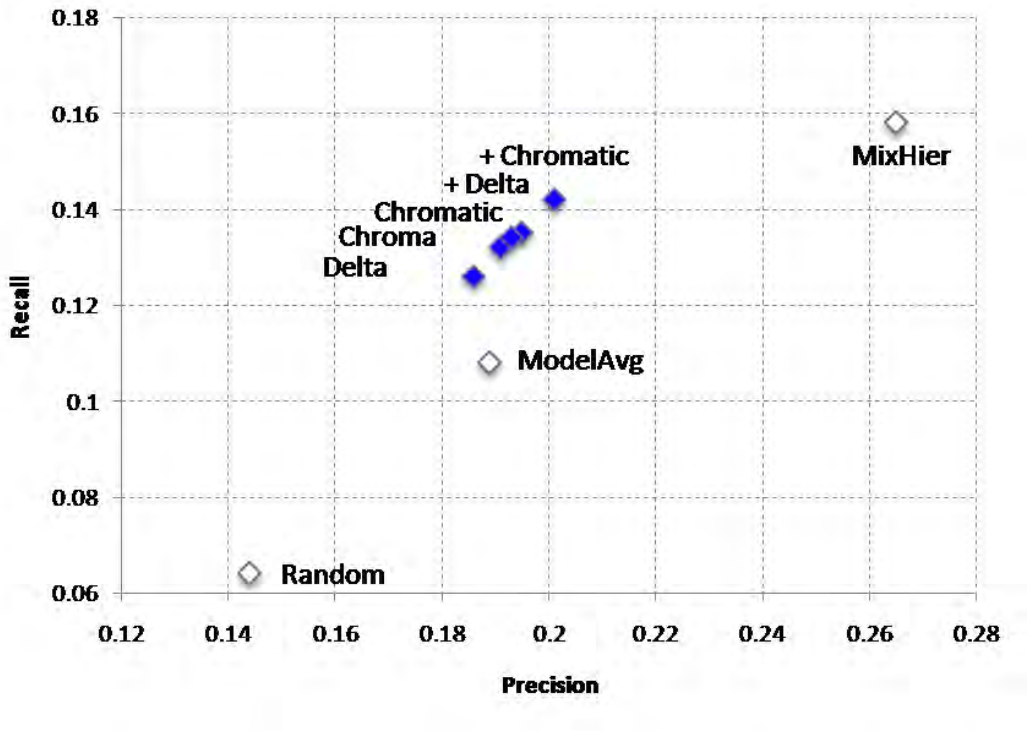


Figure 2.8: Graphical illustration of the performance of the semantic description annotation task using music fingerprint in terms of *per-word* precision and recall. Colored and white dots represent the performance of the proposed method and convensional method, respectively.

conventional method named *ModelAvg* (the improvement can be observed more clearly in recall measurement) although it is not as good as the other method *MixHier*.

2.6.4 Memory requirement

Since the goal of this work is to embed the proposed music fingerprint within a music file, improvements in terms storage memory requirements are worth discussion. While state-of-the-art machine learning based systems need to store whole chroma feature vectors and parameterized models for each song, the proposed system needs to store only the music fingerprint itself whose size is 576 Bytes (assuming double type, without any compression algorithm) for each song. In the case of using additional dynamic feature vectors, it only requires 1728 Bytes.

2.7 Chapter Summary

We introduced a new music information retrieval method using the proposed context-based music fingerprint. The proposed music fingerprint models musically meaningful aspects of a music audio signal, such as harmonic structures and their temporal dynamic information, in a compact representation. It provides an efficient way of extracting information useful for various music information retrieval applications. Through experimental evaluation with MIR frameworks, such as opus identification, composer identification, and semantic annotation, we discussed the performance of the proposed context-based music fingerprint method. The results suggest that the proposed music fingerprint is efficient in terms of complexity, both processing and storage requirements, while yielding performance accuracy that is competitive against state-of-the-art systems. New algorithms which extract dynamic information were also proposed, and it is shown that they can be incorporated to provide complementary information.

Chapter 3

General Audio Information Retrieval

3.1 Introduction

Extracting useful information from unstructured data is receiving significant attention from both research and industrial perspectives. Unstructured data can include various types of media: text, video, and audio that are typically user generated and poorly annotated in comparison to the rich information contained in them. In this chapter, we focus on unstructured audio data which can be present in various multimedia content, such as broadcasting [32], consumer videos [50], and personal sound logs [26]. The main challenge in retrieval of unstructured audio is that the context and the individual acoustic sources (for example human speech, laughter, or other environmental sounds such as car horns) are not known a-priori.

Researchers have been showing promising results in classifying generic audio clips with pre-defined descriptive categories using various machine learning approaches, such as with Gaussian mixture model (GMM) [75] and hidden Markov model (HMM) [53, 82]. For example, Slaney presented a framework to derive semantic descriptions of audio to signal features [69]. Turnbull *et. al.* applied their supervised multi-class labeling method (SML), originally devised for music information retrieval, to sound effects database [75]. In addition, the recent work from Google, Chechik *et. al.* successfully performed a large-scale content-based audio retrieval from text queries for audio clips with multiple tags [19]. Their method is scalable to a large number of audio data based on a passive-aggressive

model for image retrieval (PAMIR). Furthermore, applications like environment sound recognition aim to decode ambient sounds [21]. These types of machine learning algorithms are usually trained in a supervised manner which requires corresponding labels at the training phase. On the other hand, various unsupervised learning methods based on latent variables have also been proposed [73, 50, 84]. Sundaram *et al.* introduced a latent perceptual indexing (LPI) method based on latent semantic analysis (LSA) [73]. Lee *et al.* [50] and Zeng *et al.* [84] applied a modified version of LSA, probabilistic latent semantic analysis (pLSA), for generic audio categorization and consumer video classification using sound track, respectively.

In this chapter, our focus is on modeling context information in general audio signals within a general audio information retrieval framework. Linking audio signal to linguistic descriptions is a perennial challenge in designing content-based audio information retrieval systems. While methodologies to extract acoustic features from audio signals according to pre-defined descriptive categories have been studied intensely, various open challenges still remain. The challenges are often related to ambiguities inherent in both audio signals and linguistic descriptions used to characterize them.

In this regard, we introduce context-based approaches to address these ambiguities toward robust audio information retrieval. The central idea is to capture contextual information embedded within a collection of audio signals and linguistic descriptions in a data driven fashion. The ideas of latent acoustic topic models and intermediate audio descriptive layer are proposed for audio signal modeling and linguistic description, respectively.

3.1.1 Contributions of this work

The first major contribution of this paper is the introduction of a generative model using *context-based* information in audio, a model that is distinct from the well-known content-based methods which are based on modeling realizations of sound sources. These two approaches differ in the sense that a context-based approach seeks latent embedded rules

in the content based on surrounding acoustic properties. It should be noted that excellent advances in content-based retrieval are taking place. We believe that the content-based retrieval provides a complementary source of information and we intent to investigate hybrid systems in our future work. We apply latent topic modeling approaches to model the hidden context in unstructured audio signals. These approaches have been applied and widely used in text document processing, and here we adopt, extend, and evaluate those ideas in an audio information retrieval scenario by drawing analogies between text and audio signals.

The next major contribution of this work is the approach to mitigate the mismatch between descriptions of aural experiences at the time of annotation and the query that contains the desired information at the time of retrieval in dealing with unstructured audio signals. This will benefit addressing the interoperability issue during the annotation and retrieval processes. Our approach also provides users with sufficient flexibility in queries to express their desired information with not only audio examples or categorical descriptions but also their naïve (natural language) text queries. In this work, this is brought about by using an intermediate audio description layer (iADL) which brings descriptions of aural experience and desired sounds into one common platform.

3.1.2 Structure of the Chapter

This chapter is organized as follows. We will describe the proposed framework in Section 3.2. The description of the proposed latent acoustic topic model will be given in Section 3.3 which includes the review of the Latent Dirichlet Allocation (LDA) and detailed implementation of the proposed method. We also discuss about some drawbacks of the proposed latent acoustic topic model and proposed methodologies to overcome those drawbacks in Section 3.4. Experimental setup description and results are provided in Section 3.5 and Section 3.6, respectively. The chapter summary is given in Section 3.7.

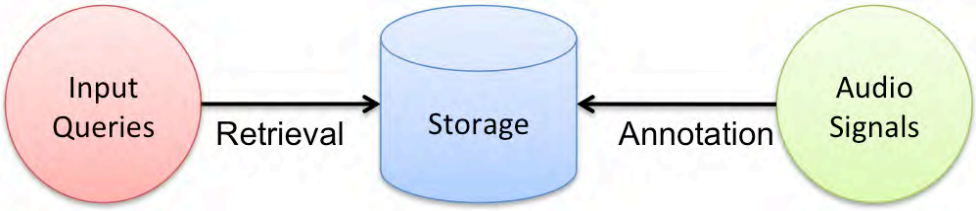


Figure 3.1: A simple diagram of audio information retrieval system.

3.2 Proposed Framework

An audio information system has two major components: annotation and retrieval. Our goal is to build a general audio information retrieval system that annotates audio signals with tags and retrieves a list of audio signals that are related to input queries. In this chapter, we propose novel approaches in both aspects to build a context-based audio information retrieval system. As illustrated in Fig. 3.1, the output of annotation process is usually stored in storage so that the retrieval process can access the storage instead of rerunning the annotation processes for the entire audio database. Although this strategy is reasonable for the system to be efficient, it is still problematic when users inquire atypical; it is often called an *out-of-vocabulary* problem. Therefore, we also introduce an intermediate audio description layer that can provide the interoperability between annotation and retrieval processes.

3.2.1 Intermediate Audio Description Layer

Before going into the detailed descriptions of the system, we introduce an intermediate audio description layer so that annotation and retrieval processes could be built accordingly. Note that processing ambiguous descriptions of embedded information in the array of heterogenous information is challenging. Furthermore, the ambiguities also exist in describing both the embedded information in audio contents and the desired information for input queries. In designing an audio information retrieval system, one should pay attention to these types of ambiguities to prevent a mismatch among them.

As we approach this problem, we introduce a two-dimensional spectrum of descriptions to illustrate the variability of descriptions with respect to how people verbally express the embedded information in sounds. The two-dimentional spectrum is depicted in Fig. 3.2. In the spectrum, we utilize two axes; one is for a psychoacoustic domain and the other is for an evaluation domain. In the psychoacoustic domain, the upper part represents the cognition process which requires users to have prior knowledge or models, while the bottom part represents the sensation process which denotes immediate responses [61]. The evaluation domain indicates the influence of personal background, such as culture, experience and social circumstances, in obtaining information from the audio signals. In the evaluation domain, the left side yields more subjective vocabulary which indicates

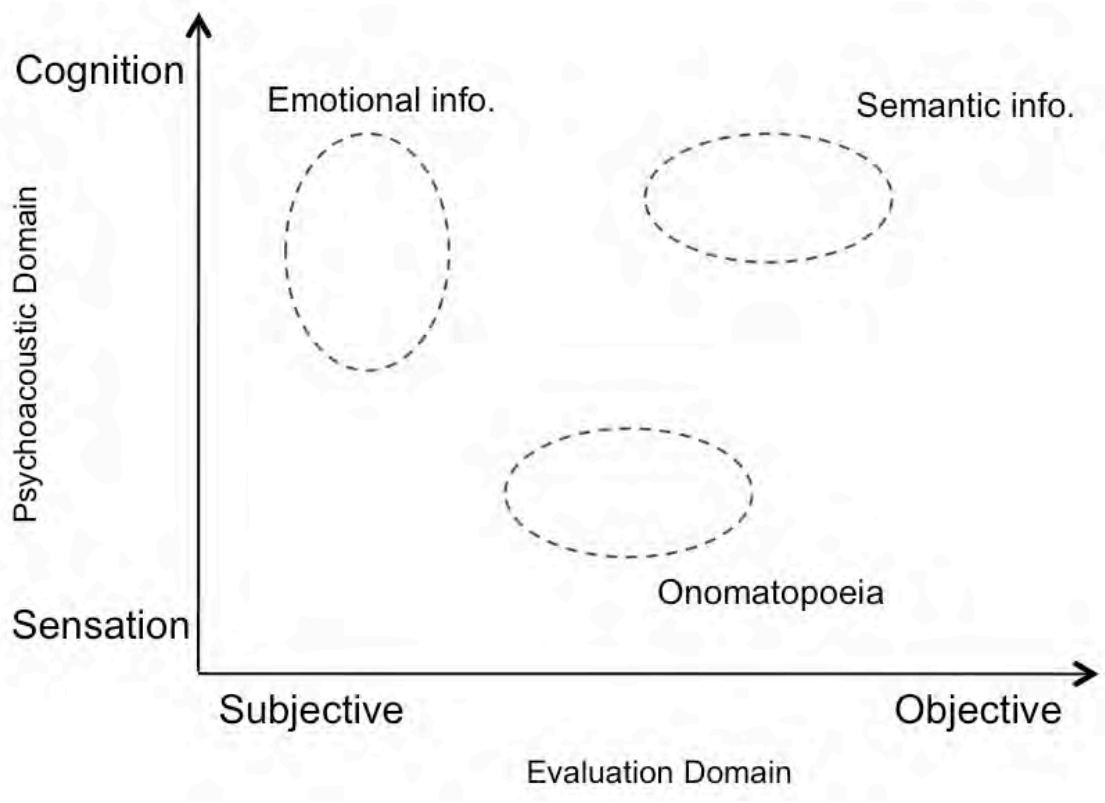


Figure 3.2: An approximated mapping of various desired information onto two-dimensional spectrum (the scales and boundaries of the examples are not exact).

strong influence of personal background. Hence, variations across the describer in the left side are greater than the right side.

In the figure, we provide three different kinds of information that exemplify the wide spectrum of desired information. Although the scales and boundaries are not exact, it provides a brief perspective of the wide variability of the information. These examples, i.e. onomatopoeia, semantic information, and sentiments, are often considered as "description of sound itself", "description of sounding situation", and "description of sound impression", respectively. They are particularly interesting because they are known to carry abundant information about sound. In their human-to-human audio information retrieval tasks, Wake and Asahi showed that users can successfully describe sounds with these three categories [78].

To mitigate the effects of ambiguities in descriptions, we introduce an intermediate audio description layer, which can carry abundant information about sounds in a set of pre-determinded categorical classes. This is motivated by the research on human-to-human communication; Wake and Asahi showed that people can successfully describe sounds with "sounding situation," "sound itself," and "sound impression" to another people [78]. These categories of descriptions are related to the examples shown in Fig. 3.2; "sound itself" with onomatopoeias, "sounding situation" with semantic information, and "sound impression" with emotional information.

Among the variety of information, we focus on two different information aspects of audio data: semantic and onomatopoeia descriptions. The semantic descriptions focus on what makes sounds, while the onomatopoeia descriptions focus on how people describe what they hear. These labels are particularly interesting because they are highly related to psychoacoustic processes, which connect physical properties and human experience of sounds; onomatopoeia labels can be considered from the perspective of the *sensation* process, and semantic labels from *cognition* or *perception* process [61]. We leave the emotional aspect of information for future work.

3.2.2 Annotation

The annotation process is to extract embedded information in given audio clips and to label with tags according to the extracted information. Dealing with general audio signals is particularly difficult due to the heterogeneity of the signals; it includes unstructured audio signals whose constructing rules are, if any, not known or difficult to estimate [58].

In the chapter, we utilize a latent topic model to describe audio context directly obtained from audio signals. It models each audio content as a distribution over a fixed number of unobservable hidden topics. Each topic, in turn, can be modeled as a distribution over a fixed number of acoustic words. The topic model algorithm was originally proposed in the framework of text information retrieval [35, 17, 70]. This idea has been successfully extended to content-based image information retrieval applications [8, 16, 83, 79]. Assuming that hidden "topics" exist behind image features, many researchers have been using the topic modeling approach in their applications. The image features are often quantized to provide discrete index numbers to resemble words in the text topic modeling approach.

Despite the advantages of the latent topic model, to the best of our knowledge, there have been only few efforts that applied topic modeling to content-based sound or audio information retrieval applications. One of the first steps can be found in [72, 73, 74]. Sundaram *et. al.* used the Latent Perceptual Index (LPI) method for classifying audio descriptions inspired by Latent Semantic Indexing (LSI) [9]. In two categories of audio descriptions, i.e., onomatopoeia and semantic descriptions, they demonstrated a promising performance using latent structure in audio information retrieval applications. Levy *et. al.* used an aspect model, which is based on probabilistic latent semantic indexing (pLSI) on music information retrieval [51]. To built the aspect model, they utilized the proposed *muswords* extracted from music audio signals and words from social tags.

In this chapter, we propose an *acoustic topic model* motivated by drawing analogies between text and sound. We hypothesize that short segments of audio signals play a similar role as words in text and that latent topics in audio signals which would be

determined by the context of audio signals. In other words, each audio clip is viewed to consist of latent acoustic topics that generate acoustic words. We use Latent Dirichlet Allocation (LDA) method [35, 17, 70, 15] to model the acoustic latent topics and perform audio information retrieval tasks.

3.2.3 Retrieval

The retrieval process is to extract a list of audio clips that are related to users' input queries. In this work, we attempt to provide users with flexibility in their queries, so that people can use both naive text descriptions and audio examples as queries. Fig. 3.3 illustrates both types of queries in a retrieval process. In case of using audio example queries, the system can annotate the input audio example with the pre-determined tags and retrieve a list of audio clips related to the extracted information from the storage. In case of using naive text queries, however, it is somewhat problematic to deal with naive text queries because users' naive descriptions to express their desired information may vary. This may cause the *out-of-vocabulary* problem issued in [19] for audio retrieval from text queries. Similar issues exist in text information retrieval applications [37, 6, 7]. In [37], Jones *et. al.* proposed a substitution method which replaces user's original descriptions for query systems. Bai *et. al.* utilizes a word similarity measurement to extend their text queries [6, 7].

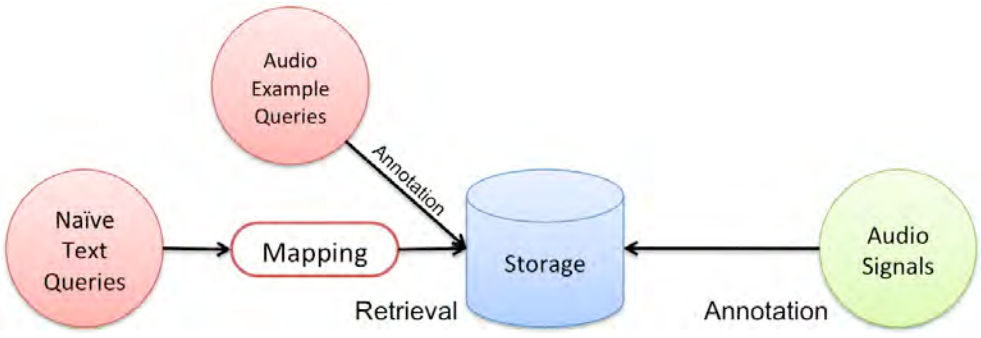


Figure 3.3: A simple diagram of audio information retrieval system (with various input queries).

To this end, we propose a transformation strategy that maps naive text queries to pre-determinded classes on the intermediate audio description layer; specifically onomatopoeia and semantic labels. As it is described earlier, the rationale behind is that onomatopoeias and semantic labels represent unique characteristics of an audio signal and the text descriptions can be mapped into both semantic labels and onomatopoeia. In turn, both annotation and retrieval processes have a shared labels in common so that they are interoperable.

3.3 Latent Acoustic Topic Model

3.3.1 Latent Dirichlet Allocation (LDA)

The topic model assumes that documents consist of hidden topics and each topic can be interpreted as a distribution over words in a dictionary [35]. This assumption enables the generative model like Latent Dirichlet allocation (LDA). Fig. 3.4 illustrates a basic concept of the LDA in a graphical representation, a three-level hierarchical Bayesian model.

Let V be the number of words in a dictionary and w be a V-dimensional vector whose elements are zero except the corresponding word index in the dictionary. A document consists of N words, and it is represented as $\mathbf{d} = \{w_1, w_2, \cdots, w_i, \cdots, w_N\}$ where w_i is the i th word in the document. A data set consists of M documents and it is represented as $S = \{\mathbf{d_1}, \mathbf{d_2}, \cdots, \mathbf{d_M}\}$.

In this work, we define k latent topics and assume that each word w_i is generated by its corresponding topic. The generative process can be described as follows:

1. For each document **d**, choose $\theta \sim Dir(\alpha)$

2. For each word w_i in document **d**,

   (a) Choose a topic $t_i \sim Multinomial(\theta)$

   (b) Choose a word w_i with a probability $p(w_i|t_i, \beta)$,
   where β denotes a $k \times V$ matrix whose elements represent the probability of a word with a given topic, i.e. $\beta_{nm} = p(w_i = m|t_i = n)$.

In LDA, the most challenging question is how to estimate or infer latent parameters like θ, **t**, α, and β while the only variable we can observe is **w**. In this section, we provide a step-by-step description of LDA.

Suppose that we have k latent topic. We define θ as a document specific k-dimensional random variable, which provides a probability for choosing topics for each word in the corresponding document. A topic for each word is determined by choosing one topic among a set of topics based on the probability that is provide by θ; this process can be considered as a multinomial process. Therefore, it is reasonable to use a Dirichlet random variable, which is a conjucate prior distribution of multinomial distribution.

A k-dimensional Dirichlet random variable θ is in the $k-1$ simplex because $\sum_{n=1}^{k} \theta_n = 1$ and $\forall n \;\; \theta_n \geq 0$. The probability of θ is given as following:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{n=1}^{k} \alpha_n)}{\prod_{n=1}^{k} \Gamma(\alpha_n)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \tag{3.1}$$

where α is a k-dimensional vector and $\Gamma(\cdot)$ is the Gamma function.

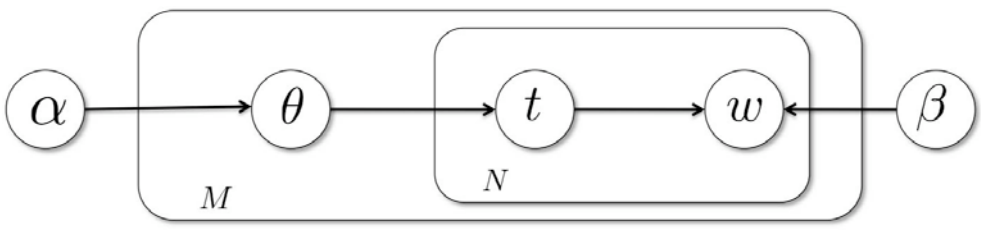


Figure 3.4: Graphical representation of the topic model using Latent Dirichlet Allocation.

For each word, the corresponding topic t is a k-dimensional multinomial random variable whose number of trials is just one; only one element t_τ is 1 where τ is the index of chosen topic, and the rest of the elements in t are zeros. Therefore, the probability can be represented as following:

$$\begin{aligned} p(t_\tau = 1|\theta) &= \frac{1!}{t_1! \cdots t_k!} \theta_1^{t_1} \cdots \theta_k^{t_k} \\ &= \theta_\tau \end{aligned} \tag{3.2}$$

where the subscripts represent the indices of elements in a vector. In general, the probability of t can be written as

$$p(t|\theta) = \prod_{n=1}^{k} (\theta_n)^{t_n}. \tag{3.3}$$

According to the chosen topic, we can draw a word form in a dictionary based on word probabilities β. The word probability β is a $k \times V$ matrix where $\beta_{ij} = p(w_j = 1|t_i = 1)$. A V-dimensional word vector w is also a multinomial random variable. It can be represented as only one element w_υ is 1 and the rest of the elements in w are zeros, where υ is the index of chosen word. The probability of the word with a given topic and word probability β can be represented as following:

$$\begin{aligned} p(w_\upsilon = 1|t_\tau = 1, \beta) &= \frac{1!}{w_1! \cdots w_V!} \beta_{\tau 1}^{w_1} \cdots \beta_{\tau V}^{w_V} \\ &= \beta_{\tau \upsilon} \end{aligned} \tag{3.4}$$

In general, the word probability can be written as

$$p(w|t_\tau = 1, \beta) = \prod_{m=1}^{V} (\beta_{\tau m})^{w_m}. \tag{3.5}$$

Let us consider that we have a set of N words in $\mathbf{w}$ and its corresponding set of N topics $\mathbf{t}$; a topic t_i represents a topic for word w_i where the subscripts denote the index

of a vector in a set. With given parameters α and β, the joint probability of θ, $\mathbf{w}$, and $\mathbf{t}$ can be derived as follows.

$$\begin{aligned} p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta) &= p(\theta|\alpha, \beta) \cdot p(\mathbf{t}|\theta, \alpha, \beta) \cdot p(\mathbf{w}|\theta, \mathbf{t}, \alpha, \beta) \\ &= p(\theta|\alpha) \cdot p(\mathbf{t}|\theta) \cdot p(\mathbf{w}|\mathbf{t}, \beta) \\ &= p(\theta|\alpha) \prod_{i=1}^{N} p(t_i|\theta) p(w_i|t_i, \beta) \end{aligned} \tag{3.6}$$

where the subscripts represent the indices of the vector in a sequence of vectors.

If we marginalize the latent components, we need to sum over topics in each word level and to integrate over θ in a document level:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \prod_{i=1}^{N} \sum_{t_i} p(t_i|\theta) p(w_i|t_i, \beta) \; d\theta \tag{3.7}$$

If we simplify the above equation in terms of model parameters using (3.3) and (3.5), the marginal distribution can be written as followings.

$$\begin{aligned} p(\mathbf{w}|\alpha, \beta) &= \int p(\theta|\alpha) \prod_{i=1}^{N} \sum_{t_i} p(t_i|\theta) p(w_i|t_i, \beta) \; d\theta \\ &= \int p(\theta|\alpha) \prod_{i=1}^{N} \sum_{n=1}^{k} p(t_{in} = 1|\theta) p(w_i|t_{in} = 1, \beta) d\theta \\ &= \int p(\theta|\alpha) \prod_{i=1}^{N} \sum_{n=1}^{k} \theta_n \; p(w_i|t_{in} = 1, \beta) d\theta \\ &= \int p(\theta|\alpha) \prod_{i=1}^{N} \sum_{n=1}^{k} \theta_n \prod_{m=1}^{V} (\beta_{nm})^{w_{im}} d\theta \\ &= \int p(\theta|\alpha) \prod_{i=1}^{N} \sum_{n=1}^{k} \prod_{m=1}^{V} (\theta_n \beta_{nm})^{w_{im}} d\theta \\ &= \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int \prod_{n=1}^{k} (\theta_n)^{\alpha_n - 1} \prod_{i=1}^{N} \sum_{n=1}^{k} \prod_{m=1}^{V} (\theta_n \beta_{nm})^{w_{im}} d\theta \end{aligned} \tag{3.8}$$

where the subscripts represent the indices of the vector in a sequence of vectors and the second subscripts represent the indices of element in the corresponding vector.

Now, the question is how to estimate or infer parameters like $\mathbf{t}$, α, and β while the only variable we can observe is w. In many estimation processes, parameters are often chosen to maximize the likelihood values of a given data $\mathbf{w}$. The likelihood can be defined as

$$l(\alpha, \beta) = \sum_{w \in \mathbf{w}} \log p(w|\alpha, \beta) \ . \tag{3.9}$$

Once α and β are estimated, the joint probability of θ and $\mathbf{t}$ with given $\mathbf{w}$ should be estimated as

$$p(\theta, \mathbf{t}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \ . \tag{3.10}$$

These processes, however, are computationally impossible because both inference and estimation require computing $p(\mathbf{w}|\alpha, \beta)$, which includes intractable integral operations (See (3.8)). To solve this problem, various approaches, such as Markov Chain Monte Carlo (MCMC) [70], gradient descent optimization method [54] and variational approximation [15] have been proposed. In this work, we try a *variational approximation method* and a *Gibbs sampling method*, a specific form of MCMC, to estimate and infer the parameters of the topic model.

The rationale behind of the variational approximation method is to minimize distance between the real distribution and the simplified distribution using Jensen's inequality [15, 24]. The simplified version has γ and ϕ, which are the Dirichlet parameter that determines θ and the multinomial parameter that generates topics respectively, as depicted in Fig. 3.5.

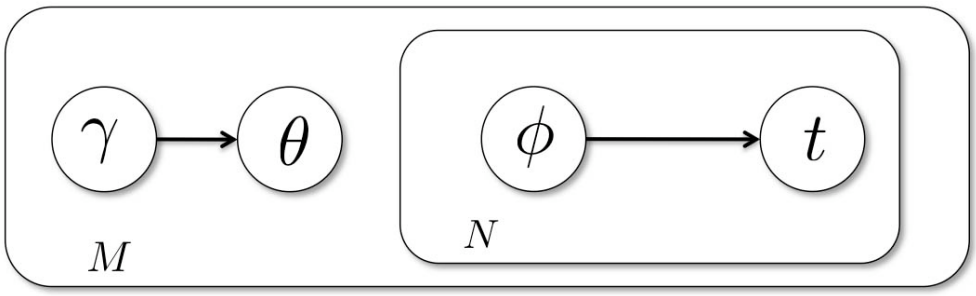


Figure 3.5: Graphical representation of the approximated topic model for variational inference method to estimate and infer the Latent Dirichlet Allocation parameters.

The mathematical details of the methods are provided in Appendix A. On the other hand, Gibbs sampling method is based on MCMC which is a iterative process of obtaining samples by allowing a Markov chain to converge to the target distribution [34, 70].

3.3.2 Implementation of Latent Acoustic Topic Model

In this subsection, we provide in-depth descriptions of the proposed algorithm. Since we utilize the latent topic model, which is originally proposed for text document modeling applications, it requires word-like, discrete indexing numbers to apply the latent topic model as it is done in image retrieval applications. In this work, we introduce the notion of *acoustic words* to tackle this problem. After extracting feature vectors that describe acoustic properties of a given segment, we assign acoustic words based on the closest word in the pre-trained acoustic words dictionary. With the extracted acoustic words, we perform the Latent Drichlet Allocation (LDA) to model hidden acoustic topics in an unsupervised way [45]. Then, we use the posterior Dirichlet parameter which describes the distribution over the hidden topics of each audio clip as a feature vector of the corresponding audio clip. Fig. 3.6 illustrates a simple notion of the proposed acoustic topic model procedure, and the detailed descriptions are given in the following subsections.

3.3.2.1 Acoustic Features

We use mel frequency cepstral coefficients (MFCC) to extract acoustic properties in a given audio signal. The MFCCs provide spectral information considering human auditory characteristics, and they have been widely used in many sound related applications, such as speech recognition and audio classification tasks. The reason we have chosen MFCC is to investigate the effects of spectral characteristics in the latent topic model approach. In this work, we applied 20 ms hamming windows with 50% overlap to extract 12-dimensional feature vectors.

3.3.2.2 Acoustic Words

With a given set of acoustic features, we trained a dictionary using a vector quantization algorithm called LBG-VQ [33]. Similar ideas to make acoustic words can be also found in [19, 73, 72]. The rationale is to cluster audio segments which have similar acoustic characteristics and to represent them as discrete indexing numbers. Once the dictionary is built, the extracted acoustic feature vectors from sound clips can be mapped to acoustic words by choosing the closest word in the dictionary. In this work, we set the number of words in the dictionary as an experimental setting. For simplicity, we choose one of values in a set $V \in \{200, 500, 1000, 2000, 4000\}$. After extracting acoustic words, we generate a *word-document co-occurance matrix*, which describes a histogram of acoustic words in individual audio clips. The word-document co-occurance matrix is fed in to the Latent Dirichlet Allociation (LDA) algorithm to model the acoustic topics.

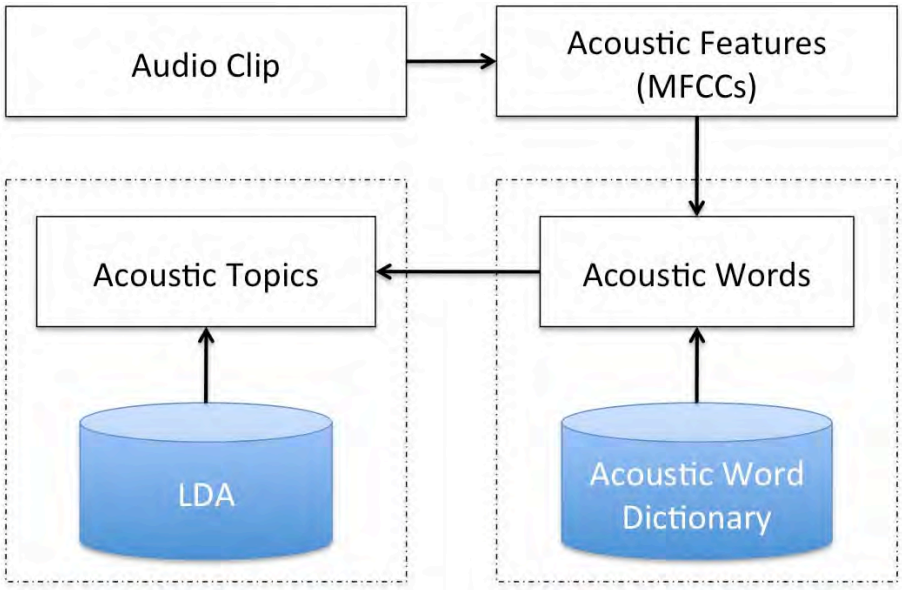
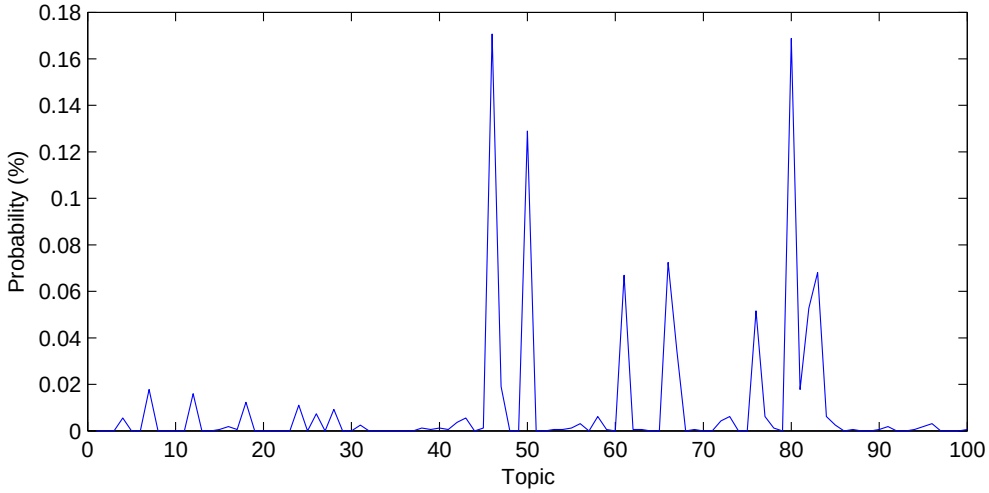
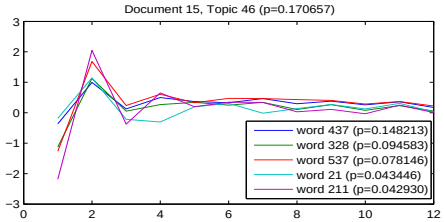


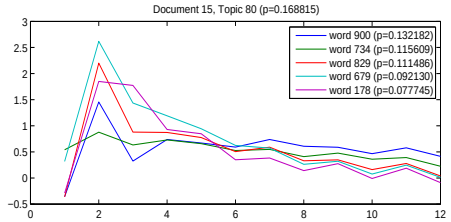
Figure 3.6: Diagram of the proposed acousic topic model algorithm



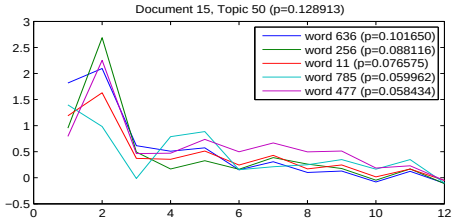
(a) topic distribution in a given audio document



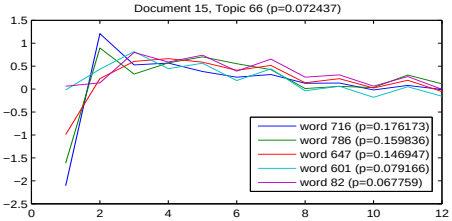
(b) the 5 most probable acoustic words in the most probable topic



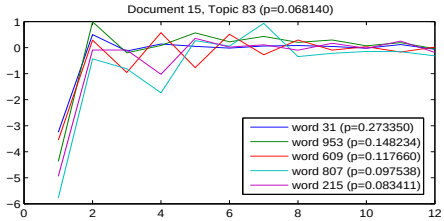
(c) the 5 most probable acoustic words in the second most probable topic



(d) the 5 most probable acoustic words in the third most probable topic



(e) the 5 most probable acoustic words in the fourth most probable topic



(f) the 5 most probable acoustic words in the fifth most probable topic

Figure 3.7: Examples of acoustic topic model: (a) topic distribution in a given audio document (b)-(f) the 5 most probable acoustic words with their probabilities in the 5 most probable topics (the number of acoustic words is 1,000 and the number of latent topic is 100).

3.3.2.3 Latent Acoustic Topic

Each sound clip is assumed to be a mixture of acoustic topics. Since the acoustic topics are hidden variables, they are learned in an unsupervised manner (although the number of latent topics should be set manually). As described in the previous section, we use both *variational inference method* and *Gibbs sampling method* to estimate and infer the parameters of the acoustic topic model.

Fig. 3.7 illustrates an example of acoustic topic modeling results (the number of acoustic words is 1,000 and the number of latent topic is 100; we use a sound clip from the database in Section 3.5 whose filename is 1-GOAT-MACHINE-MILKED-BB.wav). Fig. 3.7 (a) depicts a topic distribution in a given audio document, while Fig. 3.7 (b)-(f) represent the 5 most probable acoustic words with their probabilities in the 5 most probable topics. In Fig. 3.7 (a), there are only several topics are evidently present among 100 latent topics. In turn, as it is illustrated in Fig. 3.7 (b)-(f), each topic has a probability distribution over acoustic words (12-dimensional MFCC). With this acoustic topic model, an audio signal can be modeled as a probability distribution over the latent acoustic topics. In this work, we utilize the probability distribution as a representative feature of the audio signal.

For comparison, we use the Latent Perceptual Indexing (LPI) scheme proposed in [72] as a baseline. It is based on Singular Value Decomposition (SVD) that would reduce the feature vector dimension. The procedure is identical up to the step of generating the word-document co-occurance matrix. In the sense that the dimensions of feature vectors are reduced after the process, the topic model can be considered as a feature dimension reduction as well. However, they differ in interpretations of feature vectors that represent the corresponding audio clip. The topic model uses statistical inference, while Latent Perceptual Indexing (LPI) extracts feature vectors deterministically using Singular Value Decomposition (SVD). The differences in experimental results will be described later in this chapter.

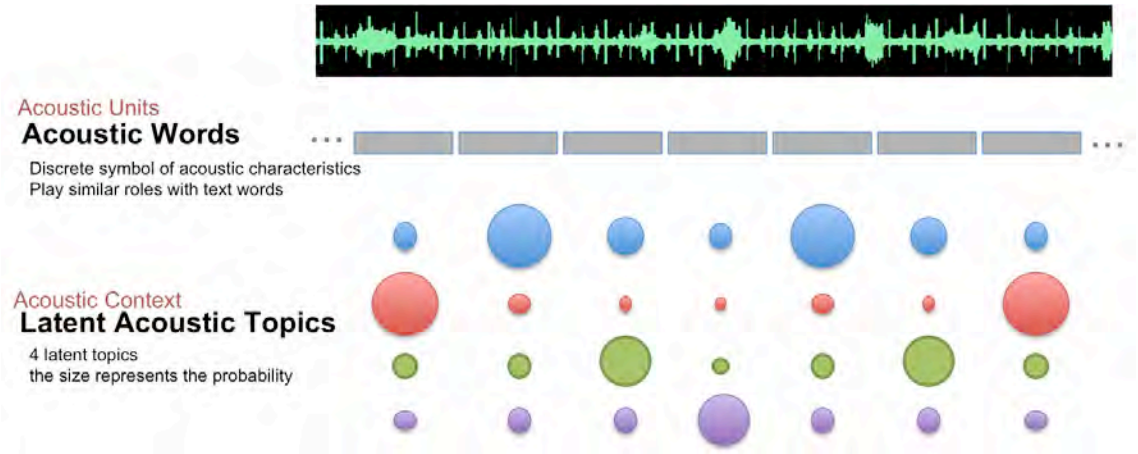


Figure 3.8: An example of latent acoustic topic model from the posterior point of view. Acoustic words represent discrete symbols of acoustic characteristics (vector quantized MFCC in this work). The size of circles indicate the probability that an acoustic word can be assigned to the corresponding latent topic (four latent acoustic topics in this example)

3.4 Modified Methodologies for Improving Latent Acoustic Topic Model

In this section, we discuss about some drawbacks of the proposed latent acoustic topic model and propose modified methodologies to improve the original idea. Particularly, we tackle the bag-of-words approach and unsupervised approach as described in the following subsections.

3.4.1 N-gram approach

Researchers have been showing promising results by treating audio signals analogous to text documents [50, 19, 73, 45, 42]. Many of them used mel-frequecy cepstral coefficients (MFCC) to capture acoustic properties and transformed the coefficients into discrete indices. Once the audio signals are represented with a sequence of discrete indices like text documents, many text modeling algorithms can be applied; Chechik *et al.* used the passive-aggressive model for image retrieval (PAMIR) [19]; Sundaram *et al.* introduced a latent perceptual indexing (LPI) method based on latent semantic analysis (LSA) [73], and Lee *et al.* applied probabilistic latent semantic analysis (pLSA) [50]. Recently,

we have proposed the acoustic topic models using latent Dirichlet allocation (LDA) to characterize unstructured audio signals [45]. Assuming that there exist latent acoustic topics and each audio clip is a mixture of those latent topics, we could demonstrate promising results in audio classification tasks.

One of the drawbacks of these algorithms, including our acoustic topic modeling scheme, is that the bag-of-words approach which does not consider temporal dynamics of features. In this work, we introduce an N-gram approach to account for temporal dynamic information of audio features. The closest work to this idea has been done by Reed and Lee [64]. For music information retrieval applications, they proposed a new iterative segmentation method based on Viterbi decoding and Baum-Welch estimation. With the proposed segments, they apply the bi-gram approach to capture the temporal dynamic information in an LSA framework.

3.4.1.1 Uni-gram approach

Once the dictionary is built, the extracted acoustic feature vectors from the test sound clips can be mapped to acoustic words by choosing the closest word in the dictionary so that individual short time segments have their assigned indices, acoustic word. In this work, we call this method uni-gram approach to contrast with the proposed N-gram approach. After extracting uni-gram words, we generate a *word-document co-occurrence matrix* which describes a histogram of acoustic words in individual audio clips. The word-document co-occurrence matrix is used with the LDA to model audio clips as a distribution of latent acoustic topics. After the LDA modeling, we use the Dirichlet parameter γ as the representative feature vector of a single sound clip to be used consequent classifiers.

3.4.1.2 N-gram approach

To model the dynamic information embedded in acoustic words, we introduce the N-gram approach which describes partial dynamics of acoustic words by considering consecutive

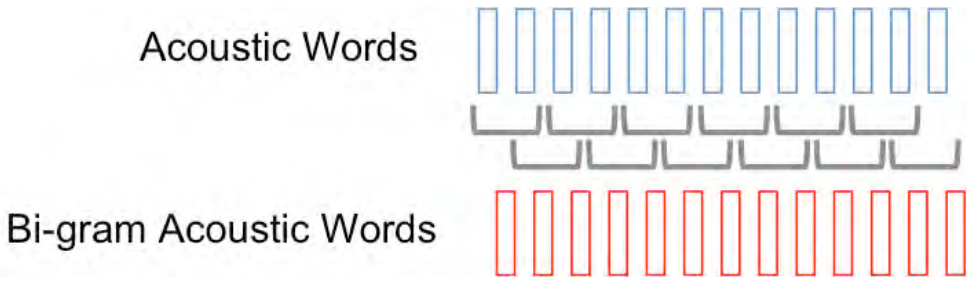


Figure 3.9: An illustration of extracting n-gram (bi-gram in this work).

words. In this work, without the loss of conceptual generality, we consider only one adjacent word to make bi-grams, since $N = 2$ case is a good starting point to explore the usefulness of local context and computational complexity (since dictionary size increases exponentially) [46].

A new acoustic dictionary $\widetilde{\mathcal{W}}$ can be built based on the bi-grams whose elements are from the original acoustic dictionary $\mathcal{W}$. The i-th word in the new dictionary $\widetilde{w_i}$ is defined as follows:

$$\widetilde{w_i} = \{(w_n, w_m) \,|w_n, w_m \in \mathcal{W}\} \tag{3.11}$$

where

$$\begin{aligned} n &= \lfloor i/V \rfloor \\ m &= \mathrm{mod}(i/V) \quad , \end{aligned} \tag{3.12}$$

$\lfloor \cdot \rfloor$ and $\mathrm{mod}(\cdot)$ represent the maximum integer that does not exceed the value of the division and the modulus of the division, respectively. Note that the size of the new dictionary is V^2.

Once the new dictionary for bi-grams is built, the extracted acoustic feature vectors from the test sound clips should be first mapped to acoustic words by choosing the closest word in the dictionary and then consider the adjacent acoustic words to generate the bi-grams. After extracting bi-gram words, we follow the same procedure as uni-gram approach; we generate a word-document co-occurrence matrix and feed into the LDA framework.

In this work, we set the number of words in the dictionary 200 for simplicity. Consequently, the size of bi-gram dictionary is 40,000.

3.4.1.3 Hybrid approach

We also introduce a way of combining both uni-gram and bi-gram approaches. In this work, we propose to make a super-vector of Dirichlet parameters after the LDA inference process; γ_{unigram} and γ_{bigram} from using uni-gram and bi-gram approaches, respectively. The dimension of the features which are fed into classifiers is, therefore, $2 \times k$ where k represents the number of latent acoustic topics.

3.4.2 Supervised Latent Acoustic Topic Model

The proposed unsupervised latent acoustic topic model requires consequent classifiers, such as k-nearest neighborhood (kNN) or support vector machines (SVM), to perform pattern recognition. Although the categorial labels are not necessary for modeling audio signals, the labels are required to train the consequent classifiers. Therefore, the classification performance also depend on the specific classifiers rather than audio modeling procedure itself.

In this subsection, we propose the supervised version of acoustic topic model to associate the categorical labels of sound clips with latent acoustic topics; specifically we apply the supervised LDA (sLDA) method introduced in [14, 80]. The rationale behind this is that considering categorical labels in learning latent variables might endow discriminant power rather than treating the acoustic topic modeling and the classification processes separately and independently [43]..

As pointed out in the previous section, the LDA-based acoustic topic model is trained in an unsupervised manner which does not require any labels during learning phase. The proposed supervised acoustic topic model utilizes a modified version of LDA as shown in Fig. 3.10 which shares most of properties with unsupervised LDA except it includes a node c that represents the category of a document and a kernel function η that transfers

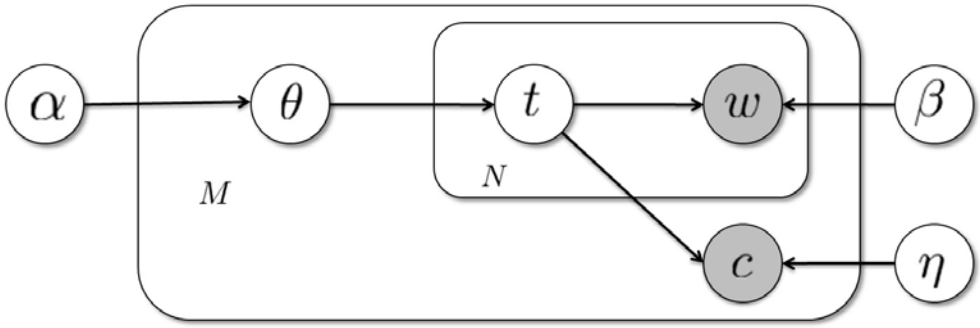


Figure 3.10: Graphical representation of topic models: supervised LDA.

the topic distribution t to the categories. The generative process can be described as follows:

1. For each document **d**, choose $\theta \sim \mathrm{Dir}(\alpha)$

2. For each word w_i in document **d**,

   (a) Choose a topic $t_i \sim \mathrm{Multinomial}(\theta)$

   (b) Choose a word w_i with a probability $p(w_i|t_i, \beta)$

3. Choose class label $c|t \sim softmax(\bar{t}, \eta)$,
   where $\bar{t}$ represents the topic frequency of a document, i.e., $\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n$. The probability of a certain class with give $\bar{t}$ and η can be represented as

$$p\left(c|\bar{t}, \eta\right) = \frac{\exp({\eta_c}^T \bar{t})}{\sum_{c'=1}^{C} \exp({\eta_{c'}}^T \bar{t})} \tag{3.13}$$

3.4.2.1 Inference

Like the conventional Latent Dirichlet Allocation method, computing exact values is not computationally feasible either because it involves intractable integral operations. To solve this problem, various approaches such as Laplace approximation and Gibbs sampling method, have been proposed. In this work, we utilize the variational inference method.

The simplified version has γ and ϕ which, respectively, are the Dirichlet parameter that determines θ and the multinomial parameter that generates topics, as depicted in Fig. 3.5. Note that this variational approximate method is valid for both unsupervised LDA and supervised LDA, since the node c in supervised LDA is not associated with any latent variable.

The joint probability of θ and $\mathbf{t}$ can be represented as

$$\begin{aligned} q(\theta, \mathbf{t}|\gamma, \phi) &= q(\theta|\gamma)q(\mathbf{t}|\phi) \\ &= q(\theta|\gamma)\prod_{i=1}^{N} q(t_i|\phi_i) \end{aligned} \tag{3.14}$$

and tries to minimize the difference between real and approximated joint probabilities using Kullback-Leibler (KL) divergence, i.e.

$$\arg\min_{\gamma,\phi} D(q(\theta, \mathbf{t}|\gamma, \phi)||p(\theta, \mathbf{t}|\mathbf{w}, c, \alpha, \beta)) \quad . \tag{3.15}$$

If we take a partial derivative with respect to γ_n and ϕ_{in}, we can obtain the following iterative process to minimize the difference between real and approximated joint probability:

$$\gamma_n = \alpha_n + \sum_{i=1}^{N} \phi_{in} \tag{3.16}$$

$$\begin{aligned} \phi_{in} \propto \beta_{nm} \exp &\left(\Psi(\gamma_n) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right) \\ &\cdot \exp\left(\frac{1}{N}\eta_{cn} - (h^T {\phi_i}^{old})^{-1} h_n\right) \end{aligned} \tag{3.17}$$

where ϕ_i^{old} represents the value of ϕ_i at the previous iteration and h represents a simplified linear function of ϕ_i (see [80] for more details).

Recall that the inference procedure for the conventional LDA is

$$\gamma_n = \alpha_n + \sum_{i=1}^{N} \phi_{in} \tag{3.18}$$

$$\phi_{in} \propto \beta_{n\tau} \exp\left(\Psi(\gamma_n) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right) \quad . \tag{3.19}$$

Note that both approaches share the same update for Dirichlet parameter γ while update for ϕ_{in} is scaled according to the kernel function η and the previous ϕ_i. The main difference between LDA and sLDA lies in this update.

3.4.2.2 Classification

With the BBC Sound Effects Library (details are given in Section 3.5.1), we perform a 5-fold classification task with the onomatopoeic and semantic labels of audio clips.

Since the models using sLDA are trained with corresponding labels, we can classify test audio clips without extra consequent classifiers. Inferring a class category from sLDA-based models requires some approximation processes as well, such as variational approximation and Jensen's inequality [80]. The inference can be written as follows.

$$\hat{c} = \arg\max p(c|w) \tag{3.20}$$

where

$$\begin{aligned} p(c|w) &\approx \int p\left(c|t\right) q(t) dt \\ &= \int \frac{\exp({\eta_c}^T \bar{t})}{\sum_{c'=1}^{C} \exp({\eta_{c'}}^T \bar{t})} q(t) dt \\ &\geq \exp\left(E_q\left[{\eta_c}^T \bar{t}\right] - E_q\left[\log\left(\sum_{c'=1}^{C} \exp\left({\eta_{c'}}^T \bar{t}\right)\right)\right]\right) \end{aligned} \tag{3.21}$$

Since the second term is common for all classes, we can infer the class which maximizes the first term, i.e.,

$$\begin{aligned} \hat{c} &= \arg\max E_q \left[{\eta_c}^T \bar{t} \right] \\ &= \arg\max {\eta_c}^T \bar{\phi} \end{aligned} \tag{3.22}$$

where $\bar{\phi} = \frac{1}{N} \sum_{n=1}^{N} \phi_n$.

3.5 Experimental Setup

3.5.1 Database

We have collected 2,140 audio clips from the BBC Sound Effects Library [3], and labeled each file with onomatopoeia, semantic labels and short descriptions. The semantic labels and short descriptions are provided with the database. The semantic labels are given as one of predetermined 21 different categories. They include *transportation*, *military*, *ambiences*, *human*, and so on. The short descriptions consist of a set of words that represent the audio clip. Total number of words in the description is 2,820 word and average number of words in each description is 7.2 words after removing stop words and punctuation marks. For onomatopoeic words, we performed subjective annotation to label individual audio clips. We asked subjects to label the corresponding audio clip among

Table 3.1: Summary of BBC Sound Effect Library.

Number of sound clips	2,140
Number of semantic categories	21
Number of onomatopoeic words	22
Number of words for descriptions	2,820
Average number of words in a description	7.2
Average number of acoustic words in an audio clip	1,294

Table 3.2: Examples of BBC sound library along with its various descriptions.

Ex 1.	Filename	1-GOAT-MACHINE-MILKED-BB.wav
	Semantic category	MACHINERY/TOOLS
	Onomatopoeia	BLEATING
	Short description	animals: goats one goat milked by machine other goats bleating occasionally - interior - abrupt end
Ex 2.	Filename	1-ENGLISH-GOAT-BLEATING-BB.wav
	Semantic category	ANIMALS
	Onomatopoeia	BLEATING
	Short description	animals: goats one old english goat bleating - occasional wind noise - interior

22 onomatopoeia descriptions. See [73] for more details about collecting onomatopoeic words. The audio clips are originally recorded with 44.1kHz (stereo) sampling rate and down-sampled to 16kHz (mono) for acoustic feature extraction. The average length of the audio clips is about 13 seconds, which can generate approximately 1,300 acoustic words for an audio clip. A summary of the database is given in Table 3.1.

Table 3.2 shows examples of BBC sound library along with various labels: semantic labels, onomatopoeic words and short text descriptions. Both examples include the sound of a goat. While the subjective annotation of onomatopoeic words are the same, the semantic categories are different. These examples show the ambiguity of information even when they include the same audio contents.

Table 3.3 shows the distribution of onomatopoeic words and semantic labels for the database. For example there are 349 audio clips whose semantic labels are "animal'.' In that category of "animal", there exist various onomatopoeic words to represent the audio clips (e.g. 62 clips for "growl" and 60 clips "meow").

Table 3.3: Distribution of onomatopoeic words and semantic labels in the BBC sound library (22 onomatopoeias and 21 semantic labels).

Onomatopoeia \ Semantic labels	ANIMALS	HUMAN	TRANSPORTATION	OFFICE	MACHINERY	ELECTRONICS	PUBLIC	POLICE	HORROR	MILITARY	AMBIENCES	NATURE	HOUSEHOLD	SCI-FI	SPORTS	DOORS	OPEN	IMPACT	MUSIC	AUTOMOBILES	EXPLOSIONS	SUM
TWEET	53	1	0	0	0	0	0	0	1	0	50	3	0	0	0	0	0	0	0	0	0	108
SQUEAK	54	10	6	7	2	1	1	1	12	0	8	4	1	1	0	4	2	0	0	9	1	124
CLATTER	26	4	47	20	13	8	0	1	0	17	7	3	1	0	8	0	1	1	0	0	0	157
GABBLE	0	33	0	15	0	0	9	0	2	0	56	0	0	0	13	0	0	0	0	0	0	128
BURR	0	4	43	5	24	0	0	8	0	17	17	0	0	0	1	0	0	0	0	24	0	143
DONG	3	6	8	4	3	5	3	5	6	0	7	1	9	14	6	0	0	4	23	1	0	108
BUZZ	13	3	26	18	18	3	2	1	10	5	17	11	5	0	11	0	0	1	0	3	4	151
BLEAT	14	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17
GROWL	62	1	1	5	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	73
HUM	2	1	72	16	44	6	4	5	6	22	15	3	7	20	12	0	0	0	0	4	1	240
TAP	33	177	1	10	0	3	1	0	0	0	0	0	0	0	13	0	2	0	0	0	0	240
BEEP	0	3	5	5	0	17	0	46	2	0	1	0	0	30	0	0	0	0	0	1	1	111
WHOOSH	2	3	2	0	0	0	0	0	11	4	2	14	1	37	9	0	0	0	0	0	1	86
BANG	1	0	1	3	0	1	0	3	4	26	0	14	0	9	2	0	1	0	0	0	4	69
HONK	2	6	15	0	2	0	1	21	0	3	3	4	0	0	1	0	0	1	0	5	0	64
TICK	1	0	2	0	0	3	0	0	1	1	2	0	4	0	0	0	0	0	1	2	0	17
THUD	3	17	14	1	7	1	0	1	11	2	0	1	0	1	12	0	2	2	0	2	0	77
CRACKLE	1	8	0	0	1	0	0	0	2	3	0	14	1	1	0	0	0	0	0	0	3	34
CRUNCH	10	4	5	5	2	0	0	1	7	0	1	1	0	1	3	0	0	7	0	0	1	48
SPLASH	3	3	26	0	0	0	2	0	2	0	1	10	8	0	10	0	0	0	0	0	1	66
MEOW	60	0	0	6	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	67
CROW	6	1	0	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	12
SUM	349	285	274	126	117	48	23	93	79	101	187	83	37	116	102	4	8	16	24	51	17	2140

3.5.2 Experimental Scenarios

In this work, we design two different tasks; audio classification and text query classification. The task of classifying audio signals can be used in the annotation process and in the retrieval process using audio example queries. On the other hand, text query classification task can be used in the retrieval process using naive text queries. This experimental setting is reasonable because annotation process and retrieval process are independent except they share the storage as it is illustrated in Fig. 3.1. In both tasks, we allow an audio clip to have two tags, i.e. onomatopoeic and semantic labels assuming that these two tags describe the audio signals.

3.5.2.1 Audio Classification

Using the acoustic topic model, we can extract a single feature vector from an audio clip. Assuming that similar sounding situation share similar distribution over a set of topics, the feature vector, i.e. a posterior Dirichlet parameter of the corresponding audio clip, represents the distribution over latent topics in the corresponding audio clip. With the feature vectors, we utilize a Support Vector Machine (SVM) with polynomial kernels as a machine learning algorithm for this application. The performances are obtained by averaging five times of 10-fold cross validation tasks.

3.5.2.2 Text Query Classification

In this work, as we discussed in Section 3.1, we attempt to provide users with flexibility in their queries, so that people can use naive text as queries. In case of using naive text queries, however, it is somewhat problematic to deal with naive text queries due to uncertainty or ambiguity; users' naive descriptions to express their desired information are more explanatory rather than categorical. Furthermore, there are numerous types of polysemic words and synonymic words in text descriptions. This phenomenon may cause *out-of-vocabulary* problems where users inquire certain information which the system does not have a corresponding model yet.

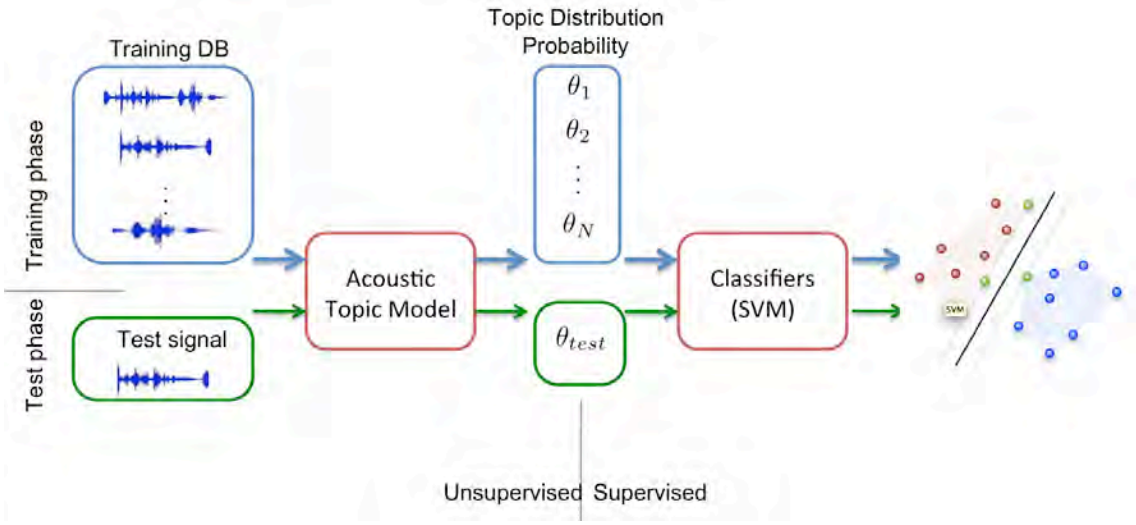


Figure 3.11: A simple diagram of audio classification task.

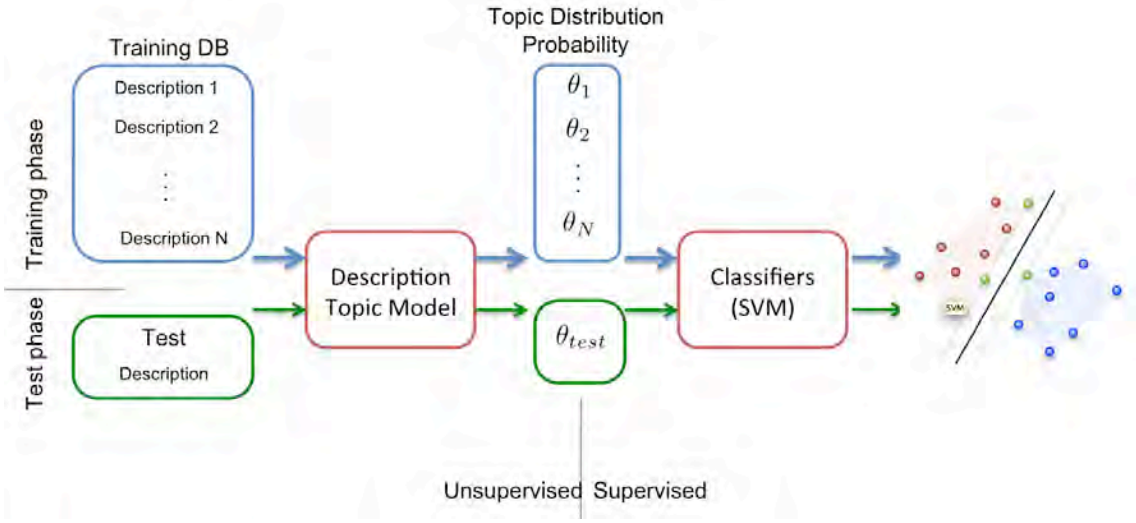


Figure 3.12: A simple diagram of text query classification task.

In this section, we introduce a new method that transforms naive text queries to pre-determinded audio descriptions on the intermediate audio description layer, i.e. onomatopoeias and semantic labels, so that people use their naive text descriptions to retrieve sounds they want. We utilize the semantic analysis methods that are used for analyzing acoustic context in the previous section: latent semantic indexing (LSI) and latent topic model. As it is described in the previous chapter, the latent topic model represents the

Topic 2		Topic 17		Topic 27		Topic 39	
word	prob.	word	prob.	word	prob.	word	prob.
animals	0.2122	**animals**	0.3333	cat	0.2905	tractor	0.0965
horse	0.1490	market	0.0583	**animal**	0.2134	tractors	0.0849
horses	0.0993	background	0.0485	individual	0.1208	**animals**	0.0772
track	0.0880	field	0.0453	miaows	0.0746	farm	0.0772
walk	0.0474	calling	0.0421	tom	0.0720	ford	0.0772

Figure 3.13: An example of words and their probability in topics. Topics that include word "animal".

probabilistic word distribution over topics [15, 35], while the LSI yields the association between words in a semantic space [9]. To this end, we follow a general way to build both topic models and latent semantic indexing. First, we build a dictionary that contains all the words in descriptions. Then, we make a *word-document co-occurance matrix* as it is done in an acoustic topic model. The word-document co-occurance matrix is fed in to both LSI and LDA algorithms to model the naive text descriptions.

For text query transformation tasks, we also apply similar settings. First, we extract a single feature vector from a description for an audio clip using latent topic model and latent semantic indexing method. Then, we utilize a SVM to classify the descriptions with the extracted feature vectors.

Fig. 3.13 shows an example of using the latent topic models to model text descriptions (we use the database introduced in Section 3.5). The figure shows the topics that has the word "animal(s)" in the 5 most probable words in the corresponding topics. Although each topic contains the word "animal(s)," surrounding words are different across the topics. It is notable that the surrounding words in individual topics are somewhat consistent in making a word cluster.

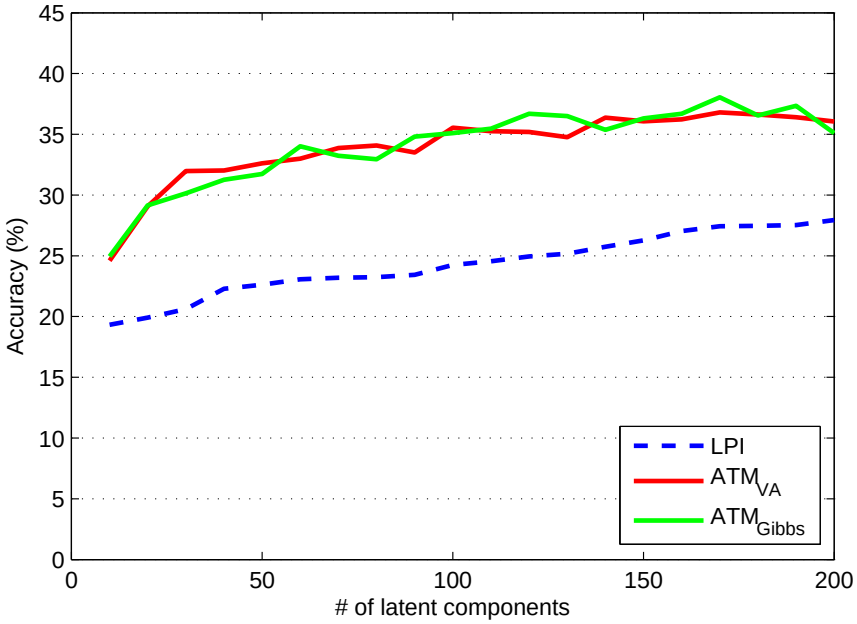
3.6 Results and discussion

3.6.1 Audio Classification

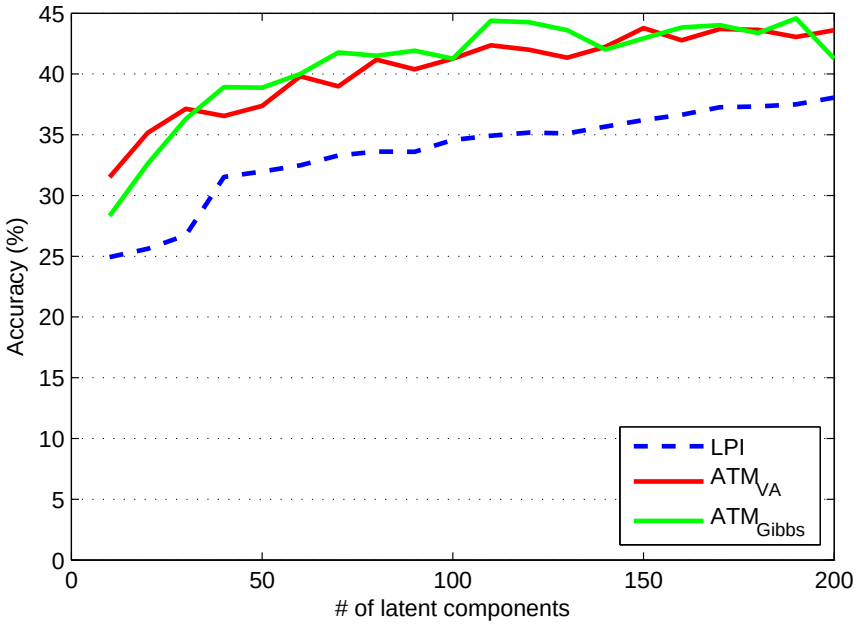
There are a couple of parameters that affect the modeling strategies in acoustic topic models such as size of acoustic dictionary and the number of latent components, etc. In this section, annotation accuracy (or classification accuracy in the sense that we consider only one tags for each information category) according to these parameters will be given to demonstrate their effects on modeling accuracy.

Fig. 3.14 shows the results of content-based audio description classification tasks using Latent Perceptual Indexing (LPI, dashed line) and Latent Dirichlet Allocation (LDA, solid lines) according to the number of latent components. The size of dictionary is set as 1,000 for this experiment. In LDA, we utilize both variational inference and Gibbs sampling methods for LDA approximation (red line for variational inference and green line for Gibbs sampling). Fig. 3.14 (a) and 3.14 (b) represent the results using onomatopoeic words and semantic labels, respectively. The number of latent components can be interpreted as the dimension of feature vector extracted from an audio clip. However, the interpretation differs in Latent Perceptual Indexing (LPI) and Latent Dirichlet Allocation (LDA). The number of latent components indicates a reduced rank after Singular Value Decomposition (SVD) in LPI, while it represents the number of hidden topics used in LDA.

The results clearly show that the proposed acoustic topic model outperforms the conventional SVD-based latent analysis method in both onomatopoeia labels and semantic labels regardless of approximation methods (no significant accuracy difference is found according to approximation methods). This significant improvement is evident regardless of the number of latent components. We argue that such results are due to utilizing LDA to analyze the hidden topics in audio clips. Note that the LPI analysis uses a deterministic approach based on SVD to map each word to the semantic space [9]. Although the semantic space is powerful to cluster the words that are highly related, the capability to predict the clusters from which the words are generated is somewhat limited in



(a) Onomatopoeic words



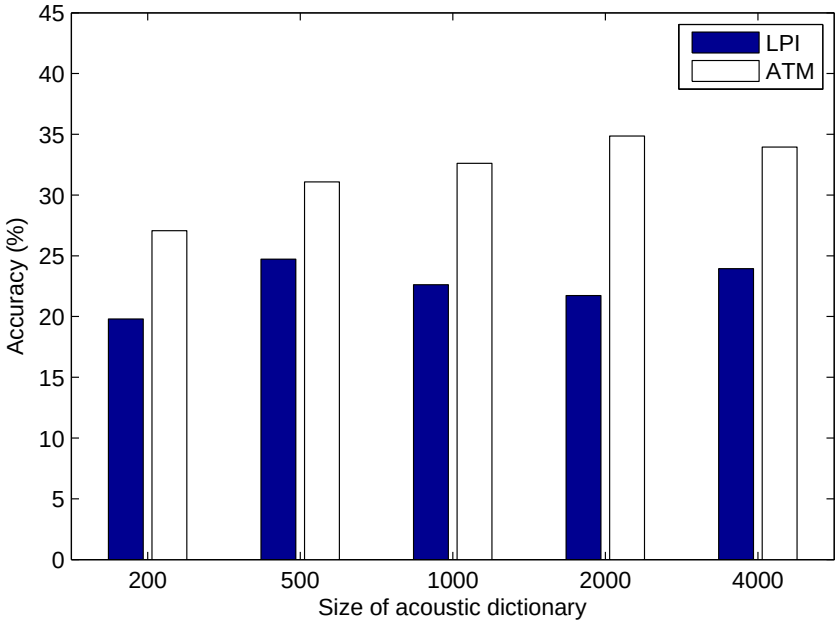
(b) Semantic labels

Figure 3.14: Classification results of acoustic words using Latent Perceptual Indexing (LPI, dashed line) and Latent Dirichlet Allocation (LDA, solid lines) according to the number of latent components: (a) onomatopoeic words and (b) semantic labels.

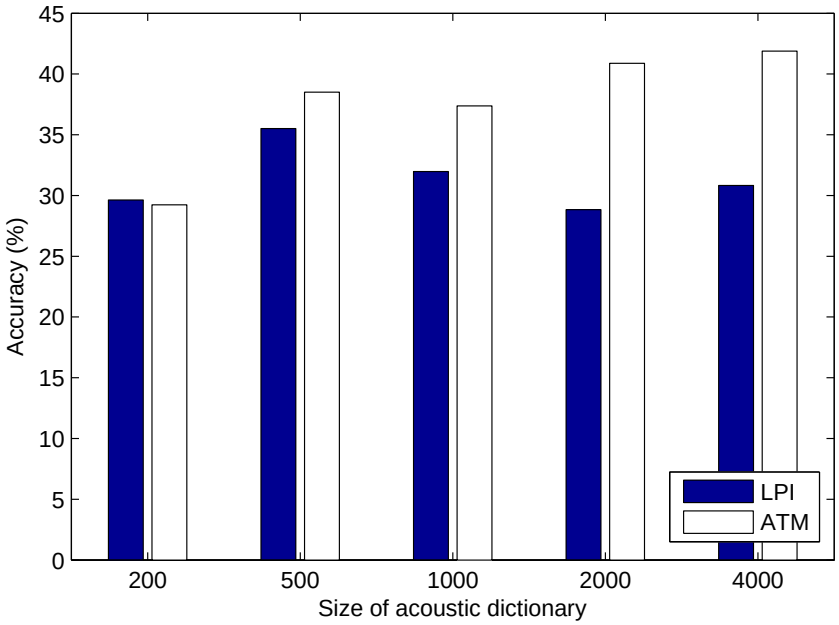
an euclidean space. With the proposed topic model, on the other hand, we are able to model the probabilities of acoustic topics that generate a specific acoustic word using a generative model.

In classifying onomatopoeia labels, the overall accuracy is lower than that of classifying semantic labels. This might be because the onomatopoeic words are for local sound contents rather than global sound contents; while the onomatopoeic words are local representations of sound clips, the topic model utilizes all the acoustic words from sound clips. Saliency detection algorithms [38] or foreground/background classification method [22] might be necessary to improve accuracy. It can be also observed that accuracy increases as the number of latent components increase. This is reasonable in the sense of feature dimension reduction; a larger feature vector usually captures more information. It should be noted, however, that there is a trade-off between accuracy and complexity. Increasing the feature vector size would increase computing power requirements exponentially as well.

We also perform the similar tasks with various sizes of dictionary. Fig. 3.15 shows the classification results according to the size of acoustic dictionary. Fig. 3.15 (a) and 3.15 (b) represent the results using onomatopoeic words and semantic labels, respectively. In this experiment, we set the number of latent components as 5% of size of acoustic dictionary for simplicity (e.g. 10 latent components for 200 acoustic words and 200 latent components for 4,000 acoustic words). As it is shown in the previous results, the results confirm that the proposed acoustic topic model outperforms the conventional SVD-based latent analysis method in both onomatopoeia labels and semantic labels. This significant improvement is evident regardless of the number of latent components (except the case of semantic labels with 200 acoustic words). However, the performances do not seem to be monotonically improving as the size of acoustic dictionary increases, especially concerning LPI cases. This is because these results are not directly comparable between the different sizes of acoustic dictionary; the numbers of latent components are also different if we apply different size of acoustic dictionary.



(a) Onomatopoeic words



(b) Semantic labels

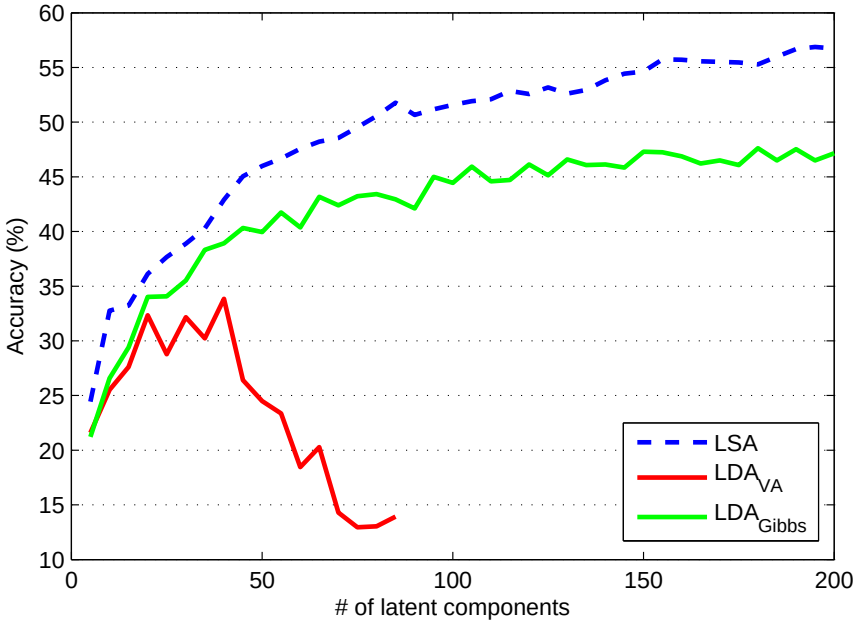
Figure 3.15: Classification results of acoustic words using Latent Perceptual Indexing (LPI) and Latent Dirichlet Allocation (LDA) according to the size of acoustic dictionary: (a) onomatopoeic words and (b) semantic labels.

3.6.2 Query Classification

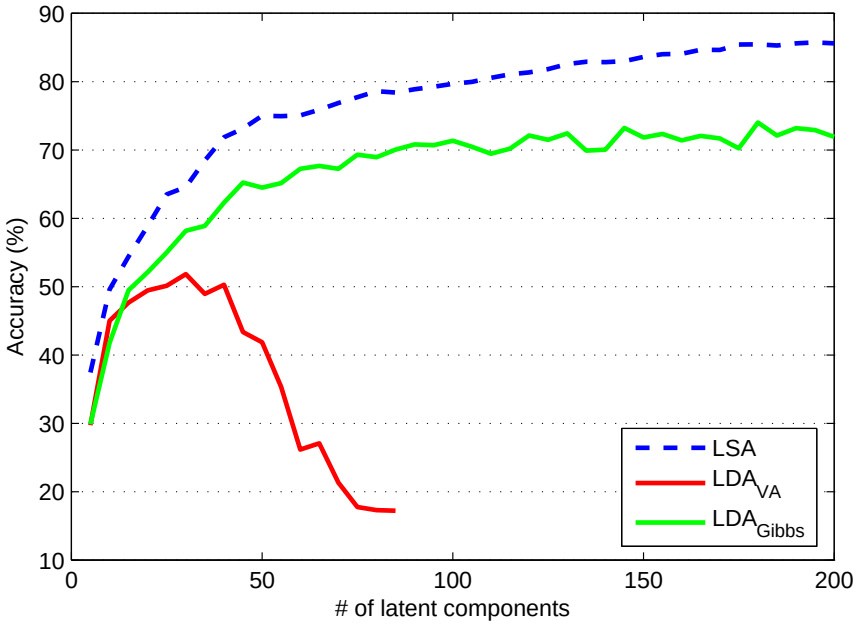
Unlike the acoustic topic model, we consider only one parameter that critically affects the modeling strategies in topic models, i.e. the number of latent components. This is because the size of vocabulary is already fixed with a given database. In this section, annotation accuracy (or, classication accuracy in the sense that we consider only one tags for each information category) according to the latent components will be given to demonstrate their effects on modeling accuracy.

Fig. 3.16 illustrates the results of classification tasks of text descriptions of audio signals using Latent Semantic Indexing (LSI, dashed line) and Latent Dirichlet Allocation (LDA, solid lines) according to the number of latent components; the size of dictionary is set as 1,000 for this experiment. In LDA, we utilize both variational inference and Gibbs sampling methods for LDA approximation (red line for variational inference and green line for Gibbs sampling). Fig. 3.16 (a) and 3.16 (b) represent the results using onomatopoeic words and semantic labels, respectively. The number of latent components can be interpreted as the dimension of feature vector extracted from a description of an audio clip.

The results clearly show that the SVD-based LSI method outperforms LDA method in classifying both onomatopoeic and semantic labels regardless of approximation methods for LDA. These significant differences are evident regardless of the number of latent components. We argue that this is because only few words are available for a description of an audio clip. As shown in Table 3.1, the average number of words in a description is 7.2. That might be too small to train the topic models which utilize a probabilistic approach while LSI utilizes a deterministic method. Furthermore, in the cases that use the variational approximation scheme, the topic model method cannot even generate any result as the number of latent components increases, while it somehow yield reasonable results with the Gibbs sampling scheme. We argue that this is because of the characteristics of



(a) Onomatopoeic words



(b) Semantic labels

Figure 3.16: Classification results of text descriptions using Latent Perceptual Indexing (LPI, dashed line) and Latent Dirichlet Allocation (LDA, solid lines) according to the number of latent components: (a) onomatopoeic words and (b) semantic labels.

approximation methods; while the Gibbs sampling method uses a simple iterative process of sampling and updates, the variational approximation method requires a sufficient number of training data to learn parameters.

In classifying onomatopoeia labels, just as the audio classification task, the overall accuracy is lower than that of classifying semantic labels. This shows the descriptions are highly related to semantic labels rather than onomatopoeic words. In LSA and LDA with Gibbs sampling method cases, It can be also observed that accuracy increases as the number of latent components increase. This is reasonable in the sense of feature dimension reduction; a larger feature vector usually captures more information. It should be noted, however, that there is a trade-off between accuracy and complexity. Increasing the feature vector size would increase computing power requirements exponentially as well.

3.6.3 Audio Classification with Modified Methodologies

3.6.3.1 N-gram approach

Fig. 3.17 shows the performance of audio classification tasks for both onomatopoeic labels and semantic labels. These two types of labels are chosen based on our previous work in [41] where the intermediate audio descriptive layer (iADL) was proposed to provide interoperability between the annotation and retrieval processes in an audio retrieval framework.

The 5-fold cross validation performance is shown as a function of number of latent components on the figure. As shown in the figure, the accuracy increases as the number of latent acoustic topics increases across various types of experimental settings. It is consistent with our previous work reported in [45] where we argued that this trend is reasonable in the sense of feature dimension reduction.

The direct comparison of performance with respect to the same number of latent acoustic topics is fair in the sense that the feature dimensions are the same which are fed into the classifier. In that sense, there is no significant performance differences by

using the bi-gram modeling approach (dash-dot lines) compared to the uni-gram approach (dashed lines). However, the direct comparison may not be fair if we consider the latent Dirichlet allocation algorithm as a dimension reduction process. For example, in the case that the number of latent acoustic topics is 100 (the feature dimension of audio clips is 100), the system only uses 0.25% of original feature vector dimension in bi-gram cases while it uses 50% in unigram cases. It is also related to the sparseness of data; for bi-gram modeling, 40,000 acoustic words are used to represent audio signals while 200 acoustic words are used for the uni-gram approach.

The solid lines show the performance using the hybrid method which makes super-vectors of feature vectors from uni-gram and bi-gram approaches. For simplicity, we use the feature vectors which are extracted using the same number of latent acoustic topics. Since we concatenate two feature vectors to make a super-vector, the dimension of a super-vector is greater than the one of original feature vectors (twice greater in this experiment). The results clearly shows the significant performance improvement by using the hybrid method which indicate that the uni-gram and bi-gram approaches represent complementary information.

We proposed the N-gram approach to model dynamic information within the text-like audio modeling scenario for information retrieval applications. Specifically, we have used the bi-gram model to consider adjacent acoustic words and built a new acoustic word dictionary for the bi-grams. Experimental results showed that the proposed N-gram approach brought significant improvements in the performance by providing complementary local dynamic information.

3.6.3.2 Supervised LDA

Fig. 3.18 illustrates the classification results of audio clips using latent acoustic topics with both LDA and sLDA (Table 3.4 shows the performance in numbers along with relative improvements).

Table 3.4: Classification results of audio clips using latent acoustic topics with LDA and supervised LDA.

Accuracy (%)	LDA	sLDA	Relative Improvement
Semantic labels	38.8	**43.4**	11.9
Onomatopoeic labels	32.1	**35.1**	9.3

As shown in the figure, the accuracy rates using the supervised acoustic topic model are higher than the ones using conventional acoustic topic model for both onomatopoeic and semantic labels (11.9% and 9.3 % relative improvements for semantic labels and onomatopoeic labels, respectively). This significant improvement is from using sLDA instead of LDA; sLDA learns its parameters according to categories of training data, while LDA does not consider the categories. Instead, LDA uses a consequence classifier (SVM, in this work) for classification tasks so that the acoustic topic modeling process is independent of descriptive categories.

The conventional acoustic model using LDA, however, has some advantages over the supervised acoustic model using sLDA, besides the fact that sLDA requires significant greater computational power than LDA does. Since LDA learns the latent variables in an unsupervised manner without considering the labels, one can apply various types of categories and classifiers without re-learning the parameters. For example, in this work, we trained two separate supervised acoustic topic models for two different descriptive categories, i.e., onomatopoeic and semantic labels, while we could train only one acoustic topic model for both descriptive categories and use consequent SVM for classification tasks.

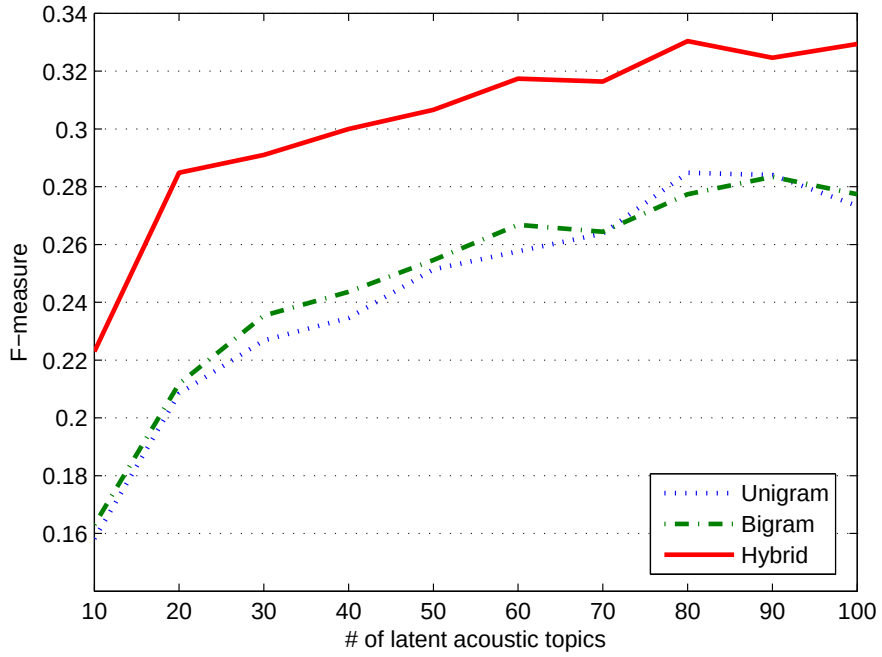
In this work, we investigated the effects of supervised acoustic topic models over the conventional acoustic topic model within the unstructured audio information retrieval framework. While the conventional acoustic topic model utilizes a latent Dirichlet allocation (LDA) method, we adopted a modified version, supervised LDA (sLDA), which

considers categorical labels during learning latent variables. The experimental results with BBC Sound Effects Library showed that the supervised acoustic topic model using sLDA outperforms the conventional acoustic topic model with LDA; it indicates that the supervised acoustic model brings benefits in terms of classification accuracy by learning parameters considering corresponding descriptive categories of audio clips rather than unsupervised learning.

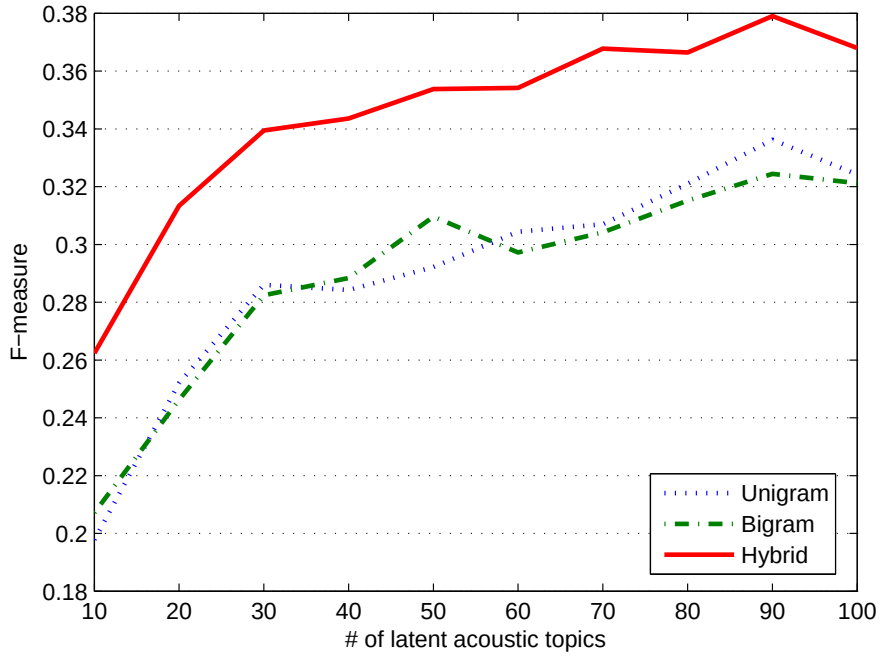
3.7 Chapter Summary

To model the context hidden in the audio signals, we proposed a latent acoustic topic model based on Latent Dirichlet Allocation (LDA), which learns hidden acoustic topics in a given audio signal in an unsupervised way. We adopted the variational inference method and the Gibbs sampling method to train the topic model and used the posterior Dirichlet parameters as a representative feature vector for an audio clip. Due to the rich acoustic information present in audio clips, the embedded information can be categorized based on the *intermediate audio description layer* which includes semantic and onomatopoeic categories; the semantic and onomatopoeic categories represent the cognition of the acoustic realization of a scene and its perceptual experience, respectively. The results of classifying theses two descriptions showed that the proposed acoustic topic model significantly outperforms the conventional SVD-based latent structure analysis method.

We also proposed the text query transformation strategy to provide the flexibility in input queries. We utilized both the latent semantic indexing (LSI) method and the latent topic model to this end; it transformed naive input queries onto the intermediate audio description layer so that the annotation process and the retrieval process are interoperable. The results suggested that LSI yields better performance than the latent topic model because of the number of words in a description.



(a) Onomatopoeic words



(b) Semantic labels

Figure 3.17: Classification results of audio clips using unigram and bigram acoustic words in the acoustic topic model framework according to the number of latent acoustic topics: (a) onomatopoeic words and (b) semantic labels.

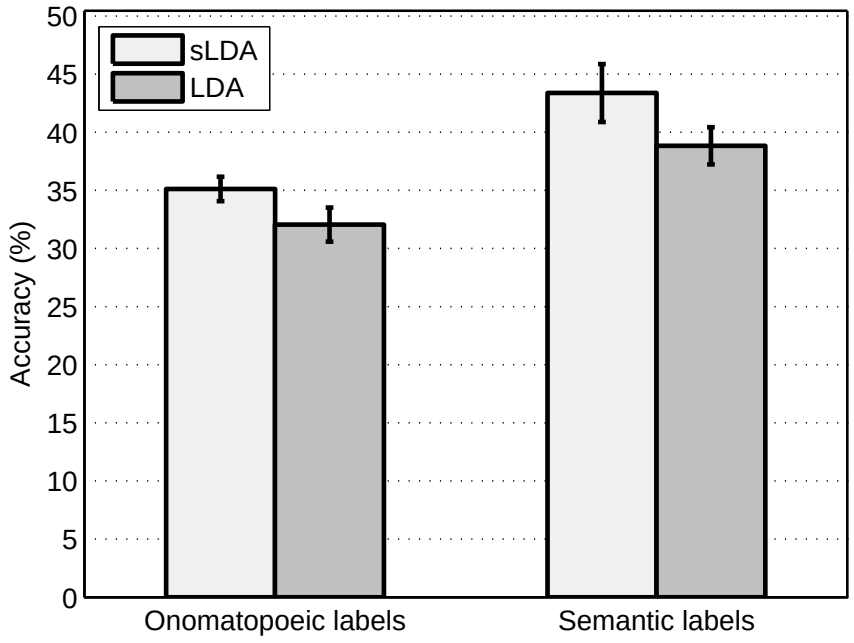


Figure 3.18: Classification results of audio clips using latent acoustic topics with LDA and supervised LDA.

Chapter 4

Concluding Remarks

4.1 Conclusion

In this dissertation, we have focused on extracting context information in audio signals toward building context-based audio information retrieval systems. We have categorized the audio signals according to how they are generated, and focus on two major categories of audio signals: well-structured audio signals such as music and general audio signals which include unstructured audio signals.

In dealing with well-structured audio signals, i.e. Chapter 2, we introduced a new music information retrieval method using the proposed context-based music fingerprint. The proposed music fingerprint models musically meaningful aspects of a music audio signal, such as harmonic structures and their temporal dynamic information, in a compact representation. It provides an efficient way of extracting information useful for various music information retrieval applications. Through experimental evaluation with MIR frameworks, such as opus identification, composer identification, and semantic annotation, we discussed the performance of the proposed context-based music fingerprint method. The results suggest that the proposed music fingerprint is efficient in terms of complexity, both processing and storage requirements, while yielding performance accuracy that is competitive against state-of-the-art systems. New algorithms which extract dynamic information were also proposed, and it is shown that they can be incorporated to provide complementary information.

In dealing with general audio signals which include unstructured audio signals, i.e. Chapter 3, we proposed an acoustic topic model based on Latent Dirichlet Allocation (LDA) which learns hidden acoustic topics in a given audio signal in an unsupervised way. We adopted the variational inference method and the Gibbs sampling method to train the topic model, and use the posterior Dirichlet parameters as a representative feature vector for an audio clip. Due to the rich acoustic information present in audio clips, they can be categorized based on the *intermediate audio description layer* which includes semantic and onomatopoeic categories; they represent the cognition of the acoustic realization of a scene and its perceptual experience, respectively. The results of classifying theses two descriptions showed that the proposed acoustic topic model significantly outperforms the conventional SVD-based latent structure analysis method. We also proposed the text query transformation strategy to provide the flexibility in input queries. We utilized both the latent semantic indexing (LSI) method and the latent topic model to this end; it transforms naive input queries onto the intermediate audio description layer so that the annotation process and the retrieval process are interoperable. The results show that LSI yields better performance than the latent topic model because of the number of words in a description.

4.2 Possible Future Work

Possible future work can be performed in two major directions: algorithmic methodologies and application domains. In the direction of developing new methodologies, we are going to extend our context-based approaches proposed in Chapter 2 and Chapter 3.

Specifically, in the direction of exploring new applications, we are interested in extending our methodologies to general multimedia databases which include music, unstructured audio, and video. To this end, we need to devise algorithms and fusion strategies to deal with multimodal data streams. Zhang *et. al.* provided a good benchmarking framework in proposing algorithms to analyze video blogs [85]. We expect that our

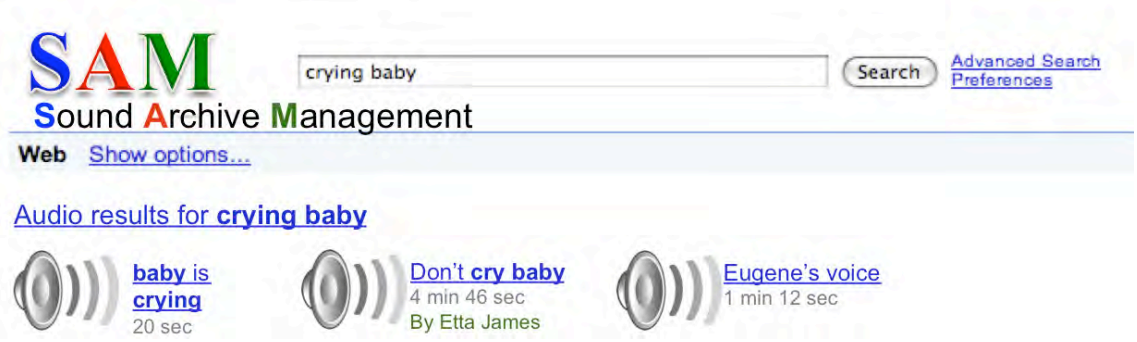


Figure 4.1: A blueprint of a sound archive management (SAM) system.

proposed context-based audio information retrieval algorithms can contribute to the multimodal approach to the multimedia databases. Secondly, we are planning to explore the community-contributed multimedia data in Web 2.0 applications. One of the examples of community-contributed multimedia data is the social tags which include sentiments and emotional responses to multimedia data [12, 71, 51]. Ratings, recommend/unrecommend, etc. are also interesting aspects of the community-contributed data.

The research in this dissertation will eventually enable to build an end-to-end sound archive management system whose blueprint is illustrated in Fig. 4.1. It will take queries from users and yield a list of relevant audio clips in either WWW framework or local computers.

References

[1] Classical archives. [Online]. Available: http://www.classicalarchives.com/

[2] Timidity++. [Online]. Available: http://timidity.sourceforge.net/

[3] The BBC sound effects library - original series. [Online]. Available: http://www.sound-ideas.com

[4] MIREX 2007. [Online]. Available: http://www.music-ir.org/mirex2007/

[5] MIREX 2008. [Online]. Available: http://www.music-ir.org/mirex2008/

[6] J. Bai, J. Y. Nie, and G. Cao, "Context-dependent term relations for information retrieval," in *Conference of Emprical Methods in Natural Language Processing*, 2006, pp. 551–559.

[7] J. Bai, J. Y. Nie, G. Cao, and H. Bouchard, "Using query contexts in information retrieval," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Reseach and Development in Information Retrieval*, 2007, pp. 15–22.

[8] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pitctures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[9] J. R. Bellegarda, "Latent semantic mapping," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 70–80, 2005.

[10] J. Bello, "Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps, and beats," in *International Symposium on Music Information Retreival*, Vienna, Austria, September 2007, pp. 239–244.

[11] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, ser. Signals and Communication Technology. Springer, 2005.

[12] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere, "Autotagger: A model for predicting social tags from acoustic features on large music database," *Journal of New Music Research*, vol. 37, no. 2, pp. 115–135, 2008.

[13] E. Bigand and R. Parncut, "Perceiving musical tension in long chord sequence," *Psychological Research*, vol. 62, no. 4, pp. 237–254, 1999.

[14] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *NIPS*, 2007.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003.

[16] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 127–134.

[17] D. M. Blei and J. D. Lafferty, "A correlated topic model of *science*," *The annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.

[18] D. A. Bray, "Information Pollution, Knowledge Overload, Limited Attention Spans, and Our Responsibilities as IS Professionals," in *Global Information Technology Management Association (GITMA) World Conference*, 2008.

[19] G. Chechik, E. Ie, M. Rehn, S. Bengio, and R. F. Lyon, "Large-scale content-based audio retrieval from text queries," in *ACM International Conference on Multimedia Information Retrieval (MIR)*, 2008.

[20] E. Chew, "Modeling tonality: Applications to music cognition," in *CogSci2001 23rd Annual Meeting of the Cognitive Science Society, Edinburg, Scotland*, 2001.

[21] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with joint time- and frequency-domain audio features," *IEEE Transactions on Speech, Audio and Language Processing*, vol. 17(6), pp. 1142–1158, 2009.

[22] ——, "A semi-supervised learning approach to online audio background detection," in *IEEE International conference on Acoustic, Speech and Signal Processing*, 2009.

[23] C.-H. Chuan and E. Chew, "Polyphonic audio key-finding using the spiral array ceg algorithm," in *International Conference on Multimedia and Expo*, 2005, pp. 21– 24.

[24] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley and Sons, New York, 1991.

[25] J. S. Downie, "The music information retrieval evaluation exchange (MIREX)," *D-Lib Magazine*, vol. 12, no. 12, 2006. [Online]. Available: http://www.dlib.org/dlib/december06/downie/12downie.html

[26] D. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *ACM workshop on Continuous Archiving and Recording of Personal Experiences CARPE-04*, 2004.

[27] D. Ellis. Music beat tracking and cover song identification. [Online]. Available: http://www.ee.columbia.edu/∼dpwe/

[28] ——, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[29] D. Ellis and C. Cotton, "The 2007 LabROSA cover song detection system," in *International Symposium on Music Information Retrieval*, 2007.

[30] D. Ellis and G. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. 1429–1432.

[31] J. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, pp. 2–10, 1999.

[32] G. Friedland, L. Gottlieb, and A. Janin, "Joke-o-mat: Browsing sitcoms punchline by punchline," in *Proceedings of ACM Multimedia*, 2009.

[33] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[34] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *National Academy of Sciences*, 2003, pp. 5228–5235.

[35] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," *Psychological Review*, vol. 114, no. 2, pp. 211–244, 2007.

[36] J. Jensen, D. Ellis, M. Chistensen, and S. Jensen, "A chroma-based tempo-insensitive distance measure for cover song identification," in *International Symposium on Music Information Retrieval*, 2007.

[37] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating query substitutions," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM, 2006, pp. 387–396.

[38] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *InterSpeech ICSLP*, 2007.

[39] S. Kim, P. Georgiou, and S. Narayanan, "A robust harmony structure modeling scheme for classical music opus identification," in *IEEE International conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2009, pp. 1961–1964.

[40] S. Kim, E. Unal, and S. Narayanan, "Music fingerprint extraction for classical music cover song identification," in *International Conference of Multimedia and Expo*, 2008, pp. 1261–1264.

[41] S. Kim, P. Georgiou, S. Narayanan, and S. Sundaram, "Using naive text queries for robust audio information retrieval system," in *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2010.

[42] S. Kim, P. Georgiou, S. Sundaram, and S. Narayanan, "Acoustic stopwords for unstructured audio information retrieval," in *European Signal Processing Conference (EUSIPCO)*, 2010.

[43] S. Kim, P. G. Georgiou, and S. Narayanan, "Supervised acoustic topic model for unstructured audio information retrieval," in *Asia Pacific Signal and Information Processing Association (APSIPA) annual summit and conference*, 2010.

[44] S. Kim and S. Narayanan, "Content-based acoustic metadata for music information retrieval," *IEEE transactions on Audio, Speech, and Language Processing*, in preparation.

[45] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic models for audio information retrieval," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2009.

[46] S. Kim, S. Sundaram, P. Georgiou, and S. Narayanan, "An n-gram model for unstructured audio signals toward information retrieval," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2010.

[47] Y. Kim and D. Perelstein, "MIREX 2007: Audio cover song detection using chroma featres and a hidden markov model," in *International Symposium on Music Information Retrieval*, 2007.

[48] K. Lee, "Identifying cover songs from audio using harmonic representation," in *International Symposium on Music Information Retrieval*, 2006.

[49] K. Lee and M. Slaney, "Acoustic chord transcription and key estimation from audio using key-dependent hmms trained on synthesized audio," *Special Issue of the IEEE transaction on Audio, Speech and Language Processing on Music Information Retrieval*, vol. 16, no. 2, pp. 291–301, 2008.

[50] K. Lee and D. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[51] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *Multimedia, IEEE Transactions on*, vol. 11, no. 3, pp. 383–395, 2009. [Online]. Available: http://dx.doi.org/10.1109/TMM.2009.2012913

[52] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 1, pp. 1–19, 2006.

[53] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Transactions on Speech and Language Processing*, 2006.

[54] R. E. Madsen and D. Kuchak, "Modeling word burstiness using the dirichlet distribution," in *International Conference on Machine Learning*, 2005.

[55] M. Mandel and D. Ellis, "Song-level features and SVM for music classification," in *International Symposium on Music Information Retrieval*, London, September 2006, pp. 594–599.

[56] A. Mardirossian and E. Chew, "Key distributions as musical fingerprints for similarity assessment," in *Proceedings of the Seventh IEEE International Symposium on Multimedia*, 2005.

[57] ——, "Music summarization via key distributions: Analyses of similarity assessment across variations," in *International Conference on Music Information Retrieval*, 2006.

[58] K. Melih, R. Gonzalez, and P. Ogunbona, "An audio representation for content based retrieval," in *IEEE TENCON - Speech and Image Technonologies for Computing and Telelcommunications*, 1997.

[59] T. Minka and J. Lafferty, "Estimating a dirichlet distribution," M.I.T., Tech. Rep., 2000.

[60] J. M. Noyes and P. J. Thomas, "Information overload: An overview," in *IEE Colloquium on Information Overload*, 1995.

[61] R. Parncut, *Harmony: A psychoacoustical approach.* Berlin: Springer-Verlag, 1989.

[62] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* Prentice Hall, 2001.

[63] L. R. Rabiner and B. Juang, *Fundamentals of speech recognition.* Prentice Hall, 1993.

[64] J. Reed and C.-H. Lee, "On the importance of modeling temporal information in music tag annotation," in *IEEE International Conference of Acoustics, Speech, and Signal Processing*, 2009.

[65] J. Serrà and E. Gómez, "A cover song identification system based on sequences of tonal descriptors," in *International Symposium on Music Information Retrieval*, 2007.

[66] J. Serrà, E. Gómez, and P. Herrera, "Improving binary similarity and local alignment for cover song detection," in *International Symposium on Music Information Retrieval*, 2008.

[67] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1138–1151, 2008.

[68] R. Shepard, "Circularity in judgments of relative pitch," *Journal of the Acoustic Society of America*, vol. 36, no. 12, pp. 2346–2353, 1964.

[69] M. Slaney, "Semantic-audio retrieval," in *IEEE International Conference of Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp. 4108–4111.

[70] M. Steyvers and T. Griffiths, *Probabilistic Topic Models.* Laurence Erlbaum, 2006.

[71] F. M. Suchanek, M. Vojnovic, and D. Gunawardena, "Social tags: Meaning and suggestions," in *17th ACM conference on Information and knowledge management*, 2009, pp. 223–232.

[72] S. Sundaram and S. Narayanan, "Audio retrieval by latent perceptual indexing," in *IEEE International Conference of Acoustics, Speech, and Signal Processing*, 2008.

[73] ——, "Classification of sound clips by two schemes: using onomatopoeia and semantic labels," in *IEEE International Conference of Multimedia and Expo*, 2008.

[74] ——, "A divide-and-conquer approach to latent perceptual indexing of audio for large web 2.0 application," in *IEEE International Conference of Multimedia and Expo*, 2009.

[75] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 467–476, Feb. 2008.

[76] E. Unal, E. Chew, P. Georgiou, and S. Narayanan, "Challenging uncertainty in query-by-humming systems: A fingerprinting approach," *Special Issue of the IEEE transaction on Audio, Speech and Language Processing on Music Information Retrieval(MIR)*, vol. 16, no. 2, pp. 359–371, 2008.

[77] E. Unal and S. Narayanan, "Statistical modeling and retrieval of polyphonic music," in *International workshop on Multimedia Signal Processing*, 2007.

[78] S. Wake and T. Asahi, "Sound retrieval with intuitive verbal expressions," in *International Conference on Auditory Display*, 1998.

[79] C. Wang, D. M. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[80] C. Wang, D. M. Blei, and L. Fie-Fei, "Simulateneous image classification and annotation," in *CVPR*, 2009.

[81] D. Warren, S. Uppenkamp, R. D. Patterson, and T. D. Griffiths, "Separating pitch chroma and pitch height in the human brain," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 17, pp. 10 038–10 042, 2003.

[82] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, "Segmentation, indexing, and retrieval for environmental and natural sounds," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[83] O. Yakhnenko and V. Honavar, "Multi-modal hierarchical dirichlet process model for predicting image annotation and image-object label correpsondence," in *the SIAM Conference on Data Mining*, 2009.

[84] Z. Zeng, H. Li, W. Liang, and S. Zhang, "A hierarchical generative model for generic audio document categorization," in *IEEE International Conference of Acoustics, Speech, and Signal Processing*, 2010.

[85] X. Zhang, C. Xu, J. Cheng, H. Lu, and S. Ma, "Effective annotation and search for video blogs with integration of context and content analysis," *Multimedia, IEEE Transactions on*, vol. 11, no. 2, pp. 272–285, Feb. 2009.

Appendix A

Variational approximation method for Latent Dirichlet Allocation: Inference

One of the most important processes in LDA might be computing the joint probability of θ and $\mathbf{t}$ with given $\mathbf{w}$. It should be estimated as

$$p(\theta, \mathbf{t}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \ . \tag{A.1}$$

As we describe earlier, it is computationally impossible to estimate the denominator of the above equation. In this work, we utilize the variational inference method introduced in [15]. Blei *et al* have shown that this approximation works reasonably well in various applications, such as document modeling and document classification.

The rationale behind the method is to minimize distance between the real distribution and the simplified distribution using Jensen's inequality [24]. The simplified version has γ and ϕ which are the Dirichlet parameter that determines θ and the multinomial parameter that generates topics respectively, as depicted in Fig. 3.5. The joint probability of θ and $\mathbf{t}$ in (A.1) can be simplified as

$$q(\theta, \mathbf{t}|\gamma, \phi) = q(\theta|\gamma) \prod_{i=1}^{N} q(t_i|\phi_i) \tag{A.2}$$

and tries to minimize the difference between real and approximated joint probabilities using Kullback-Leibler (KL) divergence, i.e.

$$\arg\min_{\gamma,\phi} D(q(\theta, \mathbf{t}|\gamma, \phi)||p(\theta, \mathbf{t}|\mathbf{w}, \alpha, \beta)) \ . \tag{A.3}$$

We can begin with the log-likelihood of the marginal distribution. We can impose a lower bound to the log-likelihood using Jensen's inequality [24]. Jensen's inequality represent if a function f is a convex function and X is a random variable, then

$$Ef(X) \geq f(EX) \tag{A.4}$$

where E denotes an expectation operator. Since the log-likehood include a logarithm function which is negative convex, i.e. concave, we can utilize the above property to impose a lower bound to the log-likelihood of the marginal distribution.

$$\begin{aligned}
\log p(\mathbf{w}|\alpha, \beta) &= \log \int \sum_{\mathbf{t}} p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta) \ d\theta \\
&= \log \int \sum_{\mathbf{t}} q(\theta, \mathbf{t}|\gamma, \phi) \frac{p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{t}|\gamma, \phi)} \ d\theta \\
&\geq \int \sum_{\mathbf{t}} q(\theta, \mathbf{t}|\gamma, \phi) \log \frac{p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{t}|\gamma, \phi)} \ d\theta \\
&= \int \sum_{\mathbf{t}} q(\theta, \mathbf{t}|\gamma, \phi) \log p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta) - \int \sum_{\mathbf{t}} q(\theta, \mathbf{t}|\gamma, \phi) \log q(\theta, \mathbf{t}|\gamma, \phi) \\
&= E_q[\log p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{t}|\gamma, \phi)]
\end{aligned} \tag{A.5}$$

The difference between the left-hand and right-hand sides can be minimized by maximizing the lower bound. Further simplification by letting the right-hand side $L(\gamma, \phi|\alpha, \beta)$ can be done as follows.

$$\begin{aligned}
L(\gamma, \phi|\alpha, \beta) &= E_q[\log p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{t}|\gamma, \phi)] \\
&= E_q[\log p(\theta|\alpha)] + E_q[\log p(\mathbf{t}|\theta)] + E_q[\log p(\mathbf{w}|\mathbf{t}, \beta)] \\
&\quad - E_q[\log q(\theta|\gamma)] - E_q[\log q(\mathbf{t}|\phi)]
\end{aligned} \tag{A.6}$$

$$
\begin{aligned}
L(\gamma, \phi|\alpha, \beta) &= E_q[\log p(\theta|\alpha)] + E_q[\log p(\mathbf{t}|\theta)] + E_q[\log p(\mathbf{w}|\mathbf{t}, \beta)] \\
&\quad - E_q[\log q(\theta|\gamma)] - E_q[\log q(\mathbf{t}|\phi)] \\
&= E_q\left[\log\left(\frac{\Gamma\left(\sum_{n=1}^{k}\alpha_n\right)}{\prod_{n=1}^{k}\Gamma(\alpha_n)}\prod_{n=1}^{k}\theta_n^{\alpha_n-1}\right)\right] + E_q\left[\log\left(\prod_{i=1}^{N}\prod_{n=1}^{k}(\theta_n)^{t_{in}}\right)\right] \\
&\quad + E_q\left[\log\left(\prod_{i=1}^{N}\prod_{n=1}^{k}\prod_{m=1}^{V}(\beta_{nm})^{w_{im}}\right)\right] - E_q\left[\log\left(\frac{\Gamma\left(\sum_{n=1}^{k}\gamma_n\right)}{\prod_{n=1}^{k}\Gamma(\gamma_n)}\prod_{n=1}^{k}\theta_n^{\gamma_n-1}\right)\right] \\
&\quad - E_q\left[\log\left(\prod_{i=1}^{N}\prod_{n=1}^{k}(\phi_{in})^{t_{in}}\right)\right] \\
&= \log\Gamma\left(\sum_{n=1}^{k}\alpha_n\right) - \sum_{n=1}^{k}\log\Gamma(\alpha_n) + \sum_{n=1}^{k}(\alpha_n - 1)E_q[\log\theta_n] \\
&\quad + \sum_{i=1}^{N}\sum_{n=1}^{k}E_q\left[t_{in}\log\theta_n\right] \\
&\quad + \sum_{i=1}^{N}\sum_{n=1}^{k}\sum_{m=1}^{V}E_q[w_{im}\log\beta_{nm}] \\
&\quad - \left\{\log\Gamma\left(\sum_{n=1}^{k}\gamma_n\right) - \sum_{n=1}^{k}\log\Gamma(\gamma_n) + \sum_{n=1}^{k}(\gamma_n - 1)E_q[\log\theta_n]\right\} \\
&\quad - \sum_{i=1}^{N}\sum_{n=1}^{k}E_q\left[t_{in}\log\phi_{in}\right]
\end{aligned}
\tag{A.7}
$$

Some parts can be even simplified as follow; the expectation of the logarithmic Dirichlet random variable θ can be written as

$$
\begin{aligned}
E_q[\log\theta_n] &= \Psi(\gamma_n) - \Psi(\gamma_0) \\
&= \Psi(\gamma_n) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)
\end{aligned}
\tag{A.8}
$$

where Ψ represents the first derivative of the log Gamma function, i.e.

$$
\Psi(\gamma) = \frac{d}{d\gamma}\log\Gamma(\gamma) \quad , \tag{A.9}
$$

$$E_q[t_{in} \log \theta_n] = E_q[t_{in}] \cdot E_q[\log \theta_n] \tag{A.10}$$

because t_{in} and θ_n are independent by the assumptions.

$$\begin{aligned} E_q[t_{in} \log \theta_n] &= E_q[t_{in}] \cdot E_q[\log \theta_n] \\ &= \int \sum_t q(\theta, \mathbf{t}|\gamma, \phi) t_{in} \; d\theta \cdot E_q[\log \theta_n] \\ &= \int \sum_t q(\theta|\gamma) \prod_{a=1}^{N} q(t_a|\phi_a) t_{in} \; d\theta \cdot E_q[\log \theta_n] \\ &= \int q(\theta|\gamma) \; d\theta \sum_t \prod_{a=1}^{N} \prod_{b=1}^{k} (\phi_{ab})^{t_{ab}} t_{in} \cdot E_q[\log \theta_n] \\ &= \phi_{in} \cdot E_q[\log \theta_n] \quad , \end{aligned} \tag{A.11}$$

and

$$\begin{aligned} E_q[w_{im} \log \beta_{nm}] &= \int \sum_t q(\theta, \mathbf{t}|\gamma, \phi) w_{im} \log \beta_{nm} \; d\theta \\ &= \int \sum_t q(\theta|\gamma) \prod_{a=1}^{N} \prod_{b=1}^{k} (\phi_{ab})^{t_{ab}} w_{im} \log \beta_{nm} \; d\theta \\ &= \phi_{in} w_{im} \log \beta_{nm} \end{aligned} \tag{A.12}$$

Finally, the approximated log-likelihood of the marginal probability can be written as

$$\begin{aligned} L(\gamma, \phi|\alpha, \beta) = & \log \Gamma \left(\sum_{n=1}^{k} \alpha_n \right) - \sum_{n=1}^{k} \log \Gamma(\alpha_n) + \sum_{n=1}^{k} (\alpha_n - 1) \left(\Psi(\gamma_n) - \Psi \left(\sum_{j=1}^{k} \gamma_j \right) \right) \\ & + \sum_{i=1}^{N} \sum_{n=1}^{k} \phi_{in} \left(\Psi(\gamma_n) - \Psi \left(\sum_{j=1}^{k} \gamma_j \right) \right) \\ & + \sum_{i=1}^{N} \sum_{n=1}^{k} \sum_{m=1}^{V} \phi_{in} w_{im} \log \beta_{nm} \\ & - \left\{ \log \Gamma \left(\sum_{n=1}^{k} \gamma_n \right) - \sum_{n=1}^{k} \log \Gamma(\gamma_n) + \sum_{n=1}^{k} (\gamma_n - 1) \left(\Psi(\gamma_n) - \Psi \left(\sum_{j=1}^{k} \gamma_j \right) \right) \right\} \\ & - \sum_{i=1}^{N} \sum_{n=1}^{k} \phi_{in} \log \phi_{in} \quad . \end{aligned} \tag{A.13}$$

To maximize the approximated log-likehood, we first use the Lagrange multiplier with respect to ϕ_{in} whose constraint is $\sum_{n=1}^{k} \phi_{in} = 1$, i.e.

$$L_{[\phi]}(\gamma, \phi|\alpha, \beta) = L(\gamma, \phi|\alpha, \beta) + \sum_{i=1}^{N} \lambda_i \left(\sum_{n=1}^{k} \phi_{in} - 1 \right) \ . \tag{A.14}$$

When we take a derivative with respect to ϕ_{in}, we can obtain

$$\frac{\partial}{\partial \phi_{in}} L_{[\phi]}(\gamma, \phi|\alpha, \beta) = \Psi(\gamma_n) - \Psi \left(\sum_{j=1}^{k} \gamma_j \right) + \sum_{m=1}^{V} w_{im} \log \beta_{nm} - \log \phi_{in} - 1 + \lambda_i \tag{A.15}$$

where $w_{i\tau} = 1$ and the other elements in w_i are zeros. Hence, it can be even simplified as

$$\frac{\partial}{\partial \phi_{in}} L_{[\phi]}(\gamma, \phi|\alpha, \beta) = \Psi(\gamma_n) - \Psi \left(\sum_{j=1}^{k} \gamma_j \right) + \log \beta_{n\tau} - \log \phi_{in} - 1 + \lambda_i \ . \tag{A.16}$$

To satisfy the derivative to be zero, ϕ_{in} can be estimated as

$$\begin{aligned} \phi_{in} &= \exp(\lambda - 1) \beta_{n\tau} \exp \left(\Psi(\gamma_n) - \Psi \left(\sum_{j=1}^{k} \gamma_j \right) \right) \\ &\propto \beta_{n\tau} \exp \left(\Psi(\gamma_n) - \Psi \left(\sum_{j=1}^{k} \gamma_j \right) \right) \ . \end{aligned} \tag{A.17}$$

Now, we maximize the approximated log-likelihood with respect to γ_n. If take the derivative of the approximated log-likelihood with respect to γ_n, we can obatin

$$\begin{aligned}\frac{\partial}{\partial\gamma_n}L_{[\gamma_n]}(\gamma,\phi|\alpha,\beta) &= (\alpha_n-1)\left(\Psi'(\gamma_n)-\Psi'\left(\sum_{j=1}^{k}\gamma_j\right)\right)+\sum_{i=1}^{N}(\phi_{in})\left(\Psi'(\gamma_n)-\Psi'\left(\sum_{j=1}^{k}\gamma_j\right)\right)\\ &\quad-\Psi\left(\sum_{i=1}^{k}\gamma_i\right)+\Psi(\gamma_n)-\left(\Psi(\gamma_n)-\Psi\left(\sum_{i=1}^{k}\gamma_i\right)\right)\\ &\quad-(\gamma_n-1)\left(\Psi'(\gamma_n)-\Psi'\left(\sum_{i=1}^{k}\gamma_i\right)\right)\\ &= (\alpha_n-\gamma_n+\sum_{i=1}^{N}\phi_{in})\left(\Psi'(\gamma_n)-\Psi'\left(\sum_{i=1}^{k}\gamma_i\right)\right).\end{aligned} \tag{A.18}$$

To make the derivative zero, the following condition should be satisfied:

$$\gamma_n = \alpha_n + \sum_{i=1}^{N}\phi_{in} \tag{A.19}$$

Therefore, in the variational inference method, an iterative procedure of (A.17) and (A.19) alternatively is required until it converges.

Appendix B

Variational approximation method for Latent Dirichlet Allocation: Parameter Estimation

In many estimation processes, parameters are often chosen to maximize the likelihood values of a given data $\mathbf{w}$. The likelihood can be defined as

$$\begin{aligned} l(\alpha, \beta) &= \sum_{w \in \mathbf{w}} \log p(w|\alpha, \beta) \\ &= \sum_{d=1}^{M} L_d(\gamma, \phi|\alpha, \beta) \end{aligned} \tag{B.1}$$

where $L_d(\gamma, \phi|\alpha, \beta)$ represents the log-likelihood of the marginal probability of the document d.

Firstly, we can perform the Lagrange multiplication to estimate β as we introduced in the previous section. In the multiplier, we have a constraint $\sum_{m=1}^{V} \beta_{nm} = 1$

$$L_{[\beta]}(\gamma, \phi|\alpha, \beta) = l(\alpha, \beta) + \sum_{n=1}^{k} \lambda_n \left(\sum_{m=1}^{V} \beta_{nm} - 1 \right). \tag{B.2}$$

If we take the derivative with respect to β_{nm}, we can obtain

$$\begin{aligned} \frac{\partial}{\partial \beta_{nm}} L_{[\beta]}(\gamma, \phi|\alpha, \beta) &= \sum_{d=1}^{M} \frac{\partial}{\partial \beta_{nm}} L_d(\gamma, \phi|\alpha, \beta) + \lambda_n \\ &= \sum_{d=1}^{M} \sum_{i=1}^{N_d} (\phi_{in})_d (w_{im})_d + \lambda_n \end{aligned} \tag{B.3}$$

where d represent the index of documents. To make this derivative zero, the following should be satisfied

$$\beta_{nm} = c \sum_{d=1}^{M} \sum_{i=1}^{N_d} (\phi_{in})_d (w_{im})_d \tag{B.4}$$

where c is a constant.

Next, we can take a derivative with respect to α that maximizes $l(\alpha, \beta)$, i.e.

$$\begin{aligned}
\frac{\partial}{\partial \alpha_n} l(\alpha, \beta) &= \sum_{d=1}^{M} \frac{\partial}{\partial \alpha_n} L_d(\gamma, \phi | \alpha, \beta) \\
&= \sum_{d=1}^{M} \left\{ \Psi(\alpha_n) - \Psi\left(\sum_{j=1}^{k} \alpha_j \right) + \Psi(\gamma_n) - \Psi\left(\sum_{j=1}^{k} \gamma_j \right) \right\} \\
&= M \left\{ \Psi(\alpha_n) - \Psi\left(\sum_{j=1}^{k} \alpha_j \right) \right\} + \sum_{d=1}^{M} \left\{ \Psi((\gamma_n)_d) - \Psi\left(\sum_{j=1}^{k} ((\gamma_j)_d) \right) \right\} .
\end{aligned} \tag{B.5}$$

This derivative, however, cannot be optimized with respect to α_n since it depends on α_j $(j \neq n)$ as well. Therefore, we take another partial derivative with respect to α_j:

$$\frac{\partial}{\partial \alpha_n \alpha_j} l(\alpha, \beta) = \delta(n, j) M \Psi'(\alpha_n) - \Psi'\left(\sum_{j=1}^{k} \alpha_j \right) , \tag{B.6}$$

which can be solve by Newton-Raphson optimization method which utilizes Hessian matrix [59].

In this parameter estimation of variational approximation method, we use expectation-maximization strategy so that it can repeatedly calculate the log-likelihood and estimate α and β until it converges.