Source Separation of Musical Instrument Sounds in Polyphonic Musical Audio Signal and Its Application

Katsutoshi Itoyama

Preface

This work was carried out at Okuno & Ogata Laboratory, Graduate School of Informatics, Kyoto University. I wish to express my gratitude to everyone who has supported this work.

First of all, I would like to express my deepest gratitude to Hiroshi G. Okuno for giving me the opportunity to work and supervising me for these years. The fact that the thesis is finished at all is in great part of his endless enthusiasm for talking about my work.

I also specially thank to Dr. Masataka Goto (AIST). The key idea in the thesis of the integrated model is inspired by the acoustic signal synthesis method of D-50, his favorite synthesizer. It may not be exaggeration to say that this study has started from his advice. He greatly supported me for theoretical and essential advices.

Professor Tatsuya Kawahara and Professor Toshiyuki Tanaka raised some essential questions about the methodology of source separation techniques of the thesis. His insightful comments enriched the content of the thesis and enlightened future directions of this research.

I want to thank students in musical information processing group at Okuno & Ogata Laboratory for lots of interesting discussions, various help, and making life at the laboratory so enjoyable: Mr. Takehiro Abe (currently with Nintendo), Mr. Kouhei Sumi (currently with Yamaha Corporation), Mr. Akira Maezawa, Mr. Naoki Yasuraoka, Mr. Shimpei Aso, and Mr. Naoki Nishikawa. I especially want to thank: Mr. Abe for always bringing enjoyable and boisterous environments to our desk; Mr. Sumi for studying chord recognition; Mr. Maezawa for helping me with statistical signal processing and English expressions; Mr. Yasuraoka for musical instrument sound analysis and synthesis, numerical and musical discussions; Mr. Aso for analysis and synthesis of singing and speech voices; and Mr. Nishikawa for discussions about music and emotion. Ms. Hiromi Okazaki and Ms. Miki Nishii (currently with Graduate School of Engineering, Kyoto University) provided secretarial support at our laboratory. The other members from other groups at Okuno & Ogata Laboratory contributed to discuss about my work, give interesting

comments, and create an enjoyable atomosphere and I thank them for experience. It was a great time and I would have never managed to complete the thesis without all the help I got.

The Japan Society for the Promotion of Science (JSPS) financially supported my life as a researcher.

Finally, I wish to thank my family: my parents and brother. I thank them for sharing so much with me, for always being there for me, encouragement, and support.

Abstract

A change of music appreciation style from *"listening to high fidelity (Hi-Fi) sounds"* to *"listening to preferred sounds"* has emerged due to evolution of digital audio processing technology for the past years. Previously, many people enjoyed passive music appreciation: e.g., they buy CD and phonograph recordings or download mp3 audio files, set the disks or files to various media players, and hit the play button. For the moment, only musical experts with signal processing expertise can enjoy active music appreciation.

To allow more music novices to enjoy active music appreciation, we developed a functional audio player, named *INTER (INstrumenT EqualizeR)*. This player enables users to change the volume of each musical instrument part in an audio mixture such as commercial CD recordings. INTER requires musical audio signals in which each musical instrument performs solo. Solo performance of each musical instrument is not generally available. Therefore, these solo musical instrument performances must be separated from an audio mixture of the musical piece. In other words, sound source separation is mandatory.

In the thesis, we focus on sound source separation that extracts all musical instrument sounds from polyphonic musical audio signal. Our goals are to design and implement a sound source separation method and to apply the method to a functional audio player which enables users to edit audio signals of existing musical pieces according to their preference, and query-by-example music information retrieval. Musical audio signals are usually polyphonic with 5 – 20 musical instruments and consist of both harmonic and inharmonic musical instrument sounds. Therefore, we tackle three technical issues in sound source separation for monaural polyphonic musical audio signal: (i) spectral modeling comprising harmonic and inharmonic instrument sounds, (ii) recognition of complex musical instrument sound mixture, and (iii) ensuring property of instrument to the spectral models. To solve the issue (i), we propose the integrated model that captures harmonic and inharmonic tone models. To solve the issue (ii), we propose a score-informed sound source separation. To solve the issue (iii), we propose a parameter estimation method

using prior distributions of the timbre parameters.

Chapter 3 presents a method for sound source separation based on maximum likelihood estimation for musical audio signals including both harmonic and inharmonic instrument sounds. We solve the issue (i) in this chapter. We define the integrated weighted mixture model consisting of harmonic and inharmonic models to represent the spectrogram of various musical instrument sounds. To decompose the magnitude spectrogram of the input audio mixture, we introduce spectral distribution functions to formulate the sound source separation problem and derive the optimal distribution function. Experimental evaluation results show that source separation performance improves by integrating the harmonic and inharmonic models.

Chapter 4 presents methods to separate musical audio signals based on maximum *A Posteriori* estimation using the integrated harmonic and inharmonic models. For prior information, we use the musical score corresponding to the audio. We solve the issues (ii) and (iii) in this chapter. We use a musical score such as a standard MIDI file (SMF) to initialize the model parameters corresponding to onset time, pitch, and duration by using the score. We introduce two approaches of instrument timbre modeling: template sounds and prior distributions of the model parameters. Template sounds are sound examples that generated by playing back each musical note of the SMF on a MIDI sound module. We initialize the model parameters by adapting them to the template sounds and then separate the observed spectrogram as we described in Chapter 3. The template sounds constrain the model parameters for each musical sound. Prior distributions of the model parameters are trained from a musical instrument sound database. The prior distributions constrain the model parameters for each musical instrument. Experimental results show that the quality of separated sounds based on the prior distributions is better than ones based on the template sounds.

Chapter 5 presents two applications that use sound source separation results. First, we describe INTER that allows users to control the volume of each instrument part within existing audio recordings in real time. Users can manipulate volume balance of the instruments and remix existing musical pieces. Second, we describe a Query-by-Example (QBE) approach in music information retrieval that allows a user to customize query examples by directly modifying the volume of different instrument parts. Our QBE system first separates all instrument parts from the audio signal of a piece with the help of its musical score, and then it lets users remix these parts to change the acoustic features

that represent the musical mood of the piece. Experimental results show that the shift was actually caused by the volume change in the vocal, guitar, and drum parts.

Chapter 6 discusses the major contributions made by this study to different research fields, particularly to sound source separation and instrument sound representation. We also discuss issues that still remain to be resolved and future directions we wish to research.

Chapter 7 concludes the thesis.

論文梗概

受動的な音楽の楽しみは「良い音（Hi-Fi な音）を聴くこと」であるのに対して，能動的な音楽の楽しみは「好みの音を聴くこと」であると言える．従来の音楽鑑賞スタイルは受動的なものが中心であったが，デジタル音響処理技術の発達により，能動的なものに変わりつつある．これまでの受動的な音楽の楽しみといえば，CD やレコードの購入や mp3 オーディオファイルのダウンロードし，ディスクやファイルをメディアプレーヤにセットし，再生ボタンを押す，といったものであった．能動的に音楽を楽しむことができるのは，作曲・編曲・楽器演奏などの技術や道具を持つ音楽のエキスパートに限られていた．

音楽の初心者でも手軽に能動的な音楽鑑賞を楽しむことができるように，我々は音楽音響信号（音楽 CD など）中の楽器音量バランスを操作することができるオーディオプレーヤ，楽器音イコライザ INTER (INstrumentT EqualizeR) を開発した．楽器音イコライザは，各楽器パートの演奏がソロで録音された音響信号を必要とする．楽曲制作者ならばこのようなデータを入手可能だが，一般のリスナには入手困難である．従って，楽器音イコライザを実現するためには，多重奏の音楽音響信号を楽器パートに分離する，すなわち音源分離が不可欠である．

本論文では，多重奏の音楽音響信号を個々の楽器音へと分離することに焦点を当てる．本研究の究極的な目標は，混合音の解析・分離手法を構築し，楽曲中の音楽的要素を自由に操作できるオーディオプレーヤを実現することである．我々が日常耳にするポピュラー音楽は，通常は多重奏で，5 から 20 程度の楽器が用いられており，楽器は調波的なもの，非調波的なものの両方がある．我々は，音源分離における以下の 3 つの課題に取り組む．(i) 調波・非調波を問わず，あらゆる楽器音を表現可能な楽器音のスペクトロモデリング，(ii) 複雑な音楽音響信号の認識，(iii) 個々の楽器音モデルの楽器特徴保持．課題 (i) を解決するため，調波・非調波統合モデルを開発した．課題 (ii) を解決するため，楽譜を援用した音源分離手法を開発した．課題 (iii) を解決するため，楽器音モデルのパラメータ事前分布を用いたパラメータ推定手法を開発した．

第 3 章では，モノラル音響信号に対して適用可能な楽器音のスペクトロモデリングである調波・非調波統合モデル，および統合モデルを用いた最尤推定に基づく音源分離・パ

ラメータ推定手法について述べる．この章では，課題 (i) を解決する．音源分離問題を振幅スペクトルの分解と定義し，分解のための分配関数を目的関数の最大化から導出する．評価実験により，調波モデルと非調波モデルを統合することで，様々な楽器音を適切にモデル化することができ，分離性能も向上することが示された．

第 4 章では，統合モデルを用いた音楽音響信号の音源分離において，楽譜を援用し，最大事後確率推定に基づいて楽器音モデルの楽器音響特性を保持する手法について述べる．この章では，課題 (ii) と (iii) を解決する．標準 MIDI ファイルなどの楽譜に基づき，モデルの発音時刻・音高・音長パラメータを初期化する．楽器音響特性保持のため，テンプレート音を用いたパラメータ推定法とパラメータ事前分布を用いたパラメータ推定法を導入する．テンプレート音は，楽譜中の音符を一つずつ演奏することで得られる音のサンプルである．音モデルをテンプレートに適応させることでモデルパラメータを初期化し，その後混合音の分離を行う．テンプレート音は，モデルごと（音符ごと）の音色パラメータ制約としてはたらく．パラメータ事前分布は，楽器音データベースを用いて楽器ごとに学習する．この事前分布は，楽器ごとの音色パラメータ制約としてはたらく．評価実験により，事前分布を用いた音源分離・パラメータ推定の方が，テンプレート音を用いた手法よりも分離性能が高いことが示された．

第 5 章では，音源分離結果を応用した 2 つのアプリケーションについて述べる．1 つは，楽器音量バランスをリアルタイムに操作可能なオーディオプレーヤ，楽器音イコライザである．楽器音量バランスの操作に伴い楽曲の雰囲気が変化するため，ユーザは好みの楽器音量バランスで楽曲を楽しむことができる．もう 1 つは，楽器音イコライザの類似楽曲検索への応用である．類似楽曲検索ではクエリとして楽曲を用いる．楽器音イコライザでクエリ楽曲をカスタマイズすることで，既存楽曲をそのままクエリとして用いるよりも多様な検索結果を得ることができる．評価実験により，歌声パート，ギターパート，ドラムパートの音量バランスを操作することで検索結果が変化し，かつその変化は楽曲のジャンルと整合していることが示唆された．

第 6 章では，音源分離や楽器音認識の分野における本研究の貢献について述べる．また，本研究では扱いきれなかった課題や，本研究の今後の方向性についても述べる．第 7 章で本論文を結ぶ．

Contents

List of Figures

List of Tables

List of Symbols

$X(t, f)$	Observed (to be separated) magnitude spectrogram
$\hat{X}_j(t, f)$	Separated spectrogram of the j-th note
$\hat{X}_{j,\mathrm{H}}(t, f)$	Separated spectrogram of the harmonic component
$\hat{X}_{j,\mathrm{I}}(t, f)$	Separated spectrogram of the inharmonic component
$\hat{X}_{j,\mathrm{H},l,m}(t, f)$	Separated spectrogram of the (l, m)-th harmonic component
$\hat{X}_{j,\mathrm{I},l,m}(t, f)$	Separated spectrogram of the (l, m)-th inharmonic component
$\xi_j(t, f)$	Template spectrogram of the j-th note
$\hat{\xi}_{j,\mathrm{H}}(t, f)$	Separated template spectrogram of the harmonic component
$\hat{\xi}_{j,\mathrm{I}}(t, f)$	Separated template spectrogram of the inharmonic component
$\hat{\xi}_{j,\mathrm{H},l,m}(t, f)$	Separated template spectrogram of the (l, m)-th harmonic component
$\hat{\xi}_{j,\mathrm{I},l,m}(t, f)$	Separated template spectrogram of the (l, m)-th inharmonic component
$Y(t, f)$	Model of the whole spectrogram
$Y_j(t, f)$	Model of the j-th note's spectrogram
$Y_{\mathrm{H}\|j}(t, f)$	Model of the harmonic component
$Y_{\mathrm{I}\|j}(t, f)$	Model of the inharmonic component
$Y_{l\|j,\mathrm{H}}(t)$	Model of the l-th temporal component of the harmonic component
$Y_{m\|j,\mathrm{H}}(f)$	Model of the m-th frequency component of the harmonic component
$Y_{l\|j,\mathrm{I}}(t)$	Model of the l-th temporal component of the inharmonic component
$Y_{m\|\mathrm{I}}(f)$	Model of the m-th frequency component of the inharmonic component
$Z(j;\, t, f)$	Distribution function which distribute $X(t, f)$ into $\hat{X}_j(t, f)$
$Z(\mathrm{H};\, j, t, f)$	Distribution function which distribute $\hat{X}_j(t, f)$ into $\hat{X}_{j,\mathrm{H}}(t, f)$

$Z(\mathrm{I};\, j, t, f)$	Distribution function which distribute $\hat{X}_j(t, f)$ into $\hat{X}_{j,\mathrm{I}}(t, f)$
$Z(l, m;\, j, \mathrm{H}, t, f)$	Distribution function which distribute $\hat{X}_{j,\mathrm{H}}(t, f)$ into $\hat{X}_{j,\mathrm{H},l,m}(t, f)$
$Z(l, m;\, j, \mathrm{I}, t, f)$	Distribution function which distribute $\hat{X}_{j,\mathrm{I}}(t, f)$ into $\hat{X}_{j,\mathrm{I},l,m}(t, f)$
w_j	Relative weight (magnitude) of the j-th note
$w_{\mathrm{H}\mid j}, w_{\mathrm{I}\mid j}$	Relative weight of the harmonic and inharmonic components
$w_{l\mid j,\mathrm{H}}$	Relative weight of the l-th temporal Gaussian functions of the harmonic component
$w_{m\mid j,\mathrm{H}}$	Relative weight of the m-th frequency Gaussian functions of the harmonic component
$w_{l\mid j,\mathrm{I}}$	Relative weight of the l-th temporal Gaussian functions of the inharmonic component
$w_{m\mid j,\mathrm{I}}$	Relative weight of the m-th frequency Gaussian functions of the inharmonic component
τ_j	Mean of the first temporal Gaussian function (onset time)
ρ_j	Standard deviation of the harmonic temporal Gaussian function
ϱ_j	Standard deviation of the inharmonic temporal Gaussian function
ϕ_j	Mean of the first frequency Gaussian function (fundamental frequency)
σ_j	Standard deviation of the harmonic frequency Gaussian function
φ, ς	Coefficients which determines the arrangement of the inharmonic frequency Gaussian functions
$\tilde{\omega}_{\mathrm{H}\mid k_j}, \tilde{\omega}_{\mathrm{I}\mid k_j}$	Prior parameters for $w_{\mathrm{H}\mid j}$ and $w_{\mathrm{I}\mid j}$
$\tilde{\omega}_{m\mid k_j,\mathrm{H}}$	Prior parameter for $w_{m\mid j,\mathrm{H}}$
$\tilde{\omega}_{m\mid k_j,\mathrm{I}}$	Prior parameter for $w_{m\mid j,\mathrm{I}}$

Chapter 1

Introduction

This chapter briefly describes the motivation, goal, issues, and approaches of the thesis.

1.1 Motivation

A change of music appreciation style from *"listening to high fidelity (Hi-Fi) sounds"* to *"listening to preferred sounds"* has emerged due to evolution of digital audio processing technology for the past years. The former is passive music appreciation and the latter is active music appreciation. Previously, many people enjoy the passive music appreciation: e.g., they buy CD and phonograph recordings or download mp3 audio files, set the disks and files to various media players, and hit the play button. Passive music appreciation involves using popular and sophisticated audio technologies. For example, 2.1 channel, 5.1 channel, or 7.1 channel sound systems provide highly realistic sensations and we can enjoy vivid musical instruments or sound sources. Active noise cancellation with a headphone may make a quiet acoustic environment. Digital audio processing technologies have been developed to deepen passive music appreciation.

On the other hand, demand for active music appreciation, which is symbolized by consumer generated media (CGM) and user generated content (UGC), has been increasing. Previously, active music appreciation has been enjoyed by a limited number of people due to it involves by particular and technical knowledge, experience, and equipment. For example, musical composition and arrangement may require knowledge of musical structure and chord progression. To enjoy performing musical instrument, adequate training and of course musical instrument itself are required. For the moment, only musical experts with signal processing expertise can enjoy active music appreciation.

To allow more music novices to enjoy active music appreciation, we developed a func-

tional audio player, named *INTER (INstrumenT EqualizeR)*. This player enables users to change the volume of each musical instrument part in an audio mixture such as commercial CD recordings. Changing the volume of the instruments is easy even for musical novices. Since the musical mood of a piece is influenced by instrumentation and the volume balance of the instruments, INTER can indirectly change musical mood and helps users *listen to their preferred sounds.*

INTER requires musical audio signals in which each musical instrument performs solo. Solo performance of each musical instrument is not generally available. To obtain these solo musical audio signals, each musical instrument part must be separated from an audio mixture of the musical piece, i.e., sound source separation is mandatory. Automatic transcription for polyphonic musical audio has been researched since the 1970s [1,2], and, since the 1980s, source separation has been tackled as an evolutionary problem related to transcription [3–10]. For example, methods of musical instrument identification [11] have been reported based on fundamental frequency (F0) estimates [12–14]. At that time, source separation methods treated only the sounds of pitched musical instruments. Beat tracking techniques [15–17] have been developed from the 1990s and the sounds of unpitched musical instruments began to be recognized and separated [18]. Until now few studies tried to recognize and separate both pitched and unpitched sounds.

1.2 Goal

In the thesis, we focus on sound source separation that extracts all solo musical instrument sounds corresponding to each musical note from polyphonic musical audio signal. Our goals are to design and implement a sound source separation method and to apply the method to a functional audio player which enable users to edit audio signals of existing musical pieces according to their preference, and query-by-example music information retrieval [19]. Musical audio signals are usually polyphonic with 5 – 20 musical instruments and consist of both harmonic and inharmonic musical instrument sounds. We deal with musical audio signals which have following properties:

Consisting of both harmonic and inharmonic musical instrument sounds.

Various musical instruments are used in popular music. Guitar, bass, and drums are basic instruments and pianoforte, synthesizer, saxophone, bowed strings, flute, and bell, as well as other instruments. These musical instruments are roughly

divided based on auditory features into two groups: one has *harmonic* sounds, e.g., pianoforte, guitar, and flute, and the other has *inharmonic* ones, e.g., drums. To separate sounds of various instruments, it is necessary to model the spectrogram of them.

Complexly-polyphonic musical audio signals.

To separate the sounds of the instruments, we have to recognize each sound in the audio mixture. In popular music, hundreds to thousands of instrument sounds exist in one piece, and five to thirty sounds are played on average. Since commercial CD recordings include stereo audio signals, the number of instrument sounds always exceeds the number of channels. In many popular songs, original audio signals are recorded in monaural for each instrument performance and mixed into stereo by changing the localization of the instruments on the basis of the volume balance of right and left.

To separate precisely the instrument sounds, we have to achieve instrument timbre representation. The characteristics of musical sounds are represented as the pitch, the duration, and the timbre. The pitch and the duration of sounds change in the same instrument, but the timbre of them is consistent. Instrument timbre representation which has high affinity to the method of source separation and spectral modeling is important.

1.3 Issues and Approaches

To achieve the above goals, we tackled the following issues:

Issue 1: Spectral modeling comprising harmonic and inharmonic sounds.

Instrument sounds have both aspects of harmonic and inharmonic sounds. Piano sounds consist of slowly decaying harmonic sounds by the stationary vibration of strings and rapidly decaying inharmonic sounds by the striking of hammers. Flute sounds consist of harmonic sounds by the vibration of the air column and inharmonic sounds by the complex airflow on the mouthpiece. Drum sounds consist of almost only inharmonic sounds by the striking of sticks and the vibration of membranes.

As we described in Section 1.1, the separation of harmonic sounds and inharmonic sounds have been separately studied. Many source separation methods aim to sepa-

rate or extract only target sounds. If harmonic sounds are separated from a mixture of harmonic and inharmonic sounds first, inharmonic sounds would be decomposed into separated harmonic sounds and the separation error would accumulate. Thus, instead of separating harmonic and inharmonic sounds independently, these sounds should be separated equally within the same framework.

Issue 2: Recognition of complex musical instrument sound mixture.

To separate an audio mixture into sound sources, information of sound sources, i.e., onset time, fundamental frequency (F0), duration, and the instrument of each instrument sound, need to be obtained by recognizing the sound of a particular instrument. As we described in Section 1.2, 5–30 instrument sounds that include both harmonic and inharmonic sounds. However, there are no instrument sound recognition methods which can analyze a complex audio mixture as popular music.

Issue 3: Ensuring property of instrument to the spectral models.

Due to differences in instrument manufacture, physical characteristics, performance styles, and audio effects, the timbre of instrument sounds will vary even if the pitch, duration, and volume of the sounds are the same. Since template sounds are generated from a sound generator which has defined individual instruments and performance styles, there is some kind of timbre difference between the template sounds and sound sources. A large difference in timbre degrades the convergence of the model parameter estimation and distorts the separated instrument sounds.

We aim to solve these issues as follows:

Solution 1: Harmonic and inharmonic integrated model.

We define a novel instrument sound model on a magnitude spectrogram, named *harmonic and inharmonic integrated model*, and separate an audio mixture into each sound source. The integrated model consists of a harmonic and an inharmonic one which represent the harmonic and inharmonic components of an instrument sound, respectively. The parameters of the integrated model represent onset time, fundamental frequency (F0), relative magnitude of harmonics, etc. By changing the parameters, the integrated model represents various magnitude spectrograms of the sound of an instrument. The parameters of multiple integrated models are estimated from the magnitude spectrogram of the audio mixture and the spectrogram is decomposed into each the sound of instrument.

Solution 2: Musical score as prior information.

We use a musical score instead of instrument sound recognition and obtain information about the sound sources from the score. We assume that the score is temporally-aligned to the audio signal.

We generate sound examples, named *template sounds*, which have similar acoustic features to the instrument sound in the mixture by playing each musical note in the score using a sound generator. Using the template sounds as an initial constraint on the model parameter estimation provides better convergence.

Solution 3: Prior distribution of model parameters.

For each musical instrument, we train probable model parameter values as distributions and estimate parameters from the magnitude spectrogram of audio mixture on the basis of a maximum *A Posteriori* estimation by using the distributions as prior information.

1.4 Problem Specification

Let t, and f be variables which represent time and frequency, respectively, and $X(t, f)$ be an observed energy distribution. Here, $X(t, f)$ is defined on $t \in \mathbb{T}$ ($\mathbb{T} = [T_0, T_1]$), and $f \in \mathbb{F}$ ($\mathbb{F} = [F_0, F_1]$). Here, the problem to be solved is distribution of the observed energy distribution $X(t, f)$ into energy distributions which belong to each auditory event, i.e., each musical note.

The observed energy distribution $X(t, f)$ at the coordinate (t, f) does not always belong to single auditory event but to the sum of the energies of multiple events. Therefore, the energy at each coordinate should not be dominated by single auditory event exclusively but shared by multiple events. Although the energy is not additive, for simplicity we assume that the energy or magnitude is additive.

Let J be the number of auditory events, we introduce a function, $Z(j;\, t, f)$, to distribute the observed energy distribution $X(t, f)$ into the j-th auditory event. The distribution function $Z(j;\, t, f)$ satisfies the condition:

$$\forall t \in \mathbb{T}, f \in \mathbb{F} : \sum_{j=1}^{J} Z(j;\, t, f) = 1, \tag{1.1}$$

and the product of them, $Z(j;\, t, f)\, X(t, f)$ means the decomposed energy distribution of the j-th event.

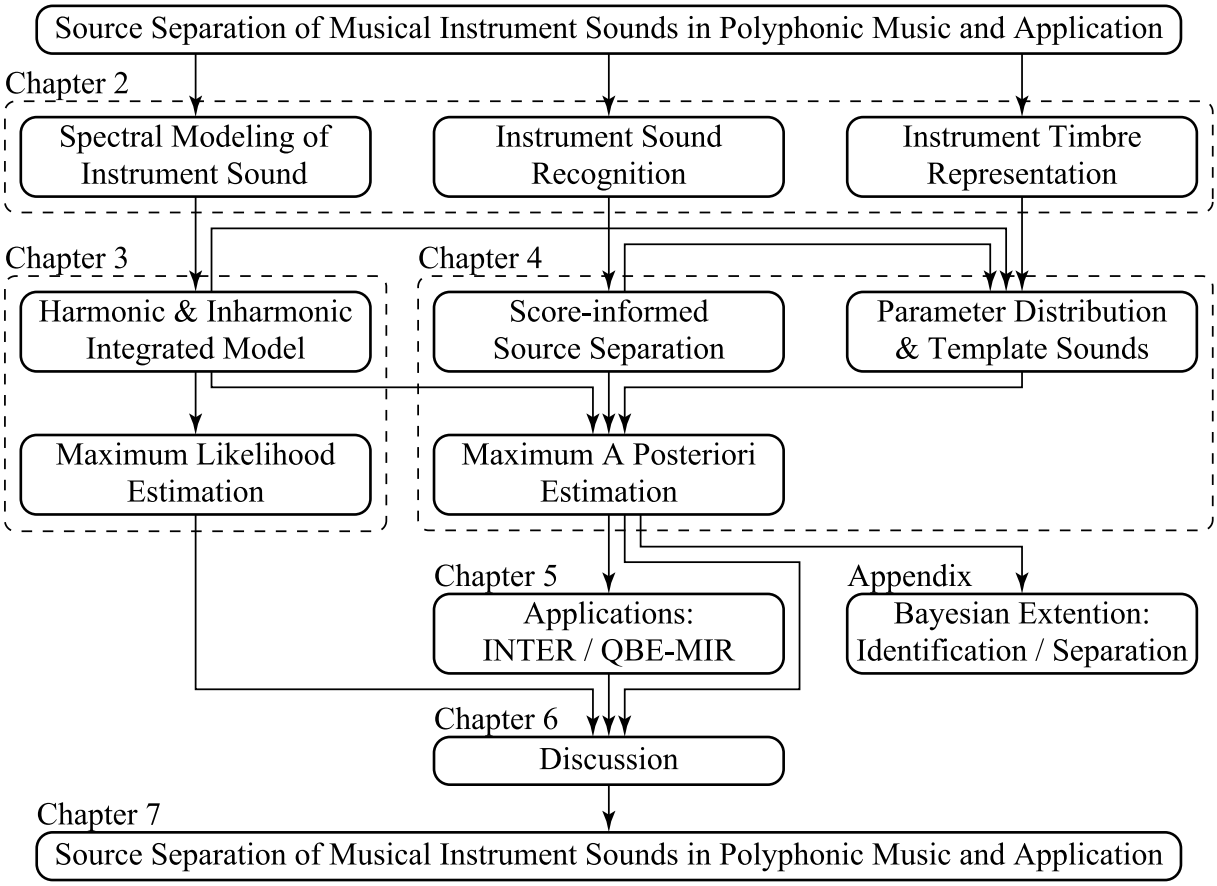


Figure 1.1: Organization of the thesis.

We assume the musical score as a standard MIDI file (SMF) which contains the following information for each musical note: instrument name given by the program number (for a pitched instrument) or the note number (for a percussive instrument), F0 given by the note number, onset time given by the tick of note-on message, and duration given by the difference of the tick from note-on to note-off messages. Other information, e.g., vibrato given by the modulation, stereo localization given by the panpot, and volume given by the expression and the velocity, are discarded. We also assume that the input audio mixture and the SMF are aligned in terms of time.

1.5 Thesis Organization

The organization of the thesis is shown in Figure 1.1. Chapter 2 provides a review of the literature in the fields of sound source separation for musical audio mixtures. Chapter 3 describes a method for sound source separation for polyphonic musical audio signals by

using the integrated weighted mixture model consisting of a harmonic-structure model and an inharmonic-structure model. Chapter 4 describes a method to separating musical audio signals using the integrated harmonic and inharmonic models and prior information based on the musical score corresponding to the audio. Chapter 5 describes two applications which uses sound source separation results, INTER and a query-by-example music information retrieval system using INTER. Chapter 6 discusses the major contributions of our studies to different research fields including sound source separation. Issues still remaining and future directions are also discussed from these standpoints. Chapter 7 concludes the thesis.

Chapter 2

Literature Review

This chapter provides a review of the literature related to sound source separation for musical audio mixtures and application using separated sources to clarify the standpoint of the thesis within related fields.

2.1 Sound Source Separation

We summarize several methods for instrument sound representation and sound source separation and discuss possible problems and benefits of them for this study. The goal of this study is sound source separation of musical audio mixtures into all instrument sounds. To achieve the source separation, it is needed to clarify the correspondence between the audio mixture and musical instrument sounds by representing the properties of the sounds, i.e., pitch, duration, and spectral shape.

2.1.1 Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) [20] is one of the most widely used method for sound source separation. NMF decomposes a non-negative $N \times M$ matrix V into $N \times R$ and $R \times M$ factor matrices W and H based on the signal model:

$$V \simeq WH. \tag{2.1}$$

V represents an observed magnitude or power spectrogram. Each column and row vector of W and H represent frequency structure and temporal magnitude change for each factorized sound source, respectively. Most spectrograms of musical instrument sounds can be well factorized since instrument sounds have stationary frequency structures denoted by the pitch and timbre. Since R is determined as $R \ll N, M$, NMF is used for

information extraction and data compression. W and H are estimated by minimizing the objective function which is defined as the Euclidean distance:

$$\|V - WH\|^2 \tag{2.2}$$

or the I-divergence:

$$\sum_{n,m} \left((V)_{n,m} \log \frac{(V)_{n,m}}{(WH)_{n,m}} - (V)_{n,m} + (WH)_{n,m} \right). \tag{2.3}$$

In the musical audio processing field, power spectrograms which have non-negative matrix representation are widely used and many power spectrograms of musical instrument sounds can be decomposed the product of two factors: frequency structure and temporal power variation. NMF has been adapted to sound source separation of musical audio signals by these two reasons.

By focusing on the above, Smaragdis *et al.* [21] adapted NMF for automatic music transcription and sound source separation. They treated musical audio mixture consisting of only harmonic instrument sounds. On the other hand, Kim and Yoo *et al.* [22,23] extended NMF, called non-negative matrix partial co-factorization (NMPCF), for an unsupervised method of separating rhythmic sources. Drum sounds were extracted by factorizing both training data which consist of only drum sounds and observed data which consist of both drum and harmonic sounds.

Extended ways of NMF for dealing with both harmonic and percussive instrument sounds were proposed by Helén *et al.* [24] and Virtanen [25]. The method of Helén *et al.* was based on two-stage processing in which the input signal is first separated into elementary time-frequency components which are then organized into sound sources. NMF is used to separate the input spectrogram into components having a fixed spectrum with time-varying gain. Each component is classified either to pitched instruments or to drums using a support vector machine (SVM). Virtanen [25] presented an unsupervised learning algorithm for the separation of sound sources in one-channel music signals which is based on factorizing the non-negative magnitude spectrogram of an input signal into a sum of components, each of which has a fixed magnitude spectrum and a time-varying gain. Temporal continuity is favored by using a cost term which is the sum of squared differences between the gains in adjacent frames, and sparseness is favored by penalizing non-zero gains.

Another extended ways of NMF for dealing with both harmonic and percussive instrument sounds were proposed by Smaragdis [26] and Schmidt *et al.* [27]. Smaragdis

presented a convolutive basis decomposition method for separating known types of sounds from monophonic mixtures. They introduced the concept of a convolutive nonnegative basis set, demonstrated how it maps to meaningful features in the case of audio spectra and demonstrated how we can use it in the context of supervised source separation. Schmidt *et al.* extended the convolutive NMF into two-dimensional convolutive version, called NMF2D, for blind separation of instruments in polyphonic music. Using a model which is convolutive in both time and frequency, they factorized a spectrogram representation of music into components corresponding to individual instruments. Based on this factorization, they separated the instruments using spectrogram masking.

An advantage of source separation and musical signal analysis by NMF is the formulation based on assuming the observed spectrogram as the sum of static spectral patterns on the two-dimensional time-frequency plane, rather than one-dimensional frequency spectrum which vary with time. On the basis of this perspective, the problem of estimation and adjustment of onset time and pitch results in the localization problem on the time-frequency domain. This perspective is also found in HTC [28], HTTC [29], and HPSS [30]. A disadvantage of NMF is that the factorized basis vectors do not always correspond to the instrument sounds. Therefore, NMF is unsuitable for source separation into each note.

2.1.2 Independent Component Analysis

Independent component analysis (ICA) [31, 32] is also widely used for sound source separation. ICA estimates source data from observed data which have instantaneous or convolutive mixture representation based only on the assumption that sources are statistically independent. Previously, ICA has been used in combination with microphone arrays for separating a few (about 2 – 4) speech signals [33], since ICA requires that the number of the observations, i.e., microphones, is equal to or larger than the number of the sources. However, typical musical audio signals are recorded in stereo and more than two musical instrument sounds are performed in most sections of the musical pieces.

Casey *et al.* [34] proposed independent subspace analysis (ISA) as an extension of ICA, ICA and ISA have been applied to sound source separation of musical audio signals. ISA separates observed audio mixtures by assuming the short-time Fourier transform coefficients for each frequency band as independent observations and virtually multiplying the number of the observations than the number of the sources. Barry *et al.* [35] and

Morita *et al.* [36] tackled the source separation by ICA on the basis of a similar problem consciousness. Casey *et al.* used a dissimilarity matrix, named ixegram, which is based on the Kullback-Leibler divergence between the estimated sources to cluster them. Dubnov [37] improved this clustering method by using a distance matrix based on a higher order statistical distance.

Source separation methods using ICA and ISA were improved for the separation without prior information. ISA has a permutation problem, i.e., ambiguity of scaling factors and their permutation. ISA treats at least hundreds of time series observations since ISA regards the STFT coefficients for each frequency as independent observations. Uhle *et al.* [38] and FitzGerald *et al.* [39] tackled this problem by performing the singular value decomposition to these time series to compress the size of dimension. Vincent [40] proposed a family of source separation methods for stereo mixtures of instrumental sources based on multilayer Bayesian network models of short-term power spectrum and interchannel phase difference, and designed a family of probabilistic mixture generative models combining modified positive ISA, localization models, and segmental models (SM). They expressed source separation as a Bayesian estimation problem and we propose efficient resolution algorithms. The resulting separation methods rely on a variable number of cues including harmonicity, spectral envelope, azimuth, note duration, and monophony.

ICA and ISA have an advantage that they can be applied to arbitrary time-series signals by assuming statistical independency of the sources. Although this assumption is correct in signals of speech mixture and simple musical pieces mixture of speech signals and simple musical audio signals, ICA and ISA are unsuitable for separating complex musical audio signals consisting of many synthesized sounds in which the independency is not correct. Additionally, by the same as NMF, estimated sources do not always correspond to the instrument sounds and musical notes.

2.1.3 Other Modeling

We describe methods of source separation of musical audio signals in signal modelings which are different to NMF and ICA. Since most of them deal with either harmonic nor inharmonic sounds exclusively, basically we cannot adopt them to this study.

Kameoka *et al.* proposed a multipitch analyzer, named harmonic temporal structured clustering (HTC), that estimates pitch, intensity, onset, duration, etc., of each underlying source in a multipitch audio signal HTC decomposes the energy patterns diffused in time-

frequency space, i.e., the power spectrum time series, into distinct clusters such that each has originated from a single source. HTC source models are defined a Gaussian kernel representation. An idea that the auditory parameter estimation problem is reduced to the localization problem by regarding the observed spectrogram as a two-dimensional time-frequency plane is helpful and we incorporated this idea in our study. Woodruff *et al.* [41,42] introduced a method, named active source estimation (ASE), that uses spatial cues from anechoic, stereo music recordings and assumptions regarding the structure of musical source signals to effectively separate mixtures of tonal music. However, these methods deal with only harmonic sounds and assume that the observed audio mixture consists of only harmonic ones.

Huang *et al.* [43] proposed a method for separating drum objects from polyphonic music signals. After a simple time domain separation method, auditory objects which are represented as the basis vectors of the NMF are classified into tonal or non-tonal components. A tonal-components tracking and attenuation (TTA) suppresses quasi-stationary auditory objects such as singing voice in the separated drum objects. Yoshii *et al.* [44,45] proposed a drum sound separation method based on drum-sound template-adaptation and harmonic structure suppression. An initial template of each drum sound, called a seed template, is prepared, and the first technique adapts it to actual drum-sound spectrograms appearing in the song spectrogram. Gillet *et al.* [46] also presented a method for music transcription and source separation with a focus on drum signals in a similar way to Adamast, which combines information from the original music signal and a drum track enhanced version obtained by source separation. Although these methods transcribe and separate drum sounds from polyphonic musical audio signals, have not discussed a combined way to harmonic sound separation.

Harmonic/percussive sound separation (HPSS) [30] is a separation method for a monaural audio signal into harmonic and percussive components without any assumptions except spectral shape. Spectrograms of harmonic and percussive sounds are represented as *horizontal and vertical* (along time-axis and frequency-axis) lines, respectively. Despite of the simplicity of the algorithm, pitched instruments and drums are well separated. Although HPSS has succeeded in dealing with both harmonic and inharmonic sounds, it is unsuitable for a detailed musical signal analysis such as separation into each instrument since HPSS adopted too simplified modeling.

2.2 Instrument Sound Recognition

An important task in the instrument sound recognition is multiple fundamental frequency estimation. The goal of this task is to estimate F0s of several dominant or all sounds in the polyphonic musical audio signal. Various multiple F0 estimation methods are proposed: based on a generative model of the harmonic structure [47, 48], based on a sinusoidal model [49], and based on a human auditory model [14, 50, 51]. However, these methods mainly deal with at most six harmonic polyphony audio signals. Since musical audio signals such as popular music are usually polyphonic with 5 – 20 musical instruments and consist of both harmonic and inharmonic musical instrument sounds, these methods cannot deal with actual music.

Another important task in the instrument sound recognition is onset time estimation. The goal of this task is to estimate the onset time for each musical instrument sound in the audio signal. Various onset time estimation methods are proposed: based on a sharp magnitude growth at the onset time [16, 52] and based on a harmonic change at the onset time of pitched sounds [53, 54]. These methods can recognize onset time of audio signals with simple rhythm, e.g., equally-spaced rhythm, while actual musical pieces sometimes complex onset patterns, e.g., trills and arpeggios.

A typical task in musical instrument timbre representation is musical instrument sound identification and classification. Most studies on instrument recognition for solo sounds [55–57] dealt comparatively with many kinds of instruments (between 10 and 30). Various acoustic features were used; some were designed based on the knowledge of musical acoustics (e.g., spectral centroid and odd/even energy ratio) [56, 57] and some were used in speech recognition (e.g., MFCCs and LPCs) [55]. The commonly used classifiers were the Gaussian [57], Gaussian mixture model (GMM) [55]. Although we cannot apply these methods and features to an analysis-synthesis system of musical instrument sounds since most of these acoustic features are irreversible, statistical classification and optimization methods by using parameters of an analysis-synthesis system should be helpful for timbre representation in the system.

2.3 Standpoint of The Thesis

In contrast to the comparison of our study with related fields above, we now compare our study with previous musical instrument recognition studies.

We consider two axes to classify the studies of sound source separation for musical audio mixture:

1. Unit of sources to be separated. "Group of instruments" means the studies which separate audio mixtures in some rough unit, e.g., harmonic and inharmonic sounds. "Specific instrument part" means the studies which separate mixtures into specific instrument parts and residuals in polyphonic mixture, e.g., vocal separation. "All instrument sound" means the studies which separate mixtures into all instrument sound.

2. Instrument sounds to be separated. "Harmonic sounds" means the studies which separate harmonic instrument sounds by assuming the audio mixture consists of only harmonic sounds. "Inharmonic sounds" means the studies which separate inharmonic instrument sounds from audio mixture which consists of only inharmonic sounds or both harmonic and inharmonic sounds. "Both sounds" means the studies which separate both harmonic and inharmonic instrument sounds from audio mixtures.

Figure 2.1 shows a classification of the studies based on these axes. Most studies are classified into "group of instruments" or "specific instrument part". Woodruff *et al.* [42] tackled to separate all instrument sounds but treated only harmonic sounds. This study is rich in originality in terms of the separation of all harmonic and inharmonic instrument sounds.

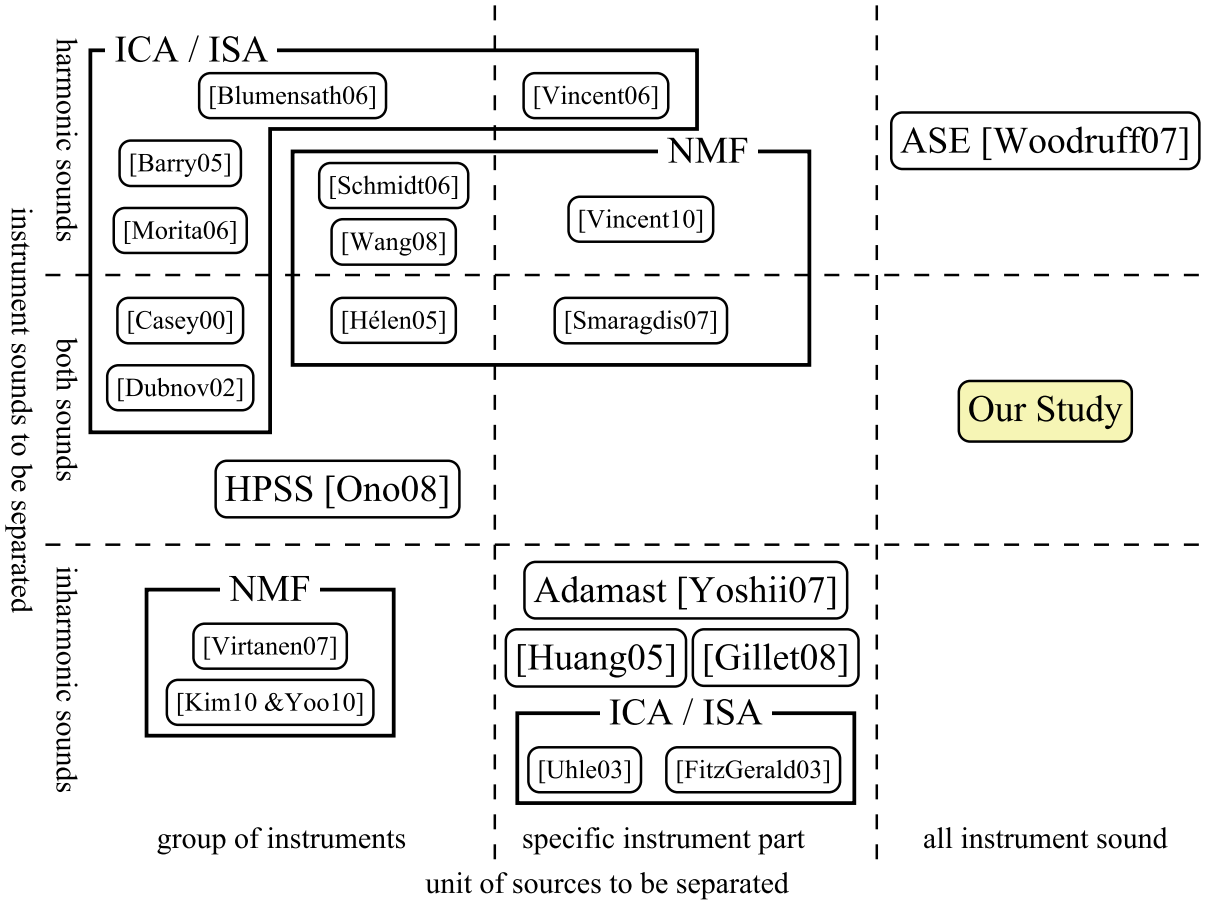


Figure 2.1: Positioning of the thesis.

Chapter 3

Separation of Harmonic and Inharmonic Instrument Sounds

This chapter describes a method for sound source separation for monaural musical audio signals. First, the properties of the musical audio signals and instrument sounds to be separated in this study are denoted. To decompose the magnitude spectrogram of the input audio mixture, we introduce spectral distribution functions and formulate the sound source separation problem and derive the optimal distribution function. We define the integrated weighted mixture model consisting of a harmonic-structure model and an inharmonic-structure model to model the spectrogram of various musical instrument sounds and derive update equations of the model parameters. An experimental evaluation result shows that source separation performance was improved by integrating the harmonic and inharmonic models.

3.1 Property of Musical Audio Signal

We deal with monaural musical audio signals, although separating monaural audio signals is more difficult than stereo or more channels but need less assumptions, e.g., recording conditions. In musical audio signals, 5–20 musical instruments are performed, 5–30 instrument sounds consisting of pitched and unpitched ones are emitted simultaneously but not always synchronized. Instrument sounds cannot be simply classified into pitched and unpitched since many instrument sounds contains both harmonic and inharmonic components, e.g., a pianoforte sound contains a harmonic component when the string is vibrating and an inharmonic component when the hammer hits the string. We exclude singing voices because they have complicated phoneme, expression, and pitch contour.

3.2 Decomposition of Magnitude Spectrogram

We target the magnitude spectrogram, the absolute value of the short-time Fourier transform (STFT) coefficients with a Gaussian window function, in which auditory events are located sparsely in the time-frequency domain at a certain level. A Gaussian window function is chosen because the spectrogram is compatible with spectral models of musical instrument sounds described below. We formulate the sound source separation problem as the decomposition of the magnitude spectrogram by assuming the additivity of the magnitude.

Let $X(t, f)$ be the observed magnitude spectrogram where t and f are the time and the frequency, respectively. Here, the problem to be solved is decomposing $X(t, f)$ into J spectrograms which belong to each musical note.

The observed magnitude at (t, f) does not always belong to a single musical instrument sound, but is derived several sounds because spectral leakage and audio effects such as reverberation. Magnitude for each (t, f) should be shared by several instrument sounds rather than be dominated by a single sound. We introduce a distribution function, $Z(j;\, t, f)$, to decompose $X(t, f)$ to the j-th musical note. The distribution function satisfies

$$\forall j = 1, \ldots, J, t \in \mathbb{T}, f \in \mathbb{F} :\ 0 \leq Z(j;\, t, f) \leq 1 \tag{3.1a}$$

and

$$\forall t \in \mathbb{T}, f \in \mathbb{F} :\ \sum_{j=1}^{J} Z(j;\, t, f) = 1, \tag{3.1b}$$

and a decomposed magnitude spectrogram for the j-th note, $\hat{X}_j(t, f)$, is represented as

$$\hat{X}_j(t, f) = Z(j;\, t, f)\, X(t, f)\,. \tag{3.2}$$

To formulate the sound source separation problem, some criteria must be defined to evaluate the performance of separation. By modeling the magnitude spectrogram of the j-th musical note as a function, $Y_j(t, f; \theta)$, with a parameter, θ, a pseudo-distance between $Y_j(t, f; \theta)$ and $\hat{X}_j(t, f)$, which is defined as:

$$\int_{\mathbb{T}} \int_{\mathbb{F}} \hat{X}_j(t, f) \log \frac{\hat{X}_j(t, f)}{Y_j(t, f; \theta)}\, df\, dt, \tag{3.3}$$

can be used for a criterion because it takes 0 if and only if $\hat{X}_j(t, f)$ and $Y_j(t, f; \theta)$ are

equal for all (t, f). Additionally, under the following condition:

$$\int_{\mathbb{T}} \int_{\mathbb{F}} X(t,f)\, df\, dt = \sum_{j=1}^{J} \int_{\mathbb{T}} \int_{\mathbb{F}} Y_j(t,f;\theta)\, df\, dt = 1, \tag{3.4}$$

the sum of Eq. (3.3) for all decomposed spectrogram:

$$\sum_{j=1}^{J} \int_{\mathbb{T}} \int_{\mathbb{F}} \hat{X}_j(t,f) \log \frac{\hat{X}_j(t,f)}{Y_j(t,f;\theta)}\, df\, dt, \tag{3.5}$$

must be non-negative by Jensen's inequality. Thus, the smaller this sum, the better the decomposition determined by the parameter θ and the distribution function $Z(j;\, t, f)$. Under conditions of Eq. (3.1) and Eq. (3.2), the source separation is achieved by calculating $Z(j;\, t, f)$ and θ which minimize Eq. (3.5).

Here, we introduce an objective function which is defined as the sum of Eq. (3.5) and a Lagrange multiplier, $\lambda(t, f)$, for the conditions of Eq. (3.1) and Eq. (3.2):

$$\begin{aligned} Q = &\sum_{j=1}^{J} \int_{\mathbb{T}} \int_{\mathbb{F}} \hat{X}_j(t,f) \log \frac{\hat{X}_j(t,f)}{Y_j(t,f;\theta)}\, df\, dt \\ &- \int_{\mathbb{T}} \int_{\mathbb{F}} \lambda(t,f) \left(\sum_{j=1}^{J} Z(j;\, t, f) - 1 \right) df\, dt. \end{aligned} \tag{3.6}$$

Although optimal both θ and $Z(j;\, t, f)$ cannot be calculated analytically since they depend each other, they converge on local optima by alternately optimizing the one with the other fixed. By solving the following simultaneous equations:

$$\frac{dQ}{dZ(j;\, t, f)} = 0 \quad \text{and} \quad \frac{dQ}{d\lambda(t,f)} = 0, \tag{3.7}$$

the optimal $Z(j;\, t, f)$ is derived as follows:

$$Z(j;\, t, f) = \frac{Y_j(t,f;\theta)}{\sum_{j'=1}^{J} Y_{j'}(t,f;\theta)}. \tag{3.8}$$

On the other hand, the optimal θ can be calculated as:

$$\arg\min_{\theta} Q = \arg\min_{\theta} \sum_{j=1}^{J} \int_{\mathbb{T}} \int_{\mathbb{F}} \hat{X}_j(t,f) \log \frac{1}{Y_j(t,f;\theta)}\, df\, dt. \tag{3.9}$$

This is determined by the definition of $Y_j(t, f; \theta)$.

To return the decomposed magnitude spectrogram to the acoustic signal, we reconstruct the phase of the spectrogram in some way and perform an inverse STFT. To reconstruct the phase, following methods can be used:

- use the phase of the complex spectrogram of the original audio mixture, and
- estimate the phase from the magnitude spectrogram [58].

3.3 Harmonic and Inharmonic Integrated Model

In this section, we formulate the spectral function model on the time-frequency domain, $Y_j(t, f; \theta)$. $Y_j(t, f; \theta)$ should be configured so that the model can represent the spectrogram of various musical instrument sounds since the input audio mixture contains various sounds, which consist of only harmonic-structure component, only inharmonic component, and both components.

The harmonic and inharmonic integrated weighted-mixture model, $Y_j(t, f)$, represents a magnitude spectrogram of the j-th musical note. We formulate this integrated model as the weighted sum of a harmonic-structure model, $Y_{\mathrm{H}|j}(t, f)$, and an inharmonic-structure model, $Y_{\mathrm{I}|j}(t, f)$, which represents the harmonic-structure and inharmonic-structure, respectively, with the weight parameters, $\left(w_{\mathrm{H}|j}, w_{\mathrm{I}|j}\right)$:

$$Y_j(t, f) = w_{\mathrm{H}|j} Y_{\mathrm{H}|j}(t, f) + w_{\mathrm{I}|j} Y_{\mathrm{I}|j}(t, f)\,. \tag{3.10}$$

The weight parameter satisfies the following constraints:

$$\forall j = 1, \ldots, J:\ 0 \leq w_{\mathrm{H}|j} \leq 1, \tag{3.11a}$$

$$\forall j = 1, \ldots, J:\ 0 \leq w_{\mathrm{I}|j} \leq 1, \tag{3.11b}$$

and

$$\forall j = 1, \ldots, J:\ w_{\mathrm{H}|j} + w_{\mathrm{I}|j} = 1. \tag{3.11c}$$

The weight parameter represents the relative magnitude of the harmonic and inharmonic components. For example, a sound of a theremin consisting of only harmonic component will make $w_{\mathrm{H}|j} \simeq 1$, a sound of a bass drum consisting of only inharmonic component will make $w_{\mathrm{I}|j} \simeq 1$, and a sound of a pianoforte consisting both harmonic and inharmonic components will make both weight parameters have a certain level.

Then, we formulate the harmonic model, $Y_{\mathrm{H}|j}(t, f)$. By assuming that the fundamental frequency (F0) of the harmonic-structure component is constant from excitation until decay, we represent the F0 as ϕ_j. The cut of the harmonic component at t should be a harmonic structure with the F0 ϕ_j as shown in Figure 3.1(c). In the STFT with a Gaussian

window function, the energy diffusion along the frequency axis can be approximated by a Gaussian function since the Gaussian window is convolved to the spectrogram. We model the frequency component of the harmonic structure as the structured mixture of Gaussian functions:

$$Y_{\mathrm{H}|j}(f) = \sum_{m=1}^{M_{\mathrm{H}}} w_{m|j,\mathrm{H}} Y_{m|j,\mathrm{H}}(f) \tag{3.12a}$$

and

$$Y_{m|j,\mathrm{H}}(f) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(f - m\phi_j)^2}{2\sigma_j^2}\right). \tag{3.12b}$$

$w_{m|j,\mathrm{H}}$ is the weight parameter which represents the relative magnitude of the m-th harmonic component and satisfies the following constraints:

$$\forall j = 1, \ldots, J;\ m = 1, \ldots, M_{\mathrm{H}} :\ 0 \leq w_{m|j,\mathrm{H}} \leq 1 \tag{3.13a}$$

and

$$\forall j = 1, \ldots, J :\ \sum_{m=1}^{M_{\mathrm{H}}} w_{m|j,\mathrm{H}} = 1. \tag{3.13b}$$

The magnitude of each harmonic component continuously increases and decreases with time. Although the magnitudes of the harmonic components change asynchronously, we assume that they change synchronously and model the change of them as $Y_{\mathrm{H}|j}(t)$. Since $Y_{\mathrm{H}|j}(t)$ must satisfy Eq. (3.4) for the separation, we assume that the integration of $Y_{\mathrm{H}|j}(t)$ over $\mathbb{T}$ is bounded and satisfies the following condition:

$$\int_{\mathbb{T}} Y_{\mathrm{H}|j}(t)\, dt = 1. \tag{3.14}$$

To represent the harmonic structure of various instrument sounds, $Y_{\mathrm{H}|j}(t)$ should have a flexible functional form which is compatible with various musical instruments rather than $Y_{\mathrm{H}|j}(t)$ is designed based on a physical acoustic mechanism, i.e., excitation and resonance, of a particular musical instrument. To achieve this, $Y_{\mathrm{H}|j}(t)$ should be continuous, be non-negative over all t, converge to 0 at $t \to \pm\infty$, be elastic, and fit various curves. Additionally, to achieve Eq. (3.9), $Y_{\mathrm{H}|j}(t)$ should be differentiable at all t. We formulate $Y_{\mathrm{H}|j}(t)$ as the following structured mixture of Gaussian functions:

$$Y_{\mathrm{H}|j}(t) = \sum_{l=0}^{L_{\mathrm{H}}-1} w_{l|j,\mathrm{H}} Y_{l|j,\mathrm{H}}(t) \tag{3.15a}$$

and

$$Y_{l|j,\mathrm{H}}(t) = \frac{1}{\sqrt{2\pi}\rho_j} \exp\left(-\frac{(t - \tau_j - l\rho_j)^2}{2\rho_j^2}\right). \tag{3.15b}$$

τ_j is the center of the first Gaussian function and correspond to the estimate of the onset time. $w_{l|j,\mathrm{H}}$ is a relative weight coefficient for the l-th Gaussian function and satisfies the following constraints:

$$\forall j = 1, \ldots, J;\ l = 0, \ldots, L_{\mathrm{H}} - 1 :\ 0 \leq w_{l|j,\mathrm{H}} \leq 1 \tag{3.16a}$$

and

$$\forall j = 1, \ldots, J :\ \sum_{l=0}^{L_{\mathrm{H}}-1} w_{l|j,\mathrm{H}} = 1. \tag{3.16b}$$

By changing the coefficients, the curve of $Y_{\mathrm{H}|j}(t)$ varies. This function is suitable for the energy diffusion along the time axis in the STFT with a Gaussian window since its basis function is a Gaussian function. In this function, L_{H} Gaussian functions share the standard deviation, ρ_j, and are equally-spaced with ρ_j. This constraint makes Gaussian functions closely-spaced and produces a property that $Y_{\mathrm{H}|j}(t)$ expands and contracts depending on ρ_j. Thus, the harmonic structure model, $Y_{\mathrm{H}|j}(t, f)$ is defined as the product of $Y_{\mathrm{H}|j}(t)$ and $Y_{\mathrm{H}|j}(f)$:

$$Y_{\mathrm{H}|j}(t, f) = Y_{\mathrm{H}|j}(t)\, Y_{\mathrm{H}|j}(f)\,. \tag{3.17}$$

Then, we formulate the inharmonic model, $Y_{\mathrm{I}|j}(t, f)$. We assume that the magnitude distribution of the inharmonic component is stable over time as the harmonic structure model. The cut of the inharmonic component at t should be an unsharp distribution as shown in Figure 3.1(c). To represent a distribution like this, it is desirable to describe the relative magnitude for each frequency band. The inharmonic component tends to diffuse according to the frequency. We split $\mathbb{F}$ into M_{I} overlapping frequency bands which are equally-spaced on a logarithmic frequency scale and model the frequency structure of the inharmonic component as the weighted mixture of distribution functions which have fixed position and diffusion:

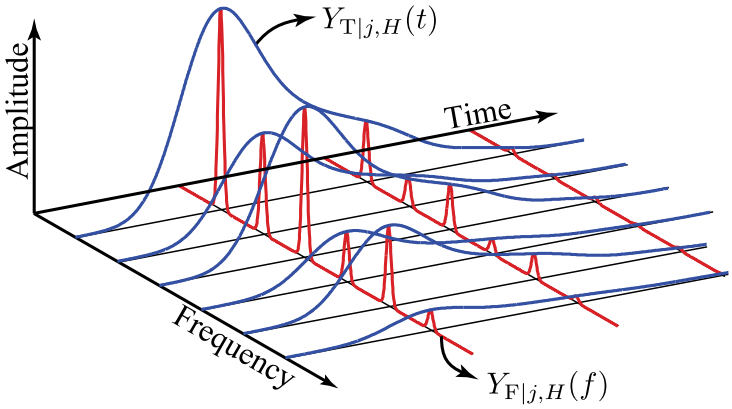
$$Y_{\mathrm{I}|j}(f) = \sum_{m=1}^{M_{\mathrm{I}}} w_{m|j,\mathrm{I}} Y_{m|\mathrm{I}}(f) \tag{3.18a}$$

and

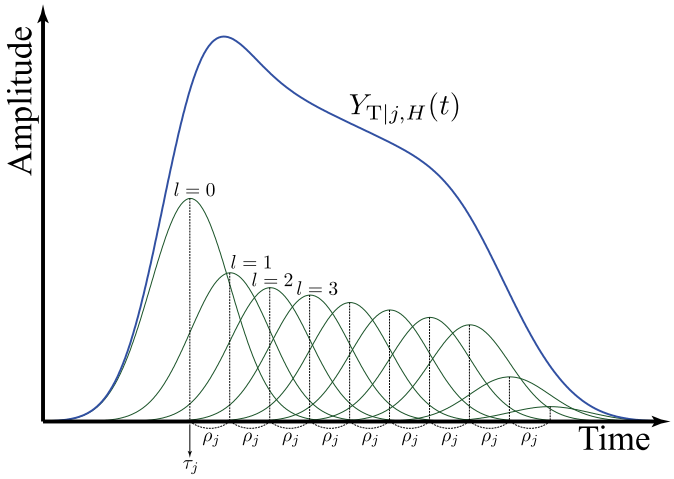
$$Y_{m|\mathrm{I}}(f) = \frac{\varphi}{\sqrt{2\pi}\,(f + \varsigma)} \exp\left(-\frac{\left(\varphi \log\left(f/\varsigma + 1\right) - m\right)^2}{2}\right). \tag{3.18b}$$

$w_{m|j,\mathrm{I}}$ is the weight parameter which represents the relative magnitude of the m-th frequency band and satisfies the following constraints:

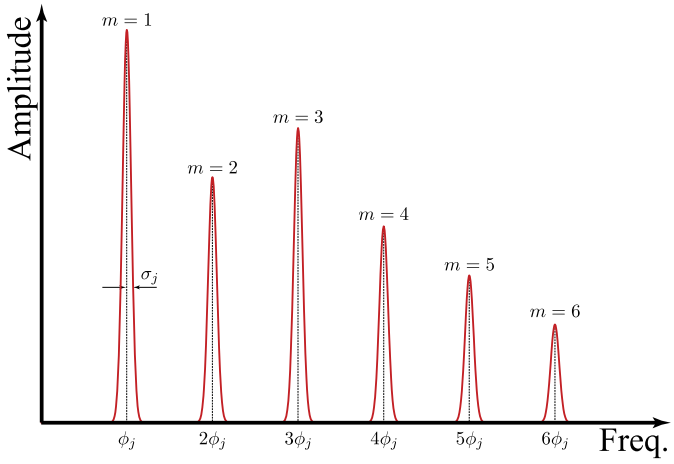
$$\forall j = 1, \ldots, J;\ m = 1, \ldots, M_{\mathrm{I}} :\ 0 \leq w_{m|j,\mathrm{I}} \leq 1 \tag{3.19a}$$



(a) Overview of harmonic tone model.



(b) Temporal structure of harmonic tone model.



(c) Frequency structure of harmonic tone model.

Figure 3.1: Overall, temporal and frequency structures of the harmonic tone model. This model consists of a two-dimensional Gaussian Mixture Model, and it is factorized into a pair of one-dimensional GMMs.

and

$$\forall j = 1, \ldots, J : \sum_{m=1}^{M_{\mathrm{I}}} w_{m|j,\mathrm{I}} = 1. \tag{3.19b}$$

φ and ς are constants which determine the frequency scale and position of the frequency bands. Note that $Y_{m|\mathrm{I}}(f)$ does not have note-dependent parameters. As shown in Figure 3.2, $Y_{m|\mathrm{I}}(f)$ is derived from the Gaussian functions:

$$\sum_{m=1}^{M_{\mathrm{I}}} w_{m|j,\mathrm{I}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(f-m)^2}{2}\right), \tag{3.20}$$

on the logarithmic frequency scale which is determined by φ and ς,

$$f' = \varphi \log\left(\frac{f}{\varsigma} + 1\right) \tag{3.21}$$

by changing the scale to the linear frequency scale from $f' = \varphi \log(f/\varsigma + 1)$ to f:

$$\begin{aligned} Y_{\mathrm{I}|j}(f) &= \frac{df'}{df} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(f'-m)^2}{2}\right) \\ &= \frac{\varphi}{\sqrt{2\pi}\,(f+\varsigma)} \exp\left(-\frac{(\varphi \log(f/\varsigma+1) - m)^2}{2}\right). \end{aligned} \tag{3.22}$$

We assume that the magnitudes of each frequency band of the inharmonic component change synchronously as the harmonic structure and model the change of them as $Y_{\mathrm{I}|j}(t)$. The functional form of $Y_{\mathrm{I}|j}(t)$ is defined as $Y_{\mathrm{H}|j}(t)$:

$$Y_{\mathrm{I}|j}(t) = \sum_{l=0}^{L_{\mathrm{I}}-1} w_{l|j,\mathrm{I}} Y_{l|j,\mathrm{I}}(t) \tag{3.23a}$$

and

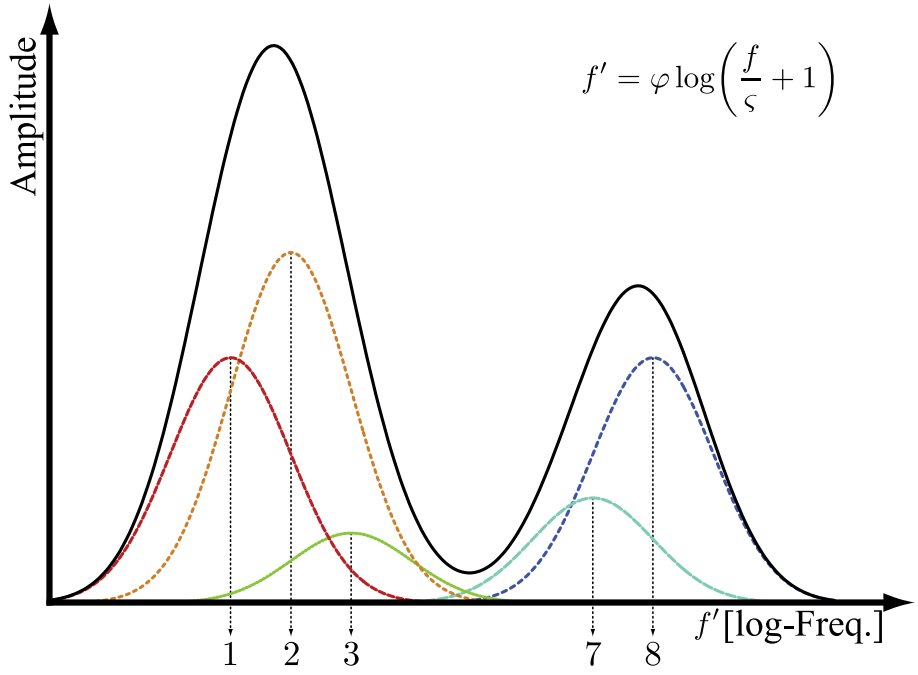
$$Y_{l|j,\mathrm{I}}(t) = \frac{1}{\sqrt{2\pi}\varrho_j} \exp\left(-\frac{(t-\tau_j - l\varrho_j)^2}{2\varrho_j^2}\right). \tag{3.23b}$$

τ_j is the shared parameter to the harmonic model to achieve the harmonic and inharmonic models of a corresponding note start at the same time. $w_{l|j,\mathrm{I}}$ is a relative weight coefficient parameter corresponding to $w_{l|j,\mathrm{H}}$ of the harmonic model and satisfies the following constraints:

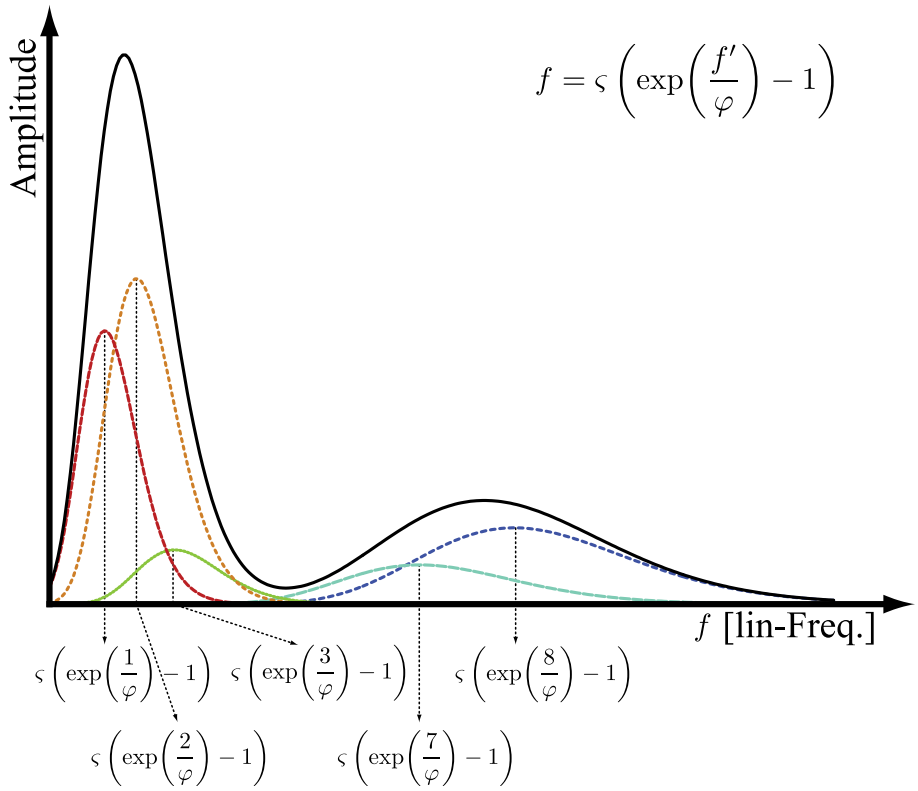
$$\forall j = 1, \ldots, J;\ l = 0, \ldots, L_{\mathrm{I}} - 1 :\ 0 \le w_{l|j,\mathrm{I}} \le 1 \tag{3.24a}$$

and

$$\forall j = 1, \ldots, J : \sum_{l=0}^{L_{\mathrm{I}}-1} w_{l|j,\mathrm{I}} = 1. \tag{3.24b}$$



(a) Equally-spaced Gaussian kernels along the log-scale frequency.



(b) Gaussian kernels obtained by changing the variables in (a).

Figure 3.2: Frequency structure of inharmonic model

Thus, the inharmonic model, $Y_{\mathrm{I}|j}(t,f)$ is defined as the product of $Y_{\mathrm{I}|j}(t)$ and $Y_{\mathrm{I}|j}(f)$:

$$Y_{\mathrm{I}|j}(t,f) = Y_{\mathrm{I}|j}(t)\, Y_{\mathrm{I}|j}(f)\,. \tag{3.25}$$

We model the observed audio mixture as the weighted mixture of J integrated models:

$$Y(t,f) = \sum_{j=1}^{J} w_j Y_j(t,f). \tag{3.26}$$

w_j is the relative magnitude parameter of the j-th musical note and satisfies the following constraints:

$$\forall j = 1,\ldots,J:\ 0 \le w_j \le 1 \tag{3.27a}$$

and

$$\sum_{j=1}^{J} w_j = 1. \tag{3.27b}$$

3.4 Model Parameter Estimation

$Y_j(t,f)$ is defined as a linear combinations of $Y_{\mathrm{H}|j}(t,f)$ and $Y_{\mathrm{I}|j}(t,f)$. $Y_{\mathrm{H}|j}(t,f)$ and $Y_{\mathrm{I}|j}(t,f)$ are also defined as linear combinations of 2-dimensional basis functions. By decomposing the decomposed magnitude spectrogram of the j-th note, $\hat{X}_j(t,f)$, into sub-spectrograms corresponding to harmonic and inharmonic components and each decomposed spectrograms into the basis functions, the optimal parameters of Eq. (3.9) can be calculated analytically. Weighted basis functions of the harmonic and inharmonic models are represented as: The weighted basis functions for the integrated model is described as:

$$Y_{j,\mathrm{H},l,m}(t,f) = w_j\, w_{\mathrm{H}|j}\, w_{l|j,\mathrm{H}}\, w_{m|j,\mathrm{H}}\, Y_{l|j,\mathrm{H}}(t)\, Y_{m|j,\mathrm{H}}(f) \tag{3.28a}$$

and

$$Y_{j,\mathrm{I},l,m}(t,f) = w_j\, w_{\mathrm{I}|j}\, w_{l|j,\mathrm{I}}\, w_{m|j,\mathrm{I}} Y_{l|j,\mathrm{I}}(t)\, Y_{m|\mathrm{I}}(f)\,. \tag{3.28b}$$

Here, we introduce new spectrogram distribution functions, $Z(\mathrm{H};\, j,t,f)$, $Z(\mathrm{I};\, j,t,f)$, $Z(l,m;\, j,\mathrm{H},t,f)$ and $Z(l,m;\, j,\mathrm{I},t,f)$. $Z(\mathrm{H};\, j,t,f)$ and $Z(\mathrm{I};\, j,t,f)$ decompose $\hat{X}_j(t,f)$ into the harmonic and inharmonic components and satisfy the following constraints:

$$\forall j,t,f:\ 0 \le Z(\mathrm{H};\, j,t,f) \le 1, \tag{3.29a}$$

$$\forall j,t,f:\ 0 \le Z(\mathrm{I};\, j,t,f) \le 1, \tag{3.29b}$$

and

$$\forall j,t,f:\ Z(\mathrm{H};\,j,t,f)+Z(\mathrm{I};\,j,t,f)=1. \tag{3.29c}$$

The decomposed spectrograms by $Z(\mathrm{H};\,j,t,f)$ and $Z(\mathrm{I};\,j,t,f)$ are denoted as:

$$\hat{X}_{j,\mathrm{H}}(t,f)=Z(\mathrm{H};\,j,t,f)\,\hat{X}_j(t,f) \tag{3.30a}$$

and

$$\hat{X}_{j,\mathrm{I}}(t,f)=Z(\mathrm{I};\,j,t,f)\,\hat{X}_j(t,f)\,. \tag{3.30b}$$

$Z(l,m;\,j,\mathrm{H},t,f)$ and $Z(l,m;\,j,\mathrm{I},t,f)$ decompose $\hat{X}_{j,\mathrm{H}}(t,f)$ and $\hat{X}_{j,\mathrm{I}}(t,f)$ into the basis functions of the harmonic and inharmonic models, respectively, and satisfies the following constraints:

$$\forall j,t,f,l,m:\ 0\le Z(l,m;\,j,\mathrm{H},t,f)\le 1, \tag{3.31a}$$

$$\forall j,t,f:\ \sum_{l=0}^{L_\mathrm{H}-1}\sum_{m=1}^{M_\mathrm{H}} Z(l,m;\,j,\mathrm{H},t,f)=1, \tag{3.31b}$$

$$\forall j,t,f,l,m:\ 0\le Z(l,m;\,j,\mathrm{I},t,f)\le 1, \tag{3.31c}$$

and

$$\forall j,t,f:\ \sum_{l=0}^{L_\mathrm{I}-1}\sum_{m=1}^{M_\mathrm{I}} Z(l,m;\,j,\mathrm{I},t,f)=1. \tag{3.31d}$$

The decomposed spectrograms by $Z(l,m;\,j,\mathrm{H},t,f)$ and $Z(l,m;\,j,\mathrm{I},t,f)$ are denoted as:

$$\hat{X}_{j,\mathrm{H},l,m}(t,f)=Z(l,m;\,j,\mathrm{H},t,f)\,\hat{X}_{j,\mathrm{H}}(t,f) \tag{3.32a}$$

and

$$\hat{X}_{j,\mathrm{I},l,m}(t,f)=Z(l,m;\,j,\mathrm{I},t,f)\,\hat{X}_{j,\mathrm{I}}(t,f)\,. \tag{3.32b}$$

For the same reason as Eq. (3.5), the sum of pseudo-distances, described as follows, must be non-negative by Jensen's inequality:

$$\begin{aligned} Q'=\sum_{j=1}^{J}\Bigg(&\sum_{l=0}^{L_\mathrm{H}-1}\sum_{m=1}^{M_\mathrm{H}}\int_{\mathbb{T}}\int_{\mathbb{F}}\hat{X}_{j,\mathrm{H},l,m}(t,f)\log\frac{\hat{X}_{j,\mathrm{H},l,m}(t,f)}{Y_{j,\mathrm{H},l,m}(t,f)}\,df\,dt\\ &+\sum_{l=0}^{L_\mathrm{I}-1}\sum_{m=1}^{M_\mathrm{I}}\int_{\mathbb{T}}\int_{\mathbb{F}}\hat{X}_{j,\mathrm{I},l,m}(t,f)\log\frac{\hat{X}_{j,\mathrm{I},l,m}(t,f)}{Y_{j,\mathrm{I},l,m}(t,f)}\,df\,dt\Bigg)\,. \end{aligned} \tag{3.33}$$

Since Q' takes 0 and Q' is equal to Q if and only of

$$\forall j,l,m,t,f:\ \hat{X}_{j,\mathrm{H},l,m}(t,f)=Y_{j,\mathrm{H},l,m}(t,f) \tag{3.34a}$$

and

$$\forall j, l, m, t, f : \hat{X}_{j,\mathrm{I},l,m}(t, f) = Y_{j,\mathrm{I},l,m}(t, f)\,, \tag{3.34b}$$

minimize of Q is also achieved by minimizing Q'. Derivation of the distribution functions, $Z(\mathrm{H};\, j, t, f)$, $Z(\mathrm{I};\, j, t, f)$, $Z(l, m;\, j, \mathrm{H}, t, f)$ and $Z(l, m;\, j, \mathrm{I}, t, f)$, is omitted since it is the same as derivation of $Z(j;\, t, f)$. To estimate the model parameters, Q' should be minimized with fixed distribution functions.

We represent integration and summation over the variables and parameters for the spectrograms by omitting the variables and parameters of them, e.g.,

$$\hat{X}_{j,\mathrm{H},l}(t) = \sum_{m=1}^{M_{\mathrm{H}}} \int_{\mathbb{F}} \hat{X}_{j,\mathrm{H},l,m}(t, f)\, df \tag{3.35a}$$

and

$$\hat{X}_j = \int_{\mathbb{T}} \int_{\mathbb{F}} \hat{X}_j(t, f)\, df\, dt. \tag{3.35b}$$

Update equations of the parameters are described as follows.

$$w_j = \hat{X}_j / X \tag{3.36}$$

$$w_{\mathrm{H}|j} = \hat{X}_{j,\mathrm{H}} / \hat{X}_j \tag{3.37}$$

$$w_{\mathrm{I}|j} = \hat{X}_{j,\mathrm{I}} / \hat{X}_j \tag{3.38}$$

$$w_{l|j,\mathrm{H}} = \hat{X}_{j,\mathrm{H},l} / \hat{X}_{j,\mathrm{H}} \tag{3.39}$$

$$w_{m|j,\mathrm{H}} = \hat{X}_{j,\mathrm{H},m} / \hat{X}_{j,\mathrm{H}} \tag{3.40}$$

$$w_{l|j,\mathrm{I}} = \hat{X}_{j,\mathrm{I},l} / \hat{X}_{j,\mathrm{I}} \tag{3.41}$$

$$w_{m|j,\mathrm{I}} = \hat{X}_{j,\mathrm{I},m} / \hat{X}_{j,\mathrm{I}} \tag{3.42}$$

$$\tau_j = \left(\sum_{l=0}^{L_{\mathrm{I}}-1} \int_{\mathbb{T}} (t - l\rho_j)\, \hat{X}_{j,\mathrm{H},l}(t)\, dt + \sum_{l=0}^{L_{\mathrm{I}}-1} \int_{\mathbb{T}} (t - l\varrho_j)\, \hat{X}_{j,\mathrm{I},l}(t)\, dt \right) \Big/ \hat{X}_j \tag{3.43}$$

$$\rho_j = \left(-a_{j,\mathrm{H}} + \sqrt{a_{j,\mathrm{H}}^2 + 4 b_{j,\mathrm{H}} \hat{X}_{j,\mathrm{H}}} \right) \Big/ 2\hat{X}_{j,\mathrm{H}} \tag{3.44a}$$

$$a_{j,\mathrm{H}} = \sum_{l=0}^{L_{\mathrm{I}}-1} \int_{\mathbb{T}} l\, (t - \tau_j)\, \hat{X}_{j,\mathrm{H},l}(t)\, dt \tag{3.44b}$$

$$b_{j,\mathrm{H}} = \sum_{l=0}^{L_{\mathrm{I}}-1} \int_{\mathbb{T}} (t - \tau_j)^2\, \hat{X}_{j,\mathrm{H},l}(t)\, dt \tag{3.44c}$$

$$\varrho_j = \left(-a_{j,\mathrm{I}} + \sqrt{a_{j,\mathrm{I}}^2 + 4 b_{j,\mathrm{I}} \hat{X}_{j,\mathrm{I}}} \right) \Big/ 2\hat{X}_{j,\mathrm{I}} \tag{3.45a}$$

$$a_{j,\mathrm{I}} = \sum_{l=0}^{L_\mathrm{I}-1} \int_{\mathbb{T}} l\left(t - \tau_j\right) \hat{X}_{j,\mathrm{I},l}(t)\, dt \tag{3.45b}$$

$$b_{j,\mathrm{I}} = \sum_{l=0}^{L_\mathrm{I}-1} \int_{\mathbb{T}} \left(t - \tau_j\right)^2 \hat{X}_{j,\mathrm{I},l}(t)\, dt \tag{3.45c}$$

$$\phi_j = \left(\sum_{m=1}^{M_\mathrm{H}} \int_{\mathbb{F}} mf \hat{X}_{j,\mathrm{H},m}(f)\, df \right) \Big/ \left(\sum_{m=1}^{M_\mathrm{H}} \int_{\mathbb{F}} m^2 \hat{X}_{j,\mathrm{H},m}(f)\, df \right) \tag{3.46}$$

$$\sigma_j = \sqrt{\sum_{m=1}^{M_\mathrm{H}} \int_{\mathbb{F}} \left(f - m\phi_j\right)^2 \hat{X}_{j,\mathrm{H},m}(f)\, df \Big/ \hat{X}_{j,\mathrm{H}}}. \tag{3.47}$$

3.5 Perspective as ML and MAP Estimation

Let input magnitude spectrogram, $X(t,f)$, be the observed probabilistic density function and model of the sound mixture, $Y(t,f;\theta)$, be the conditional probabilistic density function of the model parameters (likelihood function of the model). Standing on this perspective, the above parameter estimation has the same formulation to a maximum likelihood (ML) estimation which maximizes the logarithmic likelihood:

$$\theta_\mathrm{ML} = \arg\min_\theta Q = \arg\max_\theta \left\langle \log Y(t,f;\theta) \right\rangle_{X(t,f)}. \tag{3.48}$$

$\langle\cdot\rangle_{f(x)}$ means the expected value. Therefore, this iterative calculation of the distribution function and model parameters is equivalent to an Expectation-Maximization (EM) algorithm.

This fact leads to following extension: the estimation can be extended to a maximum *A Posteriori* (MAP) estimation by assuming the prior information for the parameters. The MAP estimation maximizes expected value of the logarithmic posteriori probability:

$$\theta_\mathrm{MAP} = \arg\max_\theta \left\langle \log Y(t,f;\theta) + \log p(\theta) \right\rangle_{X(t,f)}, \tag{3.49}$$

with statistical experience for the model parameters, represented as $p(\theta)$. The prior distribution, $p(\theta)$, performs as a penalty which prevents deviation of the parameters from the sane range. In Chapter 4, we discuss a method to estimate the model parameters on the basis of a MAP estimation by using $p(\theta)$ based on a musical score corresponding the audio mixture.

Table 3.1: Experimental conditions

Frequency analysis	
sampling rate	44.1 kHz
STFT window	4096 points Gaussian
STFT shift	441 points
Parameters	
L_{H}	30
M_{H}	30
L_{I}	30
M_{I}	30
φ	440.0
ς	1.134

3.6 Experimental Evaluation

We conducted two experimental evaluations to evaluate our new approach and determine the effectiveness of the integrated model. We compared modeling error (Eq. (3.33)) and source separation performances under the following three conditions:

1. using the integrated model (proposed method),
2. using only the harmonic structure model, and
3. using only the inharmonic model.

Performances of the source separation are calculated by a signal-to-noise ratio (SNR). The SNR of the j-th note is defined as:

$$\mathrm{SNR}_j = \frac{1}{|\mathbb{T}|}\int_{\mathbb{T}} 10\log_{10}\int_{\mathbb{F}} \frac{\dot{X}_j(t,f)^2}{\left(\dot{X}_j(t,f)-\hat{X}_j(t,f)\right)^2}\,df\,dt \tag{3.50}$$

where $\dot{X}_j(t,f)$ means the reference magnitude spectrogram of the j-th note, i.e., spectrogram before mixing-down.

We excerpted 5 musical instruments: pianoforte, guitar, violin, trumpet, and drums, from the RWC Music Database: Musical Instrument Sound (RWC-MDB-I-2001) [59]. In the first experiment, we modeled isolated musical instrument sounds by using only harmonic, only inharmonic, and integrated models, and calculated modeling errors when the parameters converged. In the second experiment, we first created 34533 instrument sound pairs excerpted from the database and mixed the sounds for each pair. This means

Table 3.2: Average modeling error of each instrument

Instrument	Integrated	Harmonic	Inharmonic
Pianoforte	1.03	9.18	1.17
Guitar	1.36	8.15	1.31
Violin	1.02	9.23	1.26
Trumpet	1.14	10.9	1.26
Drums	0.282	4.38	0.429

Table 3.3: Average SNR [dB] of each instrument

Instrument	Integrated	Harmonic	Inharmonic
Pianoforte	18.1	15.8	10.1
Guitar	20.6	20.0	14.7
Violin	37.1	36.7	26.9
Trumpet	42.3	41.7	31.5
Drums	20.0	11.6	14.4

that the experiment was conducted with $J = 2$. The mixtures were separated by a method using template sounds (described in Chapter 4). Template sounds were generated by a MIDI tone generator, MU2000. We evaluated the SNR for each separated magnitude spectrogram. The details of the experimental conditions are listed in Table 3.1.

3.6.1 Experimental Result

Instrument-wise average modeling errors of the isolated instrument sounds are shown in Table 3.2 for each spectral model. The errors by using the integrated models are less than the ones by using the harmonic and inharmonic models for each instrument. This result shows the effectiveness of the integrated models in spectral modeling of musical instrument sounds.

Instrument-wise average SNRs of the separated spectrograms are shown in Table 3.2 for each spectral model. The SNRs by using the integrated models are larger than the ones by using the harmonic and inharmonic models for each instrument. This result shows the effectiveness of the integrated models in separating instrument sounds.

Excerpting for the drums, the modeling errors by using the harmonic models are larger than the ones by using the inharmonic models. The errors by using the harmonic models tend to become large since diffusion of magnitude between frequency bins of harmonic structure and other parts. However, the SNRs by using the harmonic models

are smaller than the ones by using the inharmonic ones. The SNRs of harmonic instrument sounds tend to become large since the harmonic models make almost binary distribution functions.

3.7 Summary

We summarize this chapter as follows:

- We proposed a sound source separation method based on an integrated weighted-mixture model that represents both harmonic and inharmonic sounds. We defined the integrated model as the sum of the harmonic-structure and inharmonic models. On the basis of the iterative spectrogram separation and parameter estimation algorithm, models were adapted to the audio mixture and the mixture was separated.
- We reported our experimental results that showed effectiveness of the integrated model than the harmonic and inharmonic models.

Chapter 4

Score-informed Source Separation

This chapter describes a method to separating musical audio signals using the integrated harmonic and inharmonic models and prior information based on the musical score corresponding to the audio.

4.1 Instrument Sound Recognition in Polyphonic Musical Audio

To achieve the source separation which is the goal of this study, all musical instrument sound must be recognized in the musical audio signal. Although we defined the integrated model in Chapter 3, the separation is still not achieved by the following reasons:

- Appropriate positions of the models on the time-frequency domain, i.e., model parameters corresponding to onset time and F0, are unknown. It is hard to estimate the onset time and F0 of all instrument sound from randomly initialized model parameters although the diffusion of these parameters can be adjusted to some extent.

- The spectrogram of each instrument sound and the integrated model does not always correspond one-for-one without value ranges or norms of the parameters because the integrated model has large degrees of freedom of the parameters. Especially, it is hard to decompose a spectrogram in which multiple instrument sounds are performed in unison without any information of sound sources.

However, automatic instrument sound recognition which estimates pitch, onset time, duration, and instrument of each sound, from musical audio mixture is a hard task. Although various multipitch estimation methods [14, 28, 48, 49] have been proposed, pitch

recognition ratio for four polyphonic sounds is about 60% – 80% by these methods and recognition performance for musical CD recordings is insufficient.

4.2 Musical Score as Prior Information

We assume a musical score as a standard MIDI file (SMF) which contains following information:

- Onset time: time (tick) of the note-on message.
- Pitch: note number of the note-on message (only for pitched sounds.)
- Duration: time duration from note-on to note-off messages.
- Instrument: program number in program change message for pitched sounds or note number in the note-on message for unpitched sounds.

Based on these kinds of information, we set the model parameters as follows:

- Set τ_j from the onset time.
- Set ϕ_j from the pitch. The note number f' is converted to the pitch f by the following equation:
$$f = 440 \times 2^{(f'-69)/12}. \tag{4.1}$$
- Set ρ_j and ϱ_j from the duration.
- Set $\left(w_{\mathrm{H}|j}, w_{\mathrm{I}|j}\right)$, $w_{l|j,\mathrm{H}}$, $w_{m|j,\mathrm{H}}$, $w_{l|j,\mathrm{I}}$, and $w_{m|j,\mathrm{I}}$ by the instrument.

In the above processing, there are several parameter setting methods by the instrument. Following sections describe two parameter setting methods: a method based on template sounds and a method based on prior distributions of the model parameters.

4.3 Template Sounds

By playing back each pair of note-on and note-off messages of the SMF on a MIDI sound module, we prepared sampled sounds for each note. We call this template sounds and used this as prior information (and initial values) in the model parameter estimation.

First, we adapt each integrated model to the corresponding template sound. Let $\xi_j(t,f)$ be the magnitude spectrogram of the template sound of the j-th musical note, $\hat{\xi}_{j,\mathrm{H}}(t,f)$, $\hat{\xi}_{j,\mathrm{I}}(t,f)$, $\hat{\xi}_{j,\mathrm{H},l,m}(t,f)$, and $\hat{\xi}_{j,\mathrm{I},l,m}(t,f)$ be the decomposed template spectrogram which are defined as:

$$\hat{\xi}_{j,\mathrm{H}}(t,f) = Z(\mathrm{H};\, j,t,f)\,\xi_j(t,f)\,, \tag{4.2a}$$

$$\hat{\xi}_{j,\mathrm{I}}(t,f) = Z(\mathrm{I};\, j,t,f)\,\xi_j(t,f)\,, \tag{4.2b}$$

$$\hat{\xi}_{j,\mathrm{H},l,m}(t,f) = Z(l,m;\, j,\mathrm{H},t,f)\,\hat{\xi}_{j,\mathrm{H}}(t,f)\,, \tag{4.2c}$$

and

$$\hat{\xi}_{j,\mathrm{I},l,m}(t,f) = Z(l,m;\, j,\mathrm{I},t,f)\,\hat{\xi}_{j,\mathrm{I}}(t,f)\,. \tag{4.2d}$$

This adaptation is achieved by minimizing the sum of pseudo-distance between the decomposed magnitude spectrogram of the template sound and the basis functions of the integrated model:

$$\begin{aligned} Q_j^{(T)} = &\sum_{l=0}^{L_\mathrm{H}-1}\sum_{m=1}^{M_\mathrm{H}}\int_{\mathbb{T}}\int_{\mathbb{F}}\hat{\xi}_{j,\mathrm{H},l,m}(t,f)\log\frac{\hat{\xi}_{j,\mathrm{H},l,m}(t,f)}{Y_{j,\mathrm{H},l,m}(t,f)}\,df\,dt \\ &+\sum_{l=0}^{L_\mathrm{I}-1}\sum_{m=1}^{M_\mathrm{I}}\int_{\mathbb{T}}\int_{\mathbb{F}}\hat{\xi}_{j,\mathrm{I},l,m}(t,f)\log\frac{\hat{\xi}_{j,\mathrm{I},l,m}(t,f)}{Y_{j,\mathrm{I},l,m}(t,f)}\,df\,dt. \end{aligned} \tag{4.3}$$

Note that the optimal distribution functions and parameters derived from Eq. (4.3) are the same as the ones derived from Eq. (3.33) except for the changes of $\hat{X}_{j,\mathrm{H},l,m}(t,f)$ for $\hat{\xi}_{j,\mathrm{H},l,m}(t,f)$ and $\hat{X}_{j,\mathrm{I},l,m}(t,f)$ for $\hat{\xi}_{j,\mathrm{I},l,m}(t,f)$. By iterating alternately the source separation and parameter estimation, the weight parameters are initialized by the template sounds and the onset time, pitch, duration parameters are adapted implicitly.

After adaptation to the template sounds, we perform the source separation and parameter estimation for the audio mixture $X(t,f)$ based on the initial parameters. Parameter convergence to undesirable local optima can be avoided by separating and estimating iteratively based on both audio mixture and template sounds. To achieve this, let α be a weight parameter which satisfies $0 \le \alpha \le 1$ and we minimize the following weighted sum:

$$\alpha Q' + (1-\alpha)\sum_{j=1}^{J} Q_j^{(T)}. \tag{4.4}$$

The parameter update equations are derived by changing the input spectrograms, e.g., $\hat{X}_j(t,f)$, to the sum of the input and template spectrograms, e.g., $\alpha\hat{X}_j(t,f) + (1-\alpha)\,\xi_j(t,f)$.

The template sounds are generated for each *musical note* and affect the model of the corresponding musical note. Therefore, the template sounds perform the parameter estimation with timbre constraint for each *musical note*.

4.4 Prior Distribution of Model Parameters

Each template sound is an embodiment of a single musical instrument individual and a single instrument performance style. Instrument individuals and performance styles of the template sounds and each sound in the audio mixture are not always the same or similar. This difference makes timbre diffusion and degrades the quality of source separation. Moreover, template sounds are unable to perform as the *instrument-wise* timbre constraint which is satisfied by the model parameters.

This section describes a parameter estimation method which uses distributions of the model parameters for each instrument trained in advance as prior information. By an instrument-wise timbre constraint that deal with the timbre diffusion, avoiding parameter convergence to undesirable local optima, improving the quality of source separation, and implementing the identity of instruments can be achieved.

We set prior distributions to the following parameters: the relative magnitude of harmonic and inharmonic components, $\big(w_{\mathrm{H}|j}, w_{\mathrm{I}|j}\big)$, the relative magnitude of the harmonics, $w_{m|j,\mathrm{H}}$, and the relative magnitude of the frequency bands for the inharmonic component, $w_{m|j,\mathrm{I}}$. Other parameters are free from prior distributions because of the following reasons: $w_{l|j,\mathrm{H}}$ and $w_{l|j,\mathrm{I}}$, which determine sustain and decay of magnitude, have unfixed relativity to the time scale; τ_j, ϕ_j, ρ_j and ϱ_j do not affect timbre; and σ_j depends on the property of the STFT. In other words, we assume that the temporal structures, $w_{l|j,\mathrm{H}}$ and $w_{l|j,\mathrm{I}}$, are not instrument-specific profiles by using the prior distributions.

Since $\big(w_{\mathrm{H}|j}, w_{\mathrm{I}|j}\big)$, $w_{m|j,\mathrm{H}}$, and $w_{m|j,\mathrm{I}}$ are the multinomial-distributed parameters, we use a beta distribution, M_{H} and M_{I}-dimensional Dirichlet distributions for each parameter:

$$p\big(w_{\mathrm{H}|j}, w_{\mathrm{I}|j}\big) \propto w_{\mathrm{H}|j}^{\tilde{\omega}_{\mathrm{H}|k_j}-1} w_{\mathrm{I}|j}^{\tilde{\omega}_{\mathrm{I}|k_j}-1}, \tag{4.5}$$

$$p\big(w_{1|j,\mathrm{H}}, \ldots, w_{M_{\mathrm{H}}|j,\mathrm{H}}\big) \propto \prod_{m=1}^{M_{\mathrm{H}}} w_{m|j,\mathrm{H}}^{\tilde{\omega}_{m|k_j,\mathrm{H}}-1}, \tag{4.6}$$

and

$$p\big(w_{1|j,\mathrm{I}}, \ldots, w_{M_{\mathrm{I}}|j,\mathrm{I}}\big) \propto \prod_{m=1}^{M_{\mathrm{I}}} w_{m|j,\mathrm{I}}^{\tilde{\omega}_{m|k_j,\mathrm{I}}-1}. \tag{4.7}$$

$\tilde{\omega}_{\mathrm{H}|k_j}$, $\tilde{\omega}_{\mathrm{I}|k_j}$, $\tilde{\omega}_{m|k_j,\mathrm{H}}$, and $\tilde{\omega}_{m|k_j,\mathrm{I}}$ are parameters of the prior distributions and we constrain them to be more than 1. k_j means the index of the musical instrument which performs the j-th note. The parameters of these prior distributions are estimated on the basis of maximum likelihood [60] by using adopted models to the independent instrument sounds excerpted from musical instrument sound database.

As we described in Chapter 3, the parameter estimation based on minimizing the objective function Q' can be interpreted as a maximum likelihood estimation. A parameter estimation based on minimizing the sum of Q' and the logarithmic prior probability of the parameters is described as following:

$$\begin{aligned} Q_j^{(P)} =& -\log p\big(w_{\mathrm{H}|j}, w_{\mathrm{I}|j}\big) - \log p\big(w_{1|j,\mathrm{H}}, \ldots, w_{M_{\mathrm{H}}|j,\mathrm{H}}\big) - \log p\big(w_{1|j,\mathrm{I}}, \ldots, w_{M_{\mathrm{I}}|j,\mathrm{I}}\big) \\ =& -\big(\tilde{\omega}_{\mathrm{H}|k_j} - 1\big)\log w_{\mathrm{H}|j} - \big(\tilde{\omega}_{\mathrm{I}|k_j} - 1\big)\log w_{\mathrm{I}|j} \\ & - \sum_{m=1}^{M_{\mathrm{H}}} \big(\tilde{\omega}_{m|k_j,\mathrm{H}} - 1\big)\log w_{m|j,\mathrm{H}} - \sum_{m=1}^{M_{\mathrm{I}}} \big(\tilde{\omega}_{m|k_j,\mathrm{I}} - 1\big)\log w_{m|j,\mathrm{I}}, \end{aligned} \tag{4.8}$$

is equivalent to the MAP estimation that maximizes the logarithmic posterior probability.

Parameter update equations based on this MAP estimation are described as follows.

$$w_{\mathrm{H}|j} = \Big(\hat{X}_{j,\mathrm{H}} + \tilde{\omega}_{\mathrm{H}|k_j}\Big) \Big/ \Big(\hat{X}_j + \tilde{\omega}_{\mathrm{H}|k_j} + \tilde{\omega}_{\mathrm{I}|k_j}\Big) \tag{4.9}$$

$$w_{\mathrm{I}|j} = \Big(\hat{X}_{j,\mathrm{I}} + \tilde{\omega}_{\mathrm{I}|k_j}\Big) \Big/ \Big(\hat{X}_j + \tilde{\omega}_{\mathrm{H}|k_j} + \tilde{\omega}_{\mathrm{I}|k_j}\Big) \tag{4.10}$$

$$w_{m|j,\mathrm{H}} = \Big(\hat{X}_{j,\mathrm{H},m} + \tilde{\omega}_{m|k_j,\mathrm{H}}\Big) \Big/ \left(\hat{X}_{j,\mathrm{H}} + \sum_{m=1}^{M_{\mathrm{H}}} \tilde{\omega}_{m|j,\mathrm{H}}\right) \tag{4.11}$$

$$w_{m|j,\mathrm{I}} = \Big(\hat{X}_{j,\mathrm{I},m} + \tilde{\omega}_{m|k_j,\mathrm{I}}\Big) \Big/ \left(\hat{X}_{j,\mathrm{I}} + \sum_{m=1}^{M_{\mathrm{I}}} \tilde{\omega}_{m|j,\mathrm{I}}\right) \tag{4.12}$$

4.5 Experimental Evaluation

We conducted experiments to confirm whether the performance of the source separation using the prior distribution is better than the one using the template sounds. We separated sound mixtures which were generated by mixing musical instrument sounds in the "RWC Music Database: Musical Instrument Sound" [59] according to the SMFs of the "RWC Music Database: Jazz Music" and "RWC Music Database: Classical Music" [61] which were excerpted to be about 30 seconds. In this experiment, we compared the following two conditions:

Table 4.1: List of SMFs excerpted from RWC Music Database. Instruments are abbreviated, and are explained in Table 4.2.

Data Symbol	Instruments	Ave. # of sources
Classical No.2	VN, VL, VC, CB, TR, OB, FG, FL	6.23
Classical No.3	VN, VL, VC, CB, TR, OB, FG, CL, FL	6.51
Classical No.12	VN, VL, VC, CB, FL	4.23
Classical No.16	VN, VL, VC, CL	3.30
Classical No.17	VN, VL, VC, CL	3.76
Classical No.22	PF	4.33
Classical No.30	PF	4.94
Classical No.34	PF	5.96
Classical No.39	PF, VN	5.92
Classical No.40	PF, VN	7.54
Jazz No.1	PF	2.75
Jazz No.5	PF	6.92
Jazz No.8	EG	6.47
Jazz No.9	EG	3.23
Jazz No.16	PF, EB	3.55
Jazz No.17	PF, EB	5.19
Jazz No.23	PF, EB, TS	3.64
Jazz No.24	PF, EB, TS	6.28
Jazz No.27	PF, AG, EB, AS, TS, BS	11.71
Jazz No.28	PF, AG, EB, AS, TS, BS	5.46

1. using the prior distribution of the model parameters

2. using the template sounds

4.5.1 Experimental Conditions

We used 20 SMFs in total, which are listed in Table 4.1: ten SMFs are classical musical pieces and the other ten SMFs are jazz pieces. We prepared musical instrument sounds of 15 instruments listed in Table 4.2 from the RWC Music Database: Musical Instrument Sounds [59] with two performance styles and three instrument bodies. We generated sound mixtures for the test (evaluation) data by mixing the instrument sounds corresponding to the notes in the SMFs. Since we used two performance-style sets and three instrument bodies, six sound mixtures were generated from a SMF.

The experimental procedure was as follows:

Table 4.2: List of musical instruments. The instrument ID means the unique instrument number in the RWC Music Database: Musical Instrument Sounds [59].

Inst. name (Abbr.)	Inst. ID	Perf. style set A (Abbr.)	Perf. style set B (Abbr.)
Pianoforte (PF)	No.1	Normal (NO)	Staccato (ST)
Electric Guitar (EG)	No.13	Legato/Pick (LP)	Vibrato/Pick (VP)
Electric Bass (EB)	No.14	Normal/Pick (PN)	Normal/Two-finger (TN)
Violin (VN)	No.15	Normal (NO)	Non-vibrato (NV)
Viola (VL)	No.16	Normal (NO)	Non-vibrato (NV)
Cello (VC)	No.17	Normal (NO)	Non-vibrato (NV)
Contrabass (CB)	No.18	Normal (NO)	Non-vibrato (NV)
Trumpet (TR)	No.21	Normal (NO)	Vibrato (VI)
Alto Sax (AS)	No.26	Normal (NO)	Vibrato (VI)
Tenor Sax (TS)	No.27	Normal (NO)	Vibrato (VI)
Baritone Sax (BS)	No.28	Normal (NO)	Vibrato (VI)
Oboe (OB)	No.29	Normal (NO)	Vibrato (VI)
Fagotto (FG)	No.30	Normal (NO)	Vibrato (VI)
Clarinet (CL)	No.31	Normal (NO)	Vibrato (VI)
Flute (FL)	No.33	Normal (NO)	Vibrato (VI)

Table 4.3: Experimental conditions

Frequency Analysis	Sampling rate	16 kHz
	Analyzing method	STFT*
	STFT window	2048 points Gaussian
	STFT shift	160 points (10 ms)
Constant parameters	L_{H}	20
	M_{H}	30
	L_{I}	20
	M_{I}	30
	φ	440.0
	ς	1.134
MIDI sound generator for template sounds		Roland SD-90

* Short-time Fourier Transform

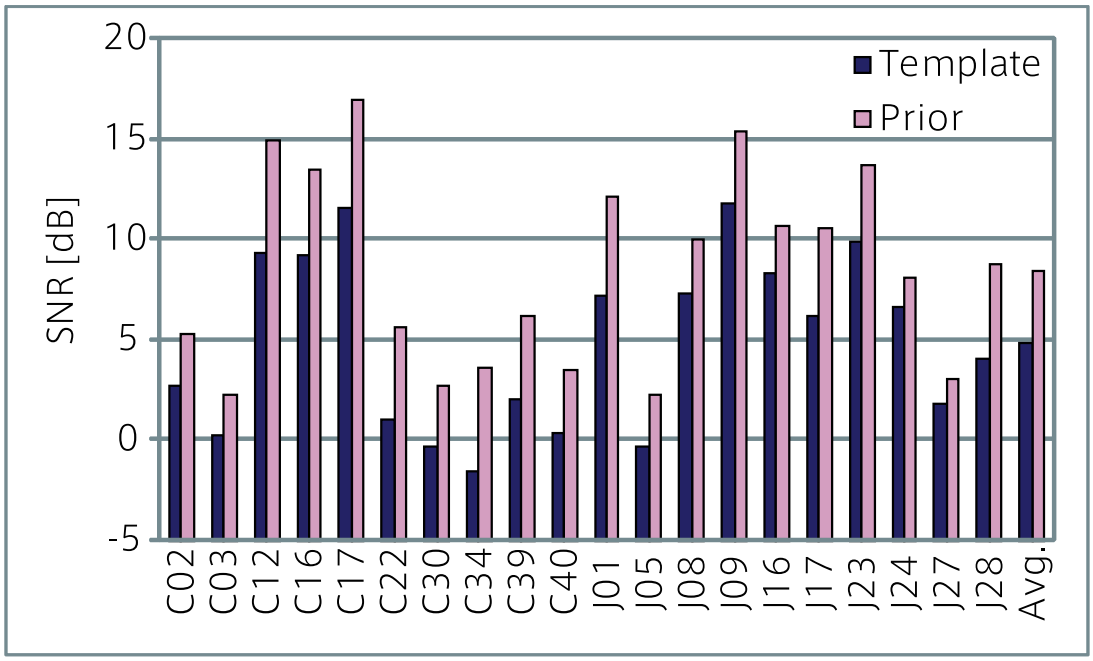


Figure 4.1: SNRs of separated signals

1. initialize the integrated model of each musical note using the corresponding template sound,
2. estimate all the model parameters from the input sound mixture, and
3. calculate the signal-to-noise ratio (SNR) for the evaluation.

SNR is defined as follow:

$$\mathrm{SNR} = \frac{1}{|\mathbb{T}|}\int_{\mathbb{T}} 10\log_{10}\int_{\mathbb{F}} \frac{\dot{X}_j(t,f)^2}{\left(\dot{X}_j(t,f) - \hat{X}_j(t,f)\right)^2}\,df\,dt, \tag{4.13}$$

where $|\mathbb{T}|$ is the duration of the time domain, and $\dot{X}_j(t,f)$ is the ground-truth magnitude spectrogram corresponding to the j-th note (i.e., the spectrogram of an actual sound before mixing). We have *original*, i.e., before mixing, source signals. If we obtain completely separated signals, the SNRs of these signals must be positive infinity, or the SNRs will decrease as the separation performance becomes worse. Other experimental conditions are shown in Table 4.3.

4.5.2 Experimental results

The average of SNRs of the sound mixtures for each musical piece is shown in Figure 4.1, and Figure 4.2 shows the SNRs for each musical instrument and performance style. The

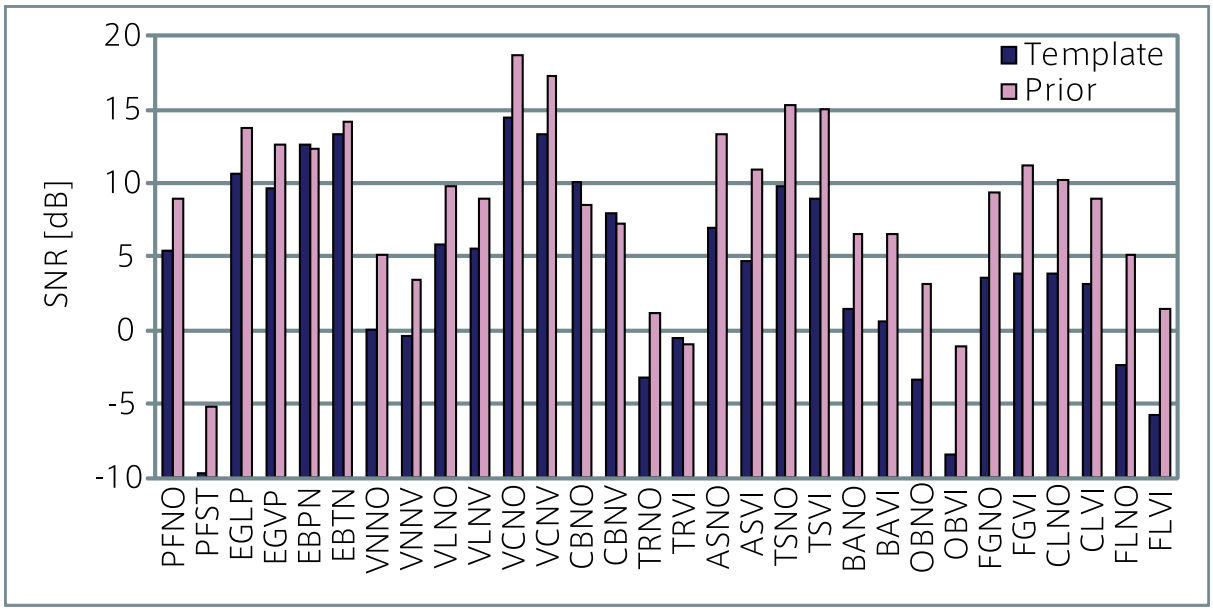


Figure 4.2: SNRs of separated signals for each musical instrument

SNRs improved from 4.89 to 8.48 dB in average by using the prior distributions. This result shows the robustness and effectiveness of our model parameter estimation method under the timbre difference between musical instrument sounds consisting of input sound mixtures and template sounds. Template sounds were generated from only one musical instrument body and performance style. These bodies and styles would be different from the ones of the input mixture signals and this difference decreased the separation performance.

The SNRs of pianoforte (PF) show a difference of more than 10 dB between the normal (NO) and the staccato (ST) styles, although the difference of other instruments between styles is at most 5 dB. Pianoforte sounds with the staccato style have long silence period because the duration of these sounds is shorter than each note in the test data. Noise in the silence period decrease the SNR even though the noise added to the separated signal is little.

The SNRs of the electric bass (EB) with the pick/normal (PN) style, contrabass (CB) with both styles, and trumpet (TR) with vibrato (VI) style decreased, as shown in Figure 4.2. This decrease is considered to be caused by the following reasons:

1. the prior distributions with inappropriate parameter values,

2. the frequency resolution in low-frequency area.

In the future, reason (1) could be corrected by using an appropriate prior distribution, such as a mixture of the Dirichlet distributions. This approach is effective in dealing with the timbre difference caused by performance styles. Reason (2) could be corrected by increasing the length of the Short-time Fourier Transform (STFT) window or using a nonlinear frequency analysis method, such as the wavelet transform.

4.6 Summary

We summarize this chapter as follows:

- Using a musical score corresponding to the audio mixture, we avoided the instrument sound recognition problem. Pitch, onset time, duration, and instrument information are extracted from the score and used for model parameter initialization and constraints for parameter estimation.

- We proposed an instrument timbre representation method by template sounds. Template sounds are generated from the SMF and MIDI sound module. The integrated models

- We proposed an instrument timbre representation method by prior distributions of the model parameters. We regarded the parameter estimation method described in Chapter 3 as a maximum likelihood estimation and realized a maximum *A Posteriori* estimation which maximizes the sum of the logarithmic conditional probability (model likelihood) and the timbre constraints corresponding to the logarithmic prior probability of the model parameters.

- We conducted an experimental evaluation to compare template-based and prior-based timbre representation methods and the result showed that the prior-based method improved the quality of the separated spectrograms.

Chapter 5

Instrument Equalizer and Its Application to Query-by-Example Music Information Retrieval

This chapter describes two applications that use sound source separation. First, we describe a music remixing interface, named *INTER (InstrumenT EqualizER)*, that allows users to control the volume of each instrument part within existing audio recordings in real time. Users can manipulate volume balance of the instruments and remix existing musical pieces. To change the volume, all instrument parts are separated from the input sound mixture. A GUI and a physical fader are combined for the user interface so that users control intuitively the volume.

Second, we describe a novel query-by-example (QBE) approach in music information retrieval that allows a user to customize query examples by directly modifying the volume of different instrument parts. The underlying hypothesis of this approach is that the musical mood of retrieved results changes in relation to the volume balance of different instruments. On the basis of this hypothesis, we aim to clarify the relationship between the change in the volume balance of a query and the genre of the retrieved pieces. Such an understanding would allow us to instruct users in how to generate alternative queries without finding other appropriate pieces. Our QBE system first separates all instrument parts from the audio signal of a piece with the help of its musical score, and then it lets users remix these parts to change the acoustic features that represent the musical mood of the piece. Experimental results showed that the shift was actually caused by the volume change in the vocal, guitar, and drum parts.

5.1 Instrument Equalizer

This section describes our music remixing interface, named *INTER*, in which a user can listen to and remix a musical piece in real time. It has sliders corresponding to different musical instruments and enables a user to manipulate the volume of each instrument part in polyphonic audio signals. The overall system is shown in Figure 5.1. It has two features for remixing audio mixtures as follows:

1. *Volume control function.* It provides the remixing function by boosting or cutting the volume of each instrument part, not by controlling the gain of a frequency band. A user can listen to the remixed sound mixture as soon as the user manipulates the volume.

2. *Interlocking with the hardware controller.* In addition to a typical mouse control on the screen, we allow a user to use a hardware controller shown in Figure 5.1 with multiple faders. It enables the user to manipulate the volume intuitively and quickly. This hardware controller makes it easy to manipulate the volume of multiple parts at the same time, while it is difficult on a mouse control.

To remix a polyphonic musical signal, the signal must be separated into each instrument part and we separate it by the separation method described in Chapter 3 and Chapter 4.

5.1.1 Internal architectures

This section describes the internal architectures of controlling the volume of each instrument part. The procedures described in this section are performed in real time under the assumption that the musical signals of each instrument part already have been obtained in advance from the target polyphonic musical signal, as described in Section 3 and Section 4. Let $x_k(t)$ and $y_k(t)$ be a separated signal and the volume of instrument k at time t, respectively. $y_k(t)$ satisfies the following condition:

$$\forall k, t : 0 \leq y_k(t) \leq 1, \tag{5.1}$$

and $y_k(t)$ is obtained from the value of volume slider k. The overview of the architecture is shown in Figure 5.2.

Figure 5.1: Instrument Equalizing System consists of GUI and physical controller.

1. *Volume control function.* The output signal, $x(t)$, is obtained as

$$x(t) = \sum_k y_k(t)\, x_k(t). \tag{5.2}$$

   Each $y_k(t)$ is obtained in real-time from the volume sliders.

2. *Interlocking with the hardware controller.* The GUI and the hardware controller communicate by MIDI. If users control the hardware fader, a MIDI message which represents the new volume is sent to the GUI, and vice versa. Since a motor is embedded in the fader, MIDI messages from the GUI move the fader to the position corresponding value of the volume.

5.1.2 Discussion

We empirically know that users feel the efficient auditory feedback towards controlling the volume balance with 8 dB of sound source separation quality by using INTER. Figure 5.3 shows a correlation between the SNR and the average number of notes for each musical piece based on the experimental result in Section 4.5. In 11 musical pieces 20 pieces, the

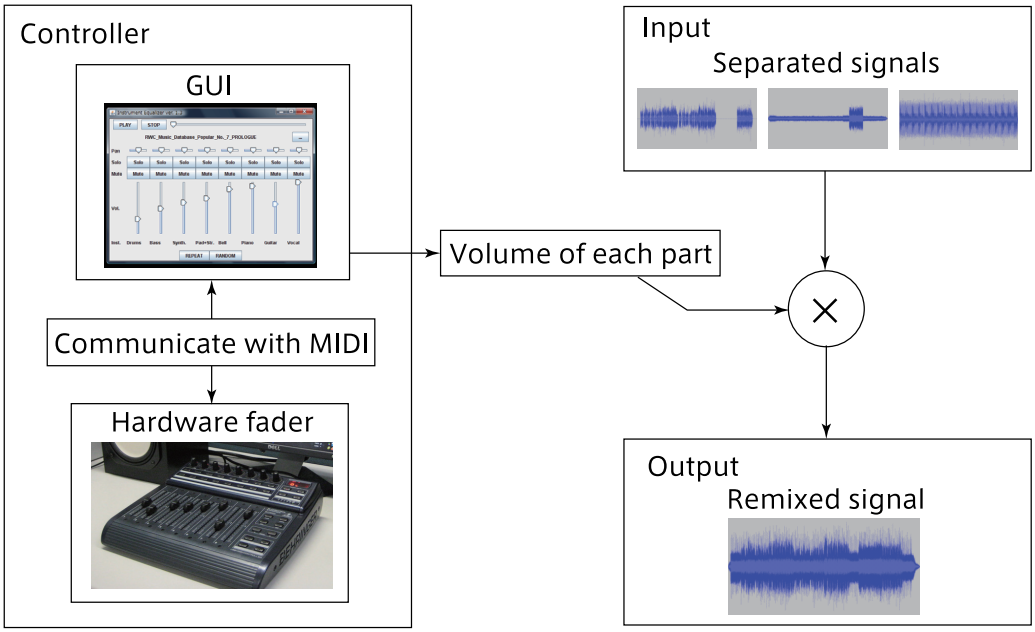


Figure 5.2: System architecture.

SNRs are exceeds 8 dB and the average numbers of notes of them are 2 – 6. Figure 5.4 shows a correlation between the averaged SNR for each time of each musical note and the number of notes in the corresponding time based on the same experiment. In the case of the number of notes is less than 8, the quality of separated sources exceeds 8 dB. These results show that INTER performs the full ability in controlling the volume balance of musical pieces whose average number of notes is less than 7. However, the source separation quality should be improved since many classic and popular songs have the average number of notes more than 7.

5.2 Query-by-Example Music Information Retrieval

One of the most promising approaches in music information retrieval is query-by-example (QBE) retrieval [62–68], where a user can receive a list of musical pieces ranked by their similarity to a musical piece (example) that the user gives as a query. This approach is powerful and useful, but the user has to prepare or find examples of favorite pieces, and it is sometimes difficult to control or change the retrieved pieces after seeing them because another appropriate example should be found and given to get better results. For example, even if a user feels that vocal or drum sounds are too strong in the retrieved pieces, it is difficult to find another piece that has weaker vocal or drum sounds while

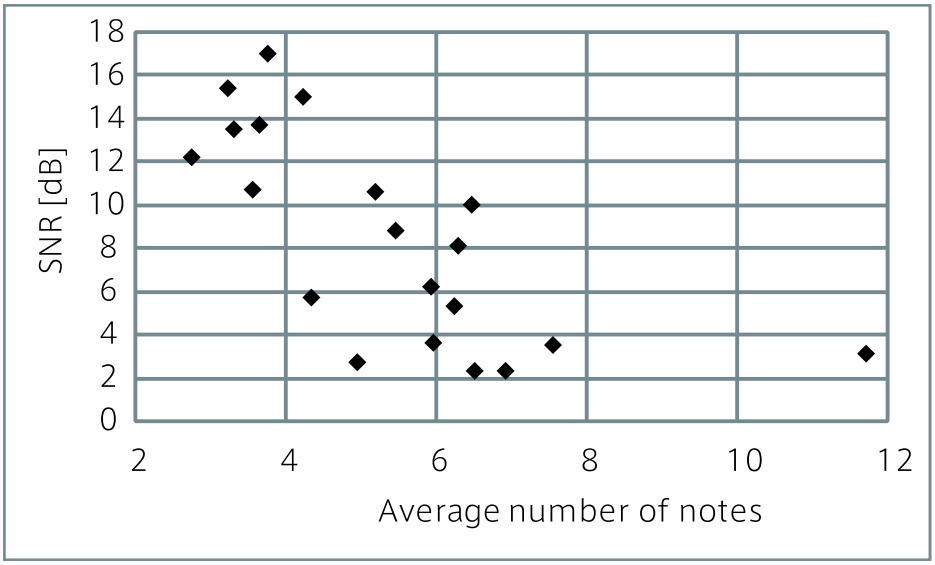


Figure 5.3: Correlation between SNR and average number of notes for each musical piece.

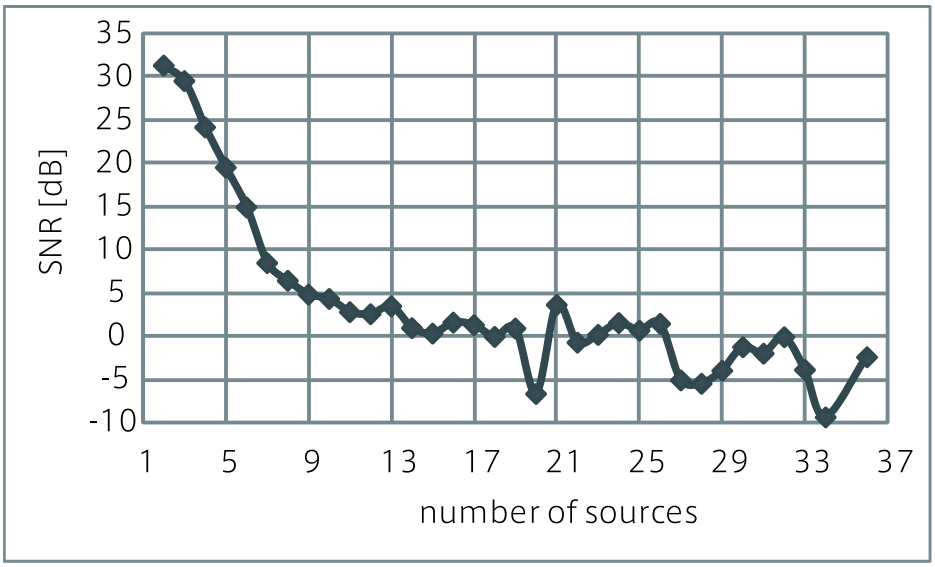


Figure 5.4: Correlation between averaged SNR for each frame of each musical note and the number of notes performed in the corresponding frame.

maintaining the basic mood and timbre of the first piece. Since finding such music pieces is now a matter of trial and error, we need more direct and convenient methods for QBE. Here we assume that QBE retrieval system takes audio inputs and treat low-level acoustic features (e.g., Mel-frequency cepstral coefficients, spectral gradient, etc.).

We solve this inefficiency by allowing a user to create new query examples for QBE by remixing existing musical pieces, i.e., changing the volume balance of the instruments. To obtain the desired retrieved results, the user can easily give alternative queries by changing the volume balance from the piece's original balance. For example, the above problem can be solved by customizing a query example so that the volume of the vocal or drum sounds is decreased. To remix an existing musical piece, we use an original sound source separation method that decomposes the audio signal of a musical piece into different instrument parts on the basis of its musical score. To measure the similarity between the remixed query and each piece in a database, we use the Earth Movers Distance (EMD) between their Gaussian Mixture Models (GMMs). The GMM for each piece is obtained by modeling the distribution of the original acoustic features, which consist of intensity and timbre.

The underlying hypothesis is that changing the volume balance of different instrument parts in a query grows diversity of the retrieved pieces. To confirm this hypothesis, we focus on the musical genre since musical diversity and musical genre have a certain level of relationship. A music database consists of various genre pieces is suitable for the purpose. We define the term *musical genre shift* as the change of musical genres in the retrieved pieces[1]. Note that this does not mean that the genre of the query piece itself can be changed. Based on this hypothesis, our research focuses on clarifying the relationship between the volume change of different instrument parts and the shift in the musical genre in order to instruct a user in how to easily generate alternative queries. To clarify this relationship, we conducted three different experiments. The first experiment examined how much change in the volume of a single instrument part is needed to cause a musical genre shift using our QBE retrieval system. The second experiment examined how the volume change of two instrument parts (a two-instrument combination for volume change) cooperatively affects the shift in genre. This relationship is explored by examining the genre distribution of the retrieved pieces. These experimental results show that the

[1]We target genres that are mostly defined by organization and volume balance of musical instruments, such as classical music, jazz, and rock. Although several genres are defined by specific rhythm patterns and singing style, e.g., waltz and hip-hop, we exclude them.

desired musical genre shift in the QBE results was easily achieved by simply changing the volume balance of different instruments in the query. The third experiment examined how the source separation performance affects the shift in the genre. The retrieved pieces using sounds separated by our method are compared with those using original sounds before mixing down in producing musical pieces. The experimental result showed that the separation performance for predictable feature shifts depends on an instrument part.

5.2.1 Query-by-Example Retrieval System

In this section, we describe our QBE retrieval system for retrieving musical pieces based on the similarity of mood between musical pieces.

Musical Genre Shift

Our original term *"musical genre shift"* means a change in the musical genre of pieces based on auditory features, which is caused by changing the volume balance of musical instruments. For example, by boosting the vocal and reducing the guitar and drums of a popular song, auditory features extracted from the modified song are similar to the features of a jazz song. The instrumentation and volume balance of musical instruments affects the musical mood. The musical genre does not have direct relation to the musical mood but musical genre shift in our QBE approach suggests that remixing query examples grow the diversity of retrieved results. As shown in Figure 5.5, by automatically separating the original recording (audio signal) of a piece into musical instrument parts, a user can change the volume balance of these parts to cause a genre shift.

Acoustic Feature Extraction

Acoustic features that represent the musical mood are designed as shown in Table 5.1 upon existing studies of mood extraction [69]. These features extracted from the magnitude spectrogram, $X(t, f)$, for each frame (100 frames per second). The spectrogram is calculated by short-time Fourier transform of the monauralized input audio signal, where t and f are the frame and frequency indices, respectively.

Overall intensity for each frame, $S_1(t)$, and intensity of each subband, $S_2(i, t)$, are defined as

$$S_1(t) = \sum_{f=1}^{F_N} X(t, f) \tag{5.3}$$

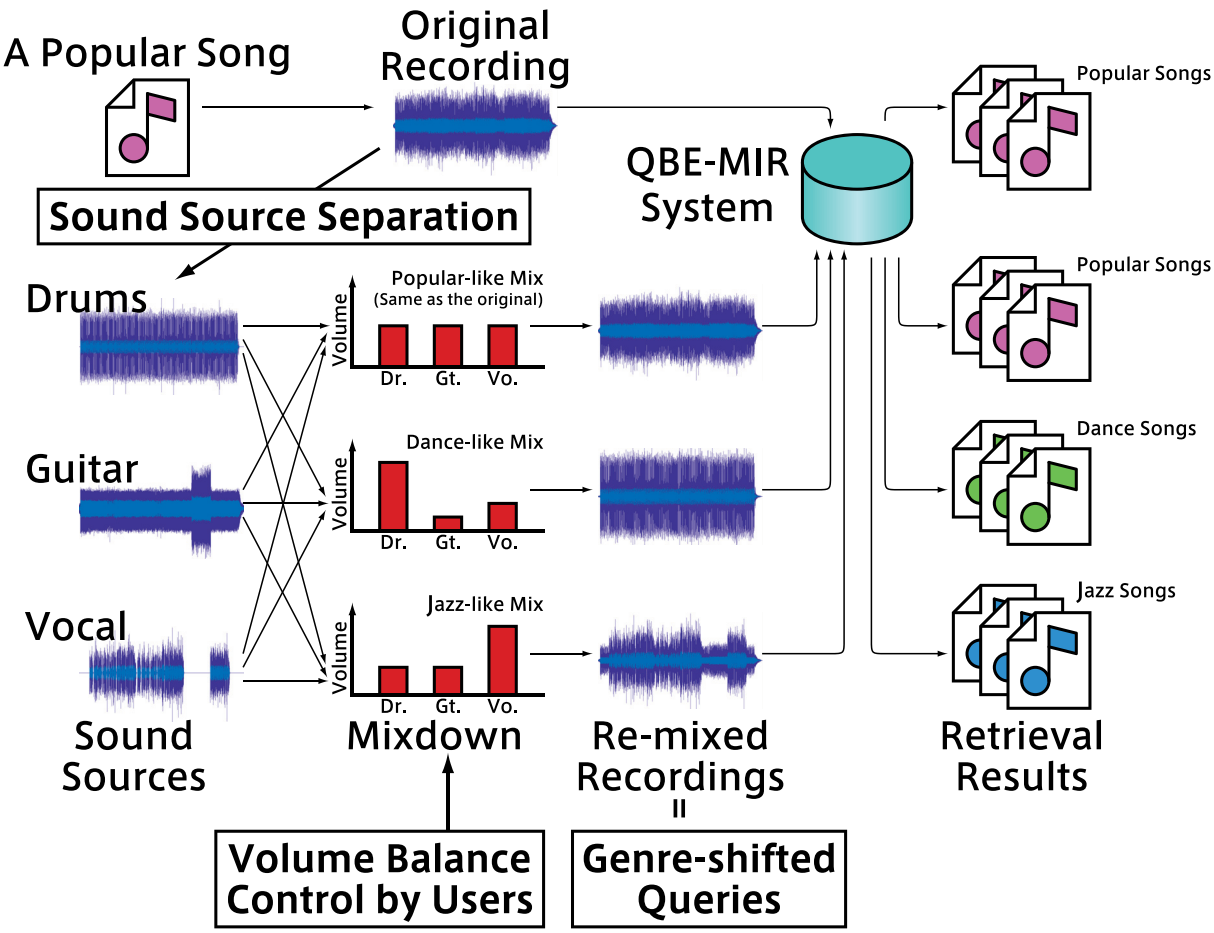
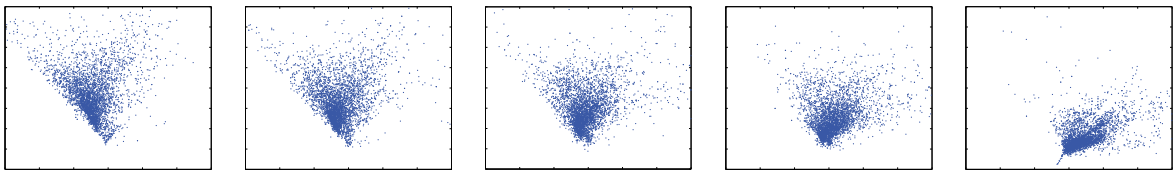


Figure 5.5: Overview of QBE retrieval system based on genre shift. Controlling the volume balance causes a genre shift of a query song, and our system returns songs that are similar to the genre-shifted query.



(a) $-\infty$ dB (b) -5 dB (c) ± 0 dB (d) $+5$ dB (e) $+\infty$ dB

Figure 5.6: Distributions of first and second principal components of extracted features from No. 1 piece of the RWC Music Database: Popular Music. Five figures show the shift of feature distribution by changing the volume of the drum part. The shift of feature distribution causes the genre shift.

Table 5.1: Acoustic features representing musical mood

Acoustic intensity features		
Dim.	Symbol	Description
1	$S_1(t)$	Overall intensity
2–8	$S_{2,1}(t), \ldots, S_{2,7}(t)$	Intensity of each subband*
Acoustic timbre features		
Dim.	Symbol	Description
9	$S_3(t)$	Spectral centroid
10	$S_4(t)$	Spectral width
11	$S_5(t)$	Spectral rolloff
12	$S_6(t)$	Spectral flux
13–19	$S_{7,1}(t), \ldots, S_{7,7}(t)$	Spectral peak of each subband*
20–26	$S_{8,1}(t), \ldots, S_{8,7}(t)$	Spectral valley of each subband*
27–33	$S_{9,1}(t), \ldots, S_{9,7}(t)$	Spectral contrast of each subband*

* 7-band octave filterbank.

and

$$S_2(i,t) = \sum_{f=F_L(i)}^{F_H(i)} X(t,f), \tag{5.4}$$

where F_N is the number of frequency bins of the magnitude spectrogram and $F_L(i)$ and $F_H(i)$ are the indices of lower and upper bounds for the i-th subband, respectively. The intensity of each subband helps to represent acoustic brightness. We use octave filterbanks that divide the magnitude spectrogram into n octave subbands:

$$\left[1, \frac{F_N}{2^{n-1}}\right), \left[\frac{F_N}{2^{n-1}}, \frac{F_N}{2^{n-2}}\right), \ldots, \left[\frac{F_N}{2}, F_N\right], \tag{5.5}$$

where n is the number of subbands, which is set to 7 in our experiments. These filterbanks cannot be constructed because they have ideal frequency response, we implemented these by division and sum of the magnitude spectrogram.

Acoustic timbre features consist of spectral shape features and spectral contrast features, which are known to be effective in detecting musical moods [69, 70]. The spectral shape features are represented by spectral centroid $S_3(t)$, spectral width $S_4(t)$, spectral rolloff $S_5(t)$, and spectral flux $S_6(t)$ as follows:

$$S_3(t) = \frac{\sum_{f=1}^{F_N} X(t,f) f}{S_1(t)}, \tag{5.6}$$

$$S_4(t) = \frac{\sum_{f=1}^{F_N} X(t,f)(f - S_3(t))^2}{S_1(t)}, \tag{5.7}$$

$$\sum_{f=1}^{S_5(t)} X(t,f) = 0.95 S_1(t), \tag{5.8}$$

and

$$S_6(t) = \sum_{f=1}^{F_N} (\log X(t,f) - \log X(t-1,f))^2. \tag{5.9}$$

The spectral contrast features are obtained as follows. Let a vector,

$$(X(i,t,1), X(i,t,2), \ldots, X(i,t,F_N(i))), \tag{5.10}$$

be the magnitude spectrogram in the t-th frame and i-th subband. By sorting these elements in descending order, we obtain another vector,

$$(X'(i,t,1), X'(i,t,2), \ldots, X'(i,t,F_N(i))), \tag{5.11}$$

where

$$X'(i,t,1) > X'(i,t,2) > \cdots > X'(i,t,F_N(i)) \tag{5.12}$$

as shown in Figure 5.7, and $F_N(i)$ is the number of the i-th subband frequency bins:

$$F_N(i) = F_H(i) - F_L(i). \tag{5.13}$$

Here, the spectral contrast features are represented by spectral peak $S_7(i,t)$, spectral valley $S_8(i,t)$, and spectral contrast $S_9(i,t)$ as follows:

$$S_7(i,t) = \log \left(\frac{\sum_{f=1}^{\beta F_N(i)} X'(i,t,f)}{\beta F_N(i)} \right), \tag{5.14}$$

$$S_8(i,t) = \log \left(\frac{\sum_{f=(1-\beta)F_N(i)}^{F_N(i)} X'(i,t,f)}{\beta F_N(i)} \right), \tag{5.15}$$

and

$$S_9(i,t) = S_7(i,t) - S_8(i,t), \tag{5.16}$$

where β is a parameter for extracting stable peak and valley values, which is set to 0.2 in our experiments.

Similarity calculation

Our QBE retrieval system needs to calculate the similarity between musical pieces, i.e., a query example and each piece in a database, on the basis of the overall mood of the piece.

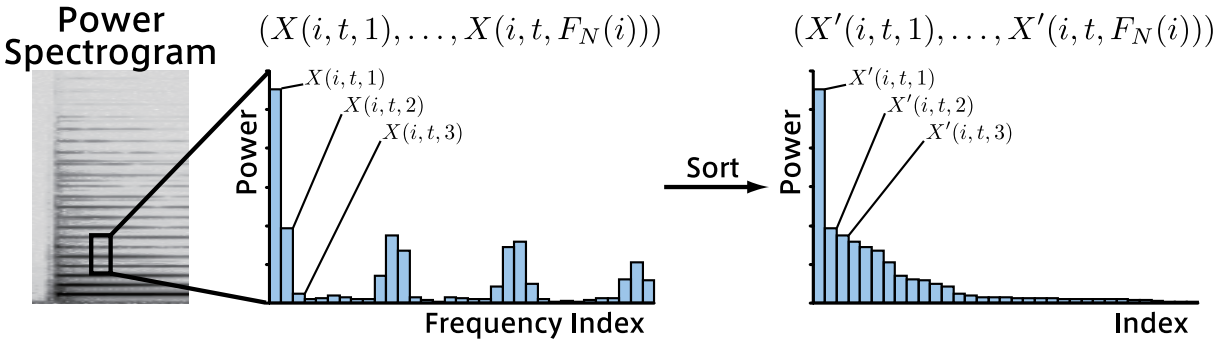


Figure 5.7: Sorted vector of magnitude spectrogram

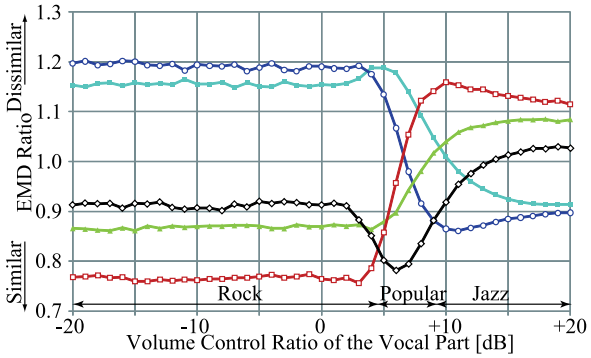
To model the mood of each piece, we use a Gaussian Mixture Model (GMM) that approximates the distribution of acoustic features. We set the number of mixtures to 8 empirically, although previous study [69] used a GMM with 16 mixtures since we used smaller database than that study for experimental evaluation. Although the dimension of the obtained acoustic features was 33, it was reduced to 9 by using the principal component analysis where the cumulative percentage of eigenvalues was 0.95.

To measure the similarity among feature distributions, we utilized Earth Movers Distance (EMD) [71]. The EMD is based on the minimal cost needed to transform one distribution into another one.

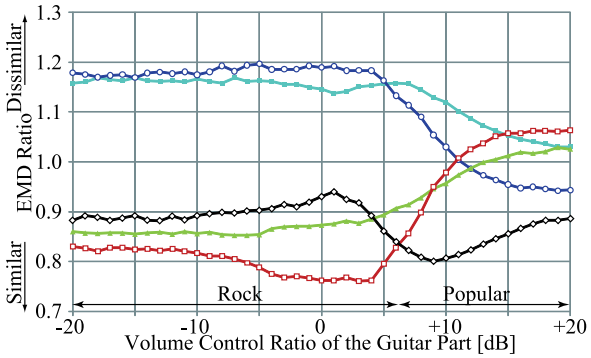
5.2.2 Experimental Evaluation

We conducted two experiments to explore the relationship between instrument volume balances and genres. Given the query musical piece in which the volume balance is changed, the genres of the retrieved musical pieces are investigated. Furthermore, we conducted an experiment to explore the influence of the source separation performance on this relationship, by comparing the retrieved musical pieces using clean audio signals before mixing down (*original*) and separated signals (*separated*).

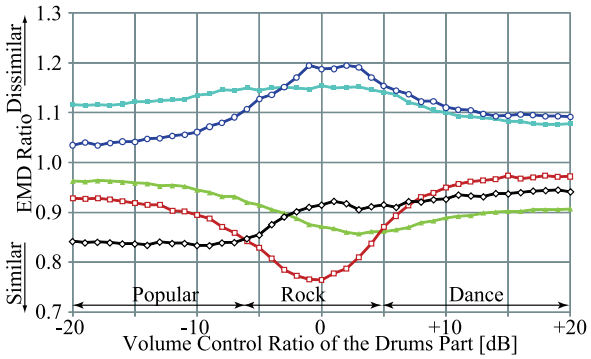
Ten musical pieces were excerpted for the query from the *RWC Music Database: Popular Music* (RWC-MDB-P-2001 No. 1–10) [61]. The audio signals of these musical pieces were separated into each musical instrument part using the standard MIDI files, which are provided as the AIST annotation [72]. The evaluation database consisted of 50 other musical pieces excerpted from the *RWC Music Database: Musical Genre* (RWC-MDB-G-2001). This excerpted database includes musical pieces in the following genres: Popular, Rock, Dance, Jazz, and Classical. The number of pieces is listed in Table 5.2.



(a) Genre shift caused by changing the volume of vocal. Genre with the highest similarity changed from rock to popular and to jazz.

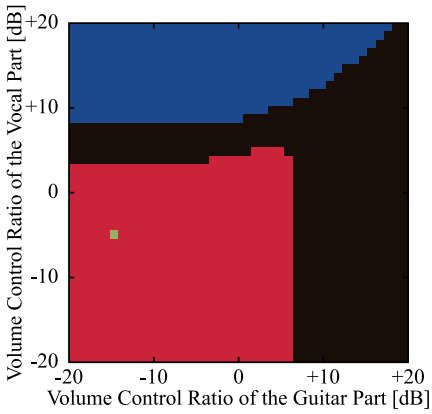


(b) Genre shift caused by changing the volume of guitar. Genre with the highest similarity changed from rock to popular.

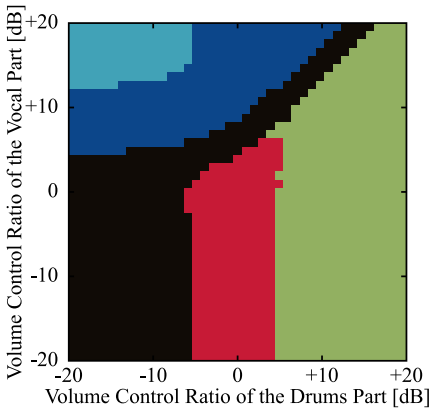


(c) Genre shift caused by changing the volume of drums. Genre with the highest similarity changed from popular to rock and to dance.

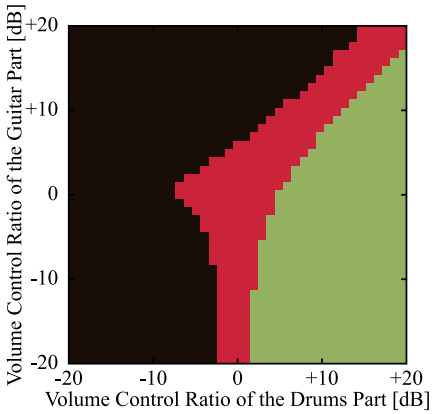
Figure 5.8: Ratio of average EMD per genre to average EMD of all genres while reducing or boosting the volume of single instrument part. Here, (a), (b), and (c) are for the vocal, guitar, and drums, respectively. Note that a smaller ratio of the EMD plotted in the lower area of the graph indicates higher similarity.



(a) Genre shift caused by changing the volume of vocal and guitar.



(b) Genre shift caused by changing the volume of vocal and drums.



(c) Genre shift caused by changing the volume of guitar and drums.

Figure 5.9: Genres that have the smallest EMD (the highest similarity) while reducing or boosting the volume of two instrument parts. The top, middle, and bottom are the cases of the vocal-guitar, vocal-drums, and guitar-drums, respectively.

Table 5.2: Number of musical pieces for each genre

Genre	Number of pieces
Popular	6
Rock	6
Dance	15
Jazz	9
Classical	14

In the experiments, we reduced or boosted the volumes of three instrument parts — vocal, guitar, and drums. To shift the genre of the musical piece by changing the volume of these parts, the part of an instrument should have sufficient duration[2]. Thus, the above three instrument parts were chosen because they satisfy the following two constraints:

1. played in all 10 musical pieces for the query, and
2. played for more than 60% of the duration of each piece.

Volume change of single instrument

The EMDs were calculated between the acoustic feature distributions of each query song and each piece in the database, while reducing or boosting the volume of these musical instrument parts between −20 and +20 dB. Figure 5.8 shows the results of changing the volume of a single instrument part. The vertical axis is the relative ratio of the EMD averaged over the 10 pieces, which is defined as

$$\text{EMD ratio} = \frac{\text{average EMD of each genre}}{\text{average EMD of all genres}}. \tag{5.17}$$

The results in Figure 5.8 clearly show that the genre shift occurred by changing the volume of any instrument part. Note that the genre of the retrieved pieces at 0 dB (giving the original queries without any changes) is the same for all three figures (a), (b), and (c). Although we used 10 popular songs excerpted from the *RWC Music Database: Popular Music* for the queries, they are considered to be rock music as the genre with the highest similarity at 0 dB because those songs actually have the true rock flavor with strong guitar and drum sounds.

By increasing the volume of the vocal from −20 dB, the genre with the highest similarity shifted from rock (−20 to 4 dB), to popular (5 to 9 dB), and to jazz (10 to 20

[2]For example, the volume of an instrument that is performed for 5 seconds in a 5-minute musical piece may not affect the genre of the piece.

dB) as shown in Figure 5.8 (a). By changing the volume of the guitar, the genre shifted from rock (−20 to 7 dB) to popular (8 to 20 dB) as shown in Figure 5.8 (b). Although it was commonly observed that the genre shifted from rock to popular in both cases of vocal and guitar, the genre shifted to jazz only in the case of vocal. These results indicate that the vocal and guitar would have different importance in jazz music. By changing the volume of the drums, genres shifted from popular (−20 to −7 dB), to rock (−6 to 4 dB), and to dance (5 to 20 dB) as shown in Figure 5.8 (c). These results indicate a reasonable relationship between the instrument volume balance and the musical genre shift, and this relationship is consistent with typical impressions of musical genres.

Volume change of two instruments (pair)

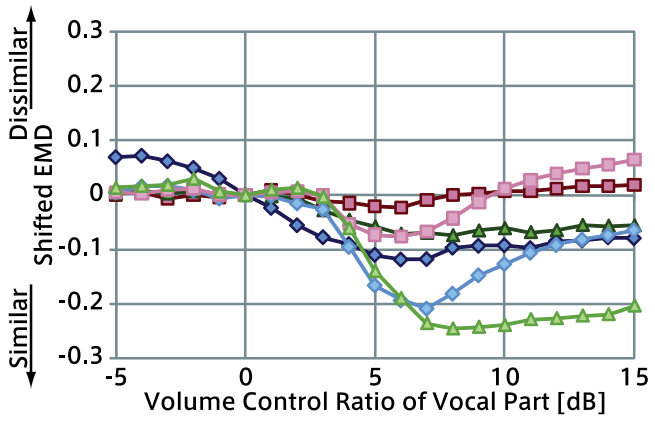
The EMDs were calculated in the same way as the previous experiment. Figure 5.9 shows the results of simultaneously changing the volume of two instrument parts (instrument pairs). If one of the parts is not changed (at 0 dB), the results are the same as those in Figure 5.8.

Although the basic tendency in the genre shifts is similar to the single instrument experiment, classical music, which does not appear as the genre with the highest similarity in Figure 5.8, appears in Figure 5.9 (b) when the vocal part is boosted and the drum part is reduced. The similarity of rock music decreased when we *separately* boosted either the guitar or the drums, but it is interesting that rock music can keep the highest similarity if both the guitar and drums are boosted *together* as shown in Figure 5.9 (c). This result closely matched with the typical impression of rock music, and it suggests promising possibilities for this technique as a tool for customizing the query for QBE retrieval.

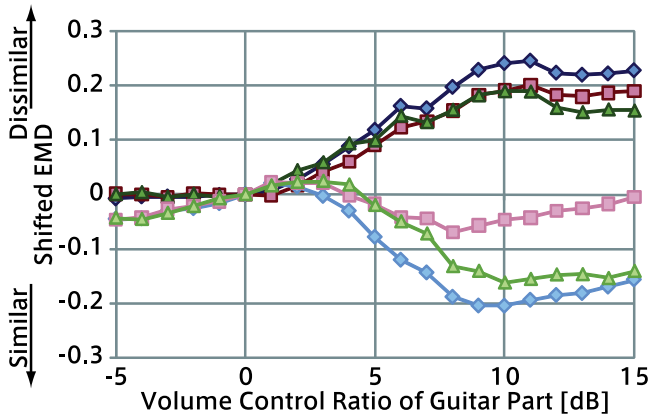
Comparison between original and separated sounds

The EMDs were calculated while reducing or boosting the volume of the musical instrument parts between −5 and +15 dB. Figure 5.10 shows the normalized EMDs that are shifted to 0 when the volume control ratio is 0 dB. Since all query songs are popular music, EMDs between query songs and popular pieces in the evaluation database tend to be smaller than the pieces of other genres. In this experiment, EMDs were normalized because we focused on the shifts in the acoustic features.

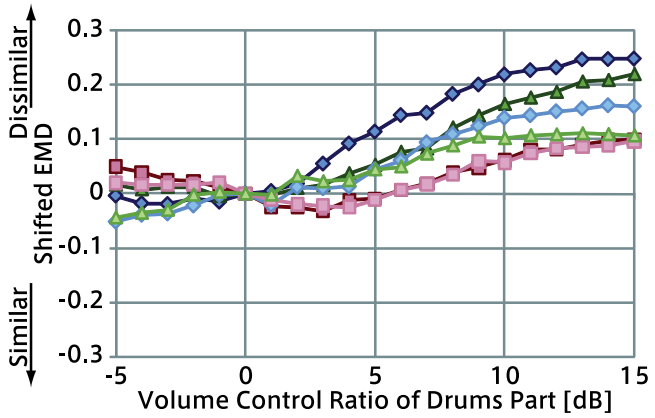
By changing the volume of the drums, the EMDs plotted in Figure 5.10(c) have similar curves in both of the *original* and *separated* conditions. On the other hand, by changing



(a) Normalized EMDs by changing the volume of vocal.



(b) Normalized EMDs by changing the volume of guitar.



(c) Normalized EMDs by changing the volume of drums.

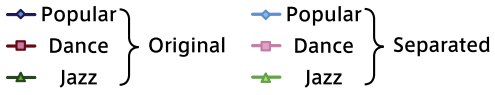


Figure 5.10: Normalized EMDs while reducing or boosting the volume. The top, middle, and bottom graphs are obtained by changing the volume of the vocal, guitar, and drum parts, respectively. Note that a smaller EMD plotted in the lower area of each graph indicates higher similarity than the one without volume controlling.

the volume of the guitar, the EMDs plotted in Figure 5.10(b) showed that a curve of the original condition is different from a curve of the separation condition. This result indicates that the shifts of features in those conditions were different. Average source separation performance of the guitar part was -1.77 dB, which was a lower value than those of vocal and drum parts. Noises included in the separated sounds of the guitar part induced this difference. By changing the volume of the vocal, the plotted EMDs of popular and dance pieces have similar curves, but the EMDs of jazz pieces have different curves, although the average source separation performance of the vocal part is the highest among these three instrument parts. This result indicates that the separation performance for predictable feature shifts depends on the instrument part.

5.2.3 Discussion

The aim of this section is achieving a QBE approach which can retrieve diverse musical pieces by boosting or reducing the volume balance of the instruments. To confirm the performance of the QBE approach, evaluation using a music database which has wide variations is necessary. A music database consists of various genre pieces is suitable for the purpose. We defined the term *musical genre shift* as the change of musical genres in the retrieved pieces since we focus on the diversity of the retrieved pieces not on musical genre change of the query example.

Although we conducted objective experiments to evaluate the effectiveness of our QBE approach, several questions remain as open questions.

1. More evidences of our QBE approach by subjective experiments are needed whether the QBE retrieval system can help users search better results.

2. In our experiments, we used only popular musical pieces as query examples. Remixing query examples except popular pieces can shift genres of retrieved results.

For source separation, we use the MIDI representation of a musical signal. Mixed and separated musical signals contain variable features: timbre difference from musical instruments' individuality, characteristic performances of instrument players such as vibrato, and environments such as room reverberation and sound effects. These features can be controlled implicitly by changing the volume of musical instruments and therefore QBE systems can retrieve various musical pieces. Since MIDI representations does not contain these features, diversity of retrieved musical pieces will decrease and users cannot evaluate

the mood difference of the pieces if we use only musical signals which are synthesized from MIDI representations.

In the experiments, we used precisely synchronized SMFs at most 50 milliseconds of onset timing error. In general, synchronization between CD recordings and their MIDI representations are not enough for separation. Previous studies on audio-to-MIDI synchronization methods [73, 74] can help this problem. We experimentally confirmed onset timing error under 200 milliseconds does not decrease source separation performance. Another problem is that the proposed separation method needs a complete musical score with melody and accompaniment instruments. A study of source separation method with a MIDI representation of specified instrument part [75] will help solving the accompaniment problem.

5.3 Summary

In Section 5.1, we described INTER which enable users to control the volume balance of musical instruments in a musical audio mixture. INTER separates the audio mixture into each instrument in advance and change the volumes in real-time. We empirically know that users feel the efficient auditory feedback towards controlling the volume balance with 8 dB of sound source separation quality by using INTER.

In Section 5.2, we described how musical genres shift by changing the volume of separated instrument parts and explained a QBE retrieval approach on the basis of such genre shift. This approach is important because it was not possible for a user to customize the QBE query in the past, which required the user to always find different pieces to obtain different retrieved results. By using the genre shift based on our original sound source separation method, it becomes easy and intuitive to customize the QBE query by simply changing the volume of instrument parts. Experimental results confirmed our hypothesis that the musical genre shifts in relation to the volume balance of instruments.

Although the current genre shift depends on only the volume balance, other factors such as rhythm patterns, sound effects, and chord progressions would also be useful for causing the shift if we could control them. In the future, we plan to pursue the promising approach proposed in this paper and develop a better QBE retrieval system that easily reflects the user's intention and preferences.

Chapter 6

Discussions

This chapter first discusses the main contributions of this study, then discusses the remaining issues and future directions.

6.1 Major Contributions

In the thesis, we pointed out three issues in musical audio signal processing by the separation-synthesis system: (1) source separation of musical instrument sounds without distinguishing harmonic and inharmonic ones, (2) instrument sound recognition in polyphonic musical audio signals, and (3) evaluation of the separation of all musical instrument sounds. To solve the first issue, we proposed the integrated harmonic and inharmonic model and the source separation method using the model. We solved the second issue by the score-informed source separation method using timbre constraints by the prior distribution of the model parameters. We tackled the third issue by developing the Instrument Equalizer and applying it to query-by-example-based music information retrieval. These approaches are based on the following two aspects of separation-synthesis-based musical audio signal processing:

- separate instrument sounds from polyphonic musical audio signals (the first and second issues,) and

- synthesize new musical audio signals by converting and combining the separated instrument sounds (the third issue.)

The main contributions of these are summarized as follows:

6.1.1 Toward Sound Source Separation

Separation of both harmonic and inharmonic sounds and separation into each instrument sound To handle harmonic and inharmonic instrument sounds, we developed the integrated harmonic and inharmonic model, and to separate complex musical audio mixture into each instrument sound, we performed the score-informed source separation. As we described in Chapter 2, previous studies related with source separation of musical audio signals are divided into two approaches: (a) describes each instrument sound and separates the mixture to them by the mixture of sinusoidals, mixture of constrained Gaussian functions, and template spectrograms and (b) separate the mixture on the basis of the statistical independency of the sources or the geometric property of the spectrogram by the ICA, NMF, and HPSS. Advantages and disadvantages of approaches (a) and (b) were complementary. This study simultaneously realized the advantages of (a) and (b).

Representation of the property of musical instrument by using the prior distribution of the model parameters Sound source separation methods which decomposes the observed mixture into each instrument sound must satisfy a essential requirement: instrument sound models correctly represent the sound of the corresponding instrument. Other properties for separation, e.g., independency of the sources, geometric property of the spectrogram, and localization of the instruments, are not essential because these properties are subsidiary to the requirement. In this study, we used prior distributions of the models parameters to satisfy the requirement.

6.1.2 Toward Musical Instrument Sound Analysis and Synthesis

Instrument Equalizer and Its Application to QBE-MIR We developed an audio player, Instrument Equalizer, and applied to a query-by-example music information retrieval system. Applications based on analysis and synthesis of musical instrument sounds realize active music appreciation [76]. Instrument Equalizer enables users to manipulate the volume of each instrument part in polyphonic audio signals by separating the audio signals into each instrument. We also realized a music information retrieval system which enables users to retrieve various musical pieces from single query piece by customizing the query piece using Instrument Equalizer.

Generative Model By analyzing musical audio signal using the integrated models, acoustic features of the instrument sounds, pitch, duration, temporal change of magnitude, and relative magnitude of the harmonics, are extracted. We developed prior distributions of the these audio features (model parameters) to represent the auditory property of each instrument in Chapter 4. Conversely, we believe that the instruments can be estimated from these acoustic features. We have developed a system which separates a mixture of instrument sounds and estimates the instruments which performed the sounds by a Bayesian extension of the integrated model [77].

We can also re-synthesize instrument sounds based on the model parameters obtained from the existing musical pieces, e.g., pitch and duration complementing, instrument timbre morphing, and phrase changing. The first application synthesizes instrument sounds with arbitrary pitch and duration from a few example sounds. The second application synthesizes sounds whose timbre is intermediate of multiple instruments. The third application synthesizes musical audio signal with a different phrase (melody) from performance features which is contained in the extracted model parameters.

6.2 Remaining Issues and Future Directions

There are many issues that remain to be resolved and future directions for research. Some of these are summarized below.

6.2.1 To Improve Versatility

Automatic extention of instrument model and its hierarchization of the model We used and trained the prior distributions of the model parameters on the basis of the assumption that the musical instruments consisting of the musical piece are given. However, it is impossible to train the distributions for all musical instrument in advance. Although instrument models for basic instruments should be trained in advance, models for other unknown instruments should be created and trained as necessary by recognizing that the instruments are different from the basic ones. In extending the instrument models, taxonomy of instruments and their groups [78–81] should be helpful. This extention can be applied to simultaneous processing of instrument identification and source separation.

Prior information based on song structure and musical genre We assumed musical audio signals as the mixtures of asynchronously emitted multiple musical instrument

sounds and we did not constrained onset time and rhythm patterns of the musical sounds. However, actual musical pieces are composed on the basis of some musical structures and constraints. Many musical pieces can be classified into some kind of musical genre and Rhythm pattens and musical instruments differ according to genre. Many popular musical pieces have common musical structure (introduction – verse – chorus – bridge – conclusion) and each component of the structure have different musical properties. By using these information obtained from musical audio signal itself, source separation without musical scores can be realized.

6.2.2 To Improve Quality of the Separated Instrument Sounds

Instrument sound model for complex spectrogram We have developed the integrated model by assuming the additivity of the magnitude spectrogram. Although this is a reasonable assumption for sparse audio mixtures, this assumption does not hold true for complex ones. Errors arising from this assumption degrade the quality of separated sources. This degrade can be solved by defining the complex spectral models and separating the complex spectrogram in which the additivity is always true.

Hidden Markov model-based temporal magnitude representation We defined the temporal magnitude variation as the weighted sum of Gaussian functions which extend and shrink according to the duration. Although this model can represent various magnitude curve, there is an unsolved problem: temporal resolution is inconsistent. The temporal resolution of the model should be defined by the resolution of time-frequency analysis, i.e., window shift in STFT, and should not be defined by the model. This problem can be solved by representing the temporal magnitude curves using models which are enable to deal with directly time-series data, e.g., hidden Markov models.

Modeling of audio effects Various kinds of audio effects, e.g., equalization, reverberation, delay, distortion, pitch modification, vocoder, and compressor, are generally used when mixing musical audio signals when composing musical pieces. By applying audio effects, the quality of musical signal analysis generally degrades. This degradation can be improved by defining the instrument sound models which can deal with audio effects and distinguishing the original spectrogram of the instrument sounds and the distortion by the effects.

Precise musical instrument model We used a simple model, Dirichlet prior distributions for the weight parameters, to constrain the property of instruments to the integrated models. This model is not enough to represent actual properties of musical instruments. For example, the timbre of many musical instrument sounds changes according to the pitch [82, 83]. The quality of separated sources can be improved by using musical instrument models which can represent precisely the timbre and auditory properties of instrument sounds, e.g., F0-dependency of timbre.

6.2.3 Other Future Directions

Instrument sound mimic Although an analysis-synthesis system is necessary for active music listening, analysis-synthesis does not necessarily mean source separation and some kinds of active music listening applications do not need source separation. Generally, the sound distortion from audio effects such as reverberation and equalization is smaller than the one from separation. This problem can be solved by synthesizing solo instrument sounds which mimic each component of the audio mixture.

Chapter 7

Conclusions

In the thesis, we dealt with spectral modeling for separating musical instrument sounds corresponding to each musical note in polyphonic music. Our goal was to separate a complex audio mixture into each musical instrument sound, to develop a functional audio player, INTER, and to apply INTER to query-by-example music information retrieval. To achieve this, we focused on two aspects of sound source separation: (1) separation of both harmonic and inharmonic musical instrument sounds and (2) separation of complex musical audio signals. We tackled three issues of (i) spectral modeling comprising harmonic and inharmonic instrument sounds, (ii) recognition of complex musical instrument sound mixture, and (iii) ensuring property of instrument to the spectral models. To solve issue (i), we proposed the integrated model that consists of a harmonic and an inharmonic tone models. To solve issue (ii), we proposed a score-informed sound source separation. To solve issue (iii), we proposed a parameter estimation method using prior distributions of the timbre parameters.

The six chapters are summarized below.

Chapter 1 described the motivation and goal of this study. We then discussed how the thesis was positioned from different viewpoints on sound source separation. We clarified the four issues and described the corresponding approaches, taking the discussions into account.

Chapter 2 reviewed state-of-the-art work in related fields. The review covered a wide range of topics, from sound source separation to instrument sound representation.

Chapter 3 presented a method for sound source separation for monaural musical audio signals which include both harmonic and inharmonic instrument sounds. We defined the integrated weighted mixture model consisting of harmonic and inharmonic models to model the spectrogram of various musical instrument sounds. To decompose the magni-

tude spectrogram of the input audio mixture, we introduce spectral distribution functions and formulate the sound source separation problem and derive the optimal distribution function. An experimental evaluation result shows that source separation performance was improved by integrating the harmonic and inharmonic models.

Chapter 4 presented methods to separating musical audio signals using the integrated harmonic and inharmonic models and prior information based on the musical score corresponding to the audio. We introduced two approaches of instrument timbre modeling: template sounds and prior distributions of the model parameters. By playing back each pair of note-on and note-off messages of the SMF on a MIDI sound module, we prepared sampled sounds for each note, template sounds. Template sounds constrain the model parameters for each model and prior distributions constrain them for each instrument. Experimental results showed that the quality of separated sounds based on the prior distributions is better than ones based on the template sounds.

Chapter 5 presented two applications that use sound source separation. First, we described a music remixing interface, INTER, that allows users to control the volume of each instrument part within existing audio recordings in real time. Users can manipulate volume balance of the instruments and remix existing musical pieces. Second, we describe a novel query-by-example (QBE) approach in music information retrieval that allows a user to customize query examples by directly modifying the volume of different instrument parts. Our QBE system first separates all instrument parts from the audio signal of a piece with the help of its musical score, and then it lets users remix these parts to change the acoustic features that represent the musical mood of the piece. Experimental results showed that the shift was actually caused by the volume change in the vocal, guitar, and drum parts.

Chapter 6 discussed the major contributions made by this study to different research fields, particularly to sound source separation and instrument sound representation. We also discussed issues that still remain to be resolved and future directions we wish to research.

We hope that our studies will trigger further attempts to develop an ultimate sound source separation and musical instrument analysis system.

Appendix A

Simultaneous Processing of Source Separation and Instrument Identification Using Bayesian Spectral Modeling

This chapter describes a method of both separating audio mixtures into sound sources and identifying the musical instruments of the sources. A statistical tone model of the magnitude spectrogram, called an integrated model, is defined and source separation and instrument identification are carried out on the basis of Bayesian inference. Since, the parameter distributions of the integrated model depend on each instrument, the instrument name is identified by selecting the one that has the maximum relative instrument weight. Experimental results showed correct instrument identification enables precise source separation even when many overtones overlap.

A.1 Introduction

Musical instrument identification in complex musical audio mixtures and sound source separation of instrument sounds are challenging problems in musical audio processing. These problems have thus far been treated independently. For example, methods of musical instrument identification [11, 28] have been reported based on fundamental frequency (F0) estimates [12–14], tempo estimates and beat tracking [15–17]. Methods of sound source separation have also been reported for separating harmonic sounds [41, 84] and separating percussive ones [34, 44]. Although methods of blind source separation and source (talker) identification have been reported [33] for multi-channel audio signals recorded by using a microphone array, these methods cannot be applied to musical audio

signals since most musical audio signals are monaural or stereo.

We believe that instrument identification and source separation rely on each other, i.e., accurate instrument identification should help source separation and high-quality source separation should simplify instrument identification. This paper reports a method of both separating audio mixtures and identifying instruments for each sound. The inputs are an audio mixture of instrument sounds, a number of mixed sounds, and the rough onset time and F0 of each sound, and the outputs are separated audio signals and the instrument name of each sound. We solved source separation as the decomposition of the input magnitude spectrogram based on the responsibility for each instrument sound, and instrument identification as the selection of the spectral tone model based on maximum *A Posteriori* approximation. Since the distributions of the tone model parameters differ by instrument, we used prior distributions of the parameters, which were trained by using a musical instrument sound database.

A.2 Bayesian Spectral Modeling

In this section, we define an extension of the integrated model defined in Chapter 3 and describe source separation and instrument identification methods based on Bayesian inference. We assume that

$$\int_{\mathbb{T}} \int_{\mathbb{F}} X(t, f)\, df\, dt = N. \tag{A.1}$$

This assumption means that we virtually sampled N data from the observed spectrogram. Let J be the number of musical instrument sounds performed in the audio mixture and K be the number of candidate musical instruments. Onset time, duration, and pitch of each sound are given and the instrument which performed the sound is unknown. Our goal is both of estimating instruments which performed each sound, i.e., instrument identification, and decomposing the input magnitude spectrogram to each sound, i.e., source separation.

A.2.1 Harmonic and Inharmonic Tone Models

Let $Y_j(t, f)$ be a spectral model which represents the magnitude spectrogram of j-th instrument sound. Since the instrument which performed the sound is unknown, we define the magnitude spectrogram model of an instrument sound as the sum of K models

with weight parameter $w_{k|j}$:

$$Y_j(t,f) = \sum_{k=1}^{K} w_{k|j} Y_{k|j}(t,f). \tag{A.2}$$

We define each $Y_{k|j}(t,f)$ as the integrated model described in Chapter 3. We also represent the magnitude spectrogram of the audio mixture by the sum of J models with weight parameter w_j:

$$Y(t,f) = \sum_{j=1}^{J} w_j Y_j(t,f). \tag{A.3}$$

Since we add a level corresponding to the sum for all instrument to the integrated model, we also add the same level to the distribution functions. The distribution functions corresponds to expected values of latent variables representing which cluster (musical note, musical instrument, harmonic or inharmonic component, and basis function) generates virtual sample data from the observed spectrogram.

We define the logarithmic likelihood of the model, described in Eq. (3.48), as follows:

$$\begin{aligned}\log p(X,Z|\theta) = \sum_{j=1}^{J}\sum_{k=1}^{K}\Bigg(&\sum_{l=0}^{L_{\mathrm{H}}-1}\sum_{m=1}^{M_{\mathrm{H}}}\int_{\mathbb{T}}\int_{\mathbb{F}} \hat{X}_{j,k,\mathrm{H},l,m}(t,f)\log\frac{\hat{X}_{j,k,\mathrm{H},l,m}(t,f)}{Y_{j,k,\mathrm{H},l,m}(t,f)}\,df\,dt \\ &+\sum_{l=0}^{L_{\mathrm{I}}-1}\sum_{m=1}^{M_{\mathrm{I}}}\int_{\mathbb{T}}\int_{\mathbb{F}} \hat{X}_{j,k,\mathrm{I},l,m}(t,f)\log\frac{\hat{X}_{j,k,\mathrm{I},l,m}(t,f)}{Y_{j,k,\mathrm{I},l,m}(t,f)}\,df\,dt\Bigg).\end{aligned} \tag{A.4}$$

A.2.2 Prior distribution

We introduce prior distributions to prevent the model parameters from deviating in source separation and instrument identification. For example, the energy distribution of the inharmonic component generally converges just after sound excitation and decreases with time, so usually $w_{l|j,k,\mathrm{I}} > w_{l'|j,k,\mathrm{I}}$ $(l < l')$. Since acoustic features, e.g., relative magnitude of the harmonic components, are different for each musical instrument, we use different prior distributions of the model parameters for each instrument. Prior distributions are trained by estimating the model parameters for isolated musical instrument sounds from a sound database with noninformative priors and averaging them. Let θ be a whole set of model parameters; the prior distributions are described as:

$$p(\theta) = p(\boldsymbol{w})\prod_{j=1}^{J} p(\boldsymbol{w}_j)\left(\prod_{k=1}^{K} p(\boldsymbol{w}_{j,k})\,p(\boldsymbol{w}_{j,k,\mathrm{H,T}})\,p(\boldsymbol{w}_{j,k,\mathrm{H,F}})\,p(\boldsymbol{w}_{j,k,\mathrm{I,T}})\,p(\boldsymbol{w}_{j,k,\mathrm{I,F}})\right)p(\tau_j)\,p(\phi_j,\sigma_j), \tag{A.5}$$

$$p(\boldsymbol{w}) = p_{\mathcal{D}}(w_1, \ldots, w_J; \tilde{\omega}_1, \ldots, \tilde{\omega}_J)\,, \tag{A.6}$$

$$p(\boldsymbol{w}_j) = p_{\mathcal{D}}\big(w_{1|j}, \ldots, w_{K|j}; \tilde{\omega}_{1|j}, \ldots, \tilde{\omega}_{K|j}\big) \tag{A.7}$$

$$p(\boldsymbol{w}_{j,k}) = p_{\mathcal{D}}\big(w_{\mathrm{H}|j,k}, w_{\mathrm{I}|j,k}; \tilde{\omega}_{\mathrm{H}|k}, \tilde{\omega}_{\mathrm{I}|k}\big)\,, \tag{A.8}$$

$$p(\boldsymbol{w}_{j,k,\mathrm{H,T}}) = p_{\mathcal{D}}\big(w_{0|j,k,\mathrm{H}}, \ldots, w_{L_{\mathrm{H}}-1|j,k,\mathrm{H}}; \tilde{\omega}_{0|k,\mathrm{H}}, \ldots, \tilde{\omega}_{L_{\mathrm{H}}-1|k,\mathrm{H}}\big)\,, \tag{A.9}$$

$$p(\boldsymbol{w}_{j,k,\mathrm{H,F}}) = p_{\mathcal{D}}\big(w_{1|j,k,\mathrm{H}}, \ldots, w_{M_{\mathrm{H}}|j,k,\mathrm{H}}; \tilde{\omega}_{1|k,\mathrm{H}}, \ldots, \tilde{\omega}_{M_{\mathrm{H}}|k,\mathrm{H}}\big)\,, \tag{A.10}$$

$$p(\boldsymbol{w}_{j,k,\mathrm{I,T}}) = p_{\mathcal{D}}\big(w_{0|j,k,\mathrm{I}}, \ldots, w_{L_{\mathrm{I}}-1|j,k,\mathrm{I}}; \tilde{\omega}_{0|k,\mathrm{I}}, \ldots, \tilde{\omega}_{L_{\mathrm{I}}-1|k,\mathrm{I}}\big)\,, \tag{A.11}$$

$$p(\boldsymbol{w}_{j,k,\mathrm{I,F}}) = p_{\mathcal{D}}\big(w_{1|j,k,\mathrm{I}}, \ldots, w_{M_{\mathrm{I}}|j,k,\mathrm{I}}; \tilde{\omega}_{1|k,\mathrm{I}}, \ldots, \tilde{\omega}_{M_{\mathrm{I}}|k,\mathrm{I}}\big)\,, \tag{A.12}$$

$$p(\tau_j) = p_{\mathcal{N}}\left(\tau_j; \tilde{\mu}_j, \left(\rho_j^{-2}\tilde{\chi} + \varrho_j^{-2}\tilde{\psi}\right)^{-1}\right), \tag{A.13}$$

and

$$p(\phi_j, \sigma_j) = p_{\mathcal{N}}\left(\phi_j; \tilde{\nu}_j, \left(\sigma_j^{-2}\tilde{\gamma}_k\right)^{-1}\right) p_{\mathcal{G}}\left(\sigma_j^{-2}; \tilde{\eta}_k, \tilde{\zeta}_k\right). \tag{A.14}$$

Prior distributions are defined as a conjugate prior of the corresponding parameters. Parameters without prior distributions, ρ_j, ϱ_j, φ, and ς, are treated as constants. The $p_{\mathcal{N}}(\cdot;\cdot,\cdot)$, $p_{\mathcal{D}}(\cdot;\cdot)$, and $p_{\mathcal{G}}(\cdot;\cdot,\cdot)$ mean probabilistic density functions of Gaussian, Dirichlet, and gamma distributions. The probabilistic density functions of these distributions are given as follows except for normalizing factors:

$$p_{\mathcal{N}}\big(x; \mu, \sigma^2\big) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{A.15}$$

$$p_{\mathcal{D}}(x_1, \ldots, x_N; \phi_1, \ldots, \phi_N) \propto \prod_{n=1}^{N} x_n^{\phi_n - 1} \tag{A.16}$$

and

$$p_{\mathcal{G}}(x; \eta, \zeta) \propto x^{\eta-1}\exp(-\zeta x)\,. \tag{A.17}$$

A graphical model of the observation is given in Fig. A.1.

A.2.3 Bayesian inference

As we described, source separation is defined as the decomposition of the input magnitude spectrogram. Decomposed spectrogram of the j-th sound, $\hat{X}_j(t,f)$, is obtained by multiplying the distribution function $Z(j;\, t, f)$ with the observed magnitude spectrogram:

$$\hat{X}_j(t,f) = Z(j;\, t,f)\, X(t,f) \tag{A.18}$$

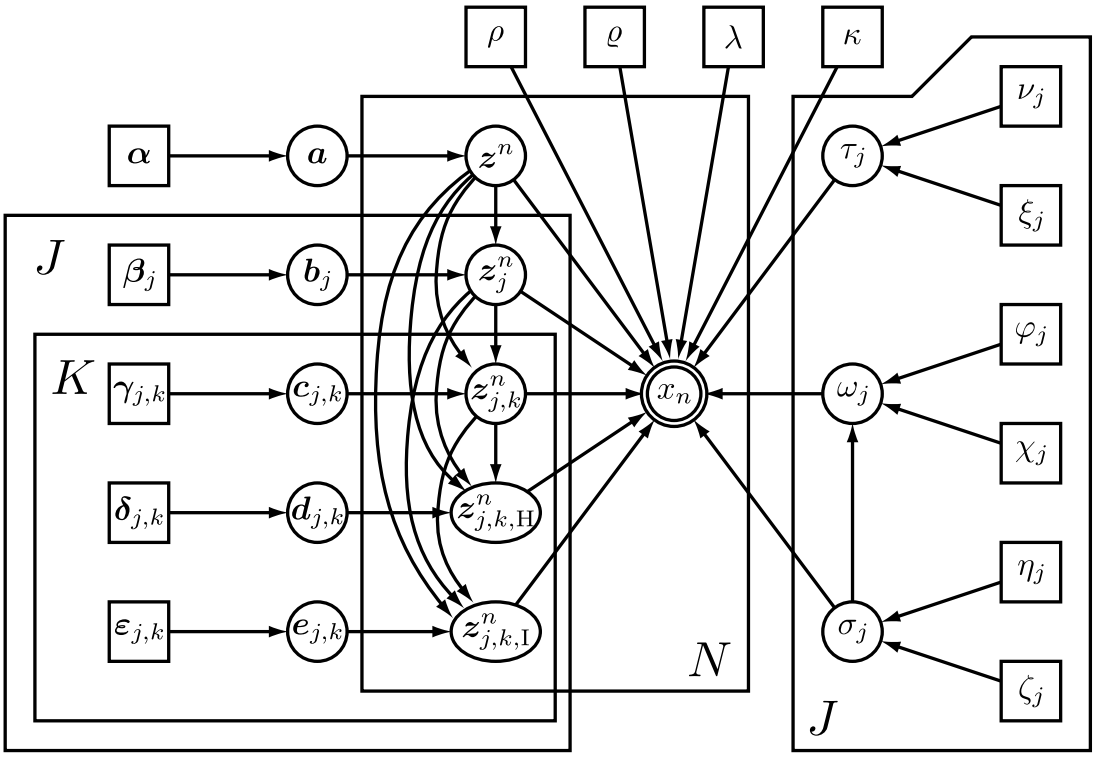


Figure A.1: Graphical model of the integrated model. n is the index of virtual sampling data from the observed spectrogram.

and

$$Z(j;\, t, f) = \arg\max_{Z} p(Z|X)\,. \tag{A.19}$$

Separated audio signals are obtained by an inverse STFT of the decomposed spectrograms.

Instrument identification is performed by model selection based on Bayesian inference. The instruments are estimated by using model selection based on maximum *A Posteriori* approximation:

$$(\text{Instrument of } j\text{-th note}) = \arg\max_{k} \left\langle w_{k|j} \right\rangle_{p(\boldsymbol{w}_j|X)}. \tag{A.20}$$

We introduce $q(Z, \theta)$ as a test distribution that approximates the true posterior distribution, $p(Z, \theta|X)$. We assume that the test distribution can be factorized as:

$$q(Z, \theta) = q(Z)\, q(\theta) \tag{A.21}$$

and

$$q(\theta) = q(\boldsymbol{w}) \prod_{j=1}^{J} q(\boldsymbol{w}_j) \left(\prod_{k=1}^{K} q(\boldsymbol{w}_{j,k})\, q(\boldsymbol{w}_{j,\mathrm{H}}k)\, q(\boldsymbol{w}_{j,\mathrm{I}}k) \right) q(\tau_j)\, q(\phi_j, \sigma_j). \tag{A.22}$$

An objective function for estimating the optimal $q(Z, \theta)$ is defined as:

$$\mathcal{F}[q] = \int\!\!\int q(Z, \theta) \log \frac{p(X, Z, \theta)}{q(Z, \theta)}\, d\theta\, dZ, \tag{A.23}$$

where $\mathcal{F}[q]$ is a functional that depends on function q. The q that maximizes $\mathcal{F}[q]$ most approximates the posterior distribution, $p(Z, \theta|X)$, under the factorization assumption.

To calculate an optimal test distribution that maximizes the objective function, we solved an Euler-Lagrange equation. When test distributions about the model parameters, $q(\theta)$, the optimal $q(Z)$ is given as:

$$q(Z(j;\, t, f)) \propto \exp \left\langle \log p(X(t, f)\,, Z(j;\, t, f)\,|\theta) \right\rangle_{q(\theta)} \tag{A.24}$$

and other distribution functions are obtained in the same way.

Let $\bar{N}_{j,k,\mathrm{H},l,m}$ and $\bar{N}_{j,k,\mathrm{I},l,m}$ be the sum of the optimal decomposed spectrograms:

$$\bar{N}_{j,k,\mathrm{H},l,m} = \int_{\mathbb{T}} \int_{\mathbb{F}} Z(l, m;\, j, k, \mathrm{H}, t, f)\, X(t, f)\, df\, dt \tag{A.25}$$

and

$$\bar{N}_{j,k,\mathrm{I},l,m} = \int_{\mathbb{T}} \int_{\mathbb{F}} Z(l, m;\, j, k, \mathrm{I}, t, f)\, X(t, f)\, df\, dt. \tag{A.26}$$

The summation or integration of them over indices, variables, and suffixes are denoted by omitting these characters, for example:

$$\bar{N}_{j,k} = \sum_{l=0}^{L_{\mathrm{H}}-1} \sum_{m=1}^{M_{\mathrm{H}}} \bar{N}_{j,k,\mathrm{H},l,m} + \sum_{l=0}^{L_{\mathrm{I}}-1} \sum_{m=1}^{M_{\mathrm{I}}} \bar{N}_{j,k,\mathrm{I},l,m}. \tag{A.27}$$

Let symbols with hat ($\hat{}$) modifier, e.g., $\hat{\omega}_j$, be parameters of the optimal $q(\theta)$, they are given as follows:

$$\hat{\omega}_j = \tilde{\omega}_j + \bar{N}_j \tag{A.28}$$

$$\hat{\omega}_{k|j} = \tilde{\omega}_{+|k} \bar{N}_{j,k} \tag{A.29}$$

$$\hat{\omega}_{\mathrm{H}|k_j} = \tilde{\omega}_{\mathrm{H}|k} + \bar{N}_{j,k,\mathrm{H}} \tag{A.30}$$

$$\hat{\omega}_{\mathrm{I}|k_j} = \tilde{\omega}_{\mathrm{I}|k} + \bar{N}_{j,k,\mathrm{I}} \tag{A.31}$$

$$\hat{\omega}_{k_j,\mathrm{H}|l,m} = \tilde{\omega}_{k_j,\mathrm{H}|l,m} + \bar{N}_{j,k,\mathrm{H},l,m} \tag{A.32}$$

$$\hat{\omega}_{k_j,\mathrm{I}|l,m} = \tilde{\omega}_{k_j,\mathrm{I}|l,m} + \bar{N}_{j,k,\mathrm{I},l,m} \tag{A.33}$$

$$\hat{\mu}_j = \frac{\left(\tilde{\chi}\tilde{\mu}_j + \bar{t}_{j,\mathrm{H}}\right)/\rho_j^2 + \left(\tilde{\psi}\tilde{\mu}_j + \bar{t}_{j,\mathrm{I}}\right)/\varrho_j^2}{\left(\tilde{\chi} + \bar{N}_{j,\mathrm{H}}\right)/\rho_j^2 + \left(\tilde{\psi} + \bar{N}_{j,\mathrm{I}}\right)/\varrho_j^2} \tag{A.34}$$

$$\hat{\chi} = \tilde{\chi} + \bar{N}_{j,\mathrm{H}} \tag{A.35}$$

$$\hat{\psi} = \tilde{\psi} + \bar{N}_{j,\mathrm{I}} \tag{A.36}$$

$$\hat{\nu}_j = \frac{\tilde{\gamma}_k\tilde{\nu}_j + \bar{f}_j^1}{\tilde{\gamma}_k + \bar{f}_j^0} \tag{A.37}$$

$$\hat{\gamma}_k = \tilde{\gamma}_k + \bar{f}_j^0 \tag{A.38}$$

$$\hat{\eta}_k = \tilde{\eta}_k + \bar{N}_{j,\mathrm{H}}/2 \tag{A.39}$$

$$\hat{\zeta}_k = \tilde{\zeta}_k + \frac{\bar{f}_j^2 - \left(\bar{f}_j^1\right)^2/\bar{f}_j^0}{2} + \frac{\tilde{\gamma}_k\bar{f}_j^0\left(\bar{f}_j^1/\bar{f}_j^0 - \tilde{\nu}_j\right)^2}{2\left(\tilde{\gamma}_k + \bar{f}_j^0\right)} \tag{A.40}$$

Auxiliary functions are defined as follows:

$$\bar{t}_{j,\mathrm{H}} = \sum_{k=1}^{K} \sum_{m=1}^{M_{\mathrm{H}}} \int_{\mathbb{T}} \int_{\mathbb{F}} (t - l\rho_j)\, \hat{X}_{j,k,\mathrm{H},l,m}(t,f)\, df\, dt \tag{A.41}$$

$$\bar{t}_{j,\mathrm{I}} = \sum_{k=1}^{K} \sum_{m=1}^{M_{\mathrm{I}}} \int_{\mathbb{T}} \int_{\mathbb{F}} (t - l\varrho_j)\, \hat{X}_{j,k,\mathrm{I},l,m}(t,f)\, df\, dt \tag{A.42}$$

$$\bar{f}_j^2 = \sum_{k=1}^{K} \sum_{l=0}^{L_{\mathrm{H}}-1} \sum_{m=1}^{M_{\mathrm{H}}} \int_{\mathbb{T}} \int_{\mathbb{F}} f^2 \hat{X}_{j,k,\mathrm{H},l,m}(t,f)\, df\, dt \tag{A.43}$$

Table A.1: Musical Instruments

Inst. name (Abbr.)	# of tones
Acoustic piano (PF)	1584
Violin (VN)	2304
Trumpet (TR)	1964
Alto sax (AS)	891
Clarinet (CL)	1080
Fagotto (FG)	1079
Marimba (MB)	909
Vibraphone (VI)	1332

Table A.2: Constants of the integrated model

Symbol	Value
M_H	30
N_H	100
N_H	30
N_H	100
ρ_j	0.05 sec.
ϱ_j	0.05 sec.
φ	440.0 Hz
ς	1.134

$$\bar{f}_j^1 = \sum_{k=1}^{K} \sum_{l=0}^{L_{\mathrm{H}}-1} \sum_{m=1}^{M_{\mathrm{H}}} \int_{\mathbb{T}} \int_{\mathbb{F}} m f \hat{X}_{j,k,\mathrm{H},l,m}(t,f)\, df\, dt \tag{A.44}$$

$$\bar{f}_j^0 = \sum_{k=1}^{K} \sum_{l=0}^{L_{\mathrm{H}}-1} \sum_{m=1}^{M_{\mathrm{H}}} \int_{\mathbb{T}} \int_{\mathbb{F}} m^2 \hat{X}_{j,k,\mathrm{H},l,m}(t,f)\, df\, dt \tag{A.45}$$

Since derivation of the test distribution of the latent variables and the parameters depend on each other, they cannot be solved in closed form. In order to estimate the optimal distribution, we update them by alternatively repeating.

A.3 Experimental Evaluation

We conducted an experiment to evaluate the efficiency of our source separation and instrument identification methods. Given audio mixtures that consisted of two or three musical instrument sounds excerpted from the *RWC Music Database: Musical Instrument Sound* [59], the audio mixtures were separated into sources and instruments were estimated. As shown in Table A.1, eight musical instruments were excerpted from the

Table A.3: Experimental results for instrument identification and source separation, which show averaged log spectral distances in instrument sounds. Bold characters mean top two numbers.

Inst.	Acc. rate [%]		Log spect. dist. ($\times 10^{-2}$)	
name	2 sounds	3 sounds	2 sounds	3 sounds
PF	63.4	28.0	3.25	3.88
VN	**87.8**	76.5	2.43	3.31
TR	79.6	61.5	2.78	3.40
AS	39.1	12.7	3.29	3.97
CL	85.1	**79.4**	**1.59**	**2.12**
FG	**91.7**	**85.1**	**1.83**	**2.39**
MB	48.6	28.2	5.25	5.34
VI	67.6	53.9	6.43	5.86
Avg.	72.1	54.8	3.12	3.65

database and sounds were divided into subsets for 10-fold cross validation. The prior distribution of each instrument was created by averaging the model parameters estimated from the training data (nine subsets). Audio mixtures were produced from the combination of the instrument sounds for each data subset except pairs consisting of the same instrument sounds. The constant parameters in the integrated models were set as listed in Table A.2. The performance of instrument identification and source separation were respectively evaluated by using the accuracy rate and log spectral distance defined as:

$$\sqrt{\sum_{t=0}^{T}\sum_{f=0}^{F}\left|20\log_{10}\frac{X_{\mathrm{org}}(t,f)}{X_{\mathrm{sep}}(t,f)}\right|^2}\Big/TF \tag{A.46}$$

Table A.3 summarizes the accuracy rate of instrument identification and log spectral distance for the source instruments. The fagotto (FG), violin (VN), and clarinet (CL) have a high accuracy rate for identification and short log spectral distances. This suggests that correct instrument identification help to improve source separation. It is easier to decompose audio mixture of two sounds than mix of three sounds and decreasing the number of sounds increases the accuracy rate of identification on average. This suggests precise source separation increases the accuracy of instrument identification. The marimba (MB) and vibraphone (VI) have larger spectral distances than the other instruments. These instrument sounds have percussive properties and are sensitive to the diffusion of onset time. The spectral distances can decrease by accurately estimating the onset time.

Fig. A.2 shows the relationship between the pitch differences in two instrument sounds

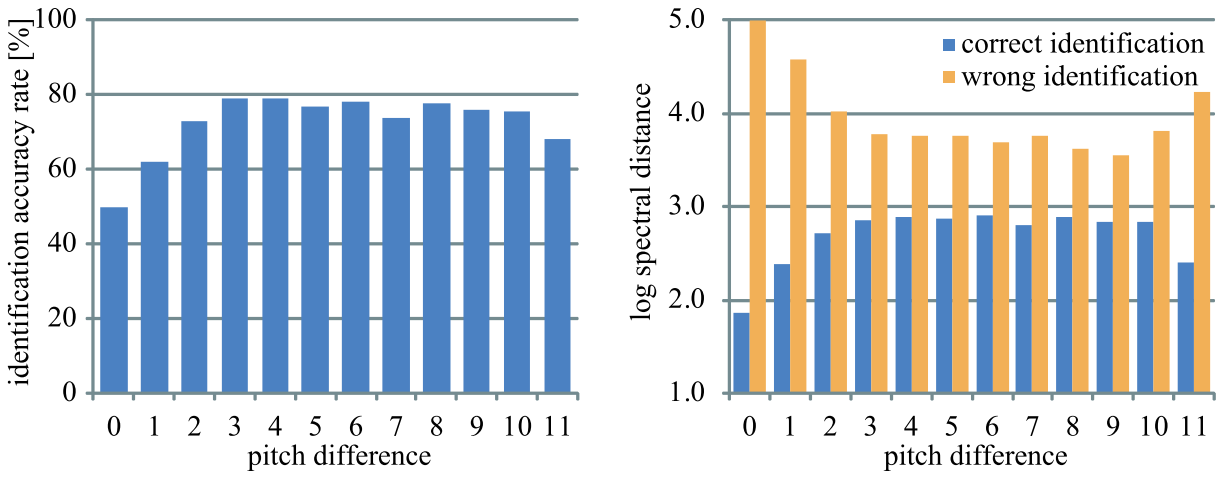


Figure A.2: Relationship between pitch (MIDI note number) difference in two instrument sounds to accuracy rate of instrument identification (left) and between the differences to log spectral distances of separated sounds whose instruments are correctly or incorrectly estimated (right).

when two sounds are mixed to the accuracy of identification and the log spectral distance of separated sounds. The pitch difference is based on the difference of MIDI pitch numbers. The spectral distances are shown in cases of correct and incorrect instrument identification. When pitch differences are 0 (unison), 1, 4 (perfect fourth), 5 (perfect fifth), and 11, many overtones overlap and this overlap decreases the accuracy of identification. This suggests that the overlap of overtones degrades the accuracy of source separation. Spectral distances also degrade when pitch differences are 0, 1, and 11 when instruments are identified incorrectly. However, when instruments are identified correctly, spectral distances did not increase when many overtones overlapped. This suggests that correct instrument identification enables precise source separation even when many overtones overlap.

A.4 Summary

We reported a method of simultaneously processing sound source separation and musical instrument identification using Bayesian spectral modeling. We defined the integrated harmonic and inharmonic tone models, decomposed the observed magnitude spectrogram by using the expectation value of the latent variable, and identified the instrument for each sound in the audio mixture by selecting the instrument based on maximum *A Posteriori* approximation. The experimental results revealed that the accuracy of instrument identification and source separation rely on each other and correct instrument identification

enables precise source separation even when many overtones overlap.

Bibliography

[1] J. A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, 1(4):32–38, 1977.

[2] C. Chafe, J. Kashima, B. Mont-Reynaud, and J. Smith. Techniques for note identification in polyphonic music. In *ICMC1985*, pages 399–405, 1985.

[3] Anssi P. Klapuri. Multipitch estimation and sound separation by the spectral smoothness principle. In *ICASSP '01*, volume V, pages 3381–3384, 2001.

[4] Tuomas Virtanen and Anssi Klapuri. Separation of harmonic sound sources using sinusoidal modeling. In *ICASSP2000*, volume II, pages 765–768, 2000.

[5] Tuomas Virtanen and Anssi Klapuri. Separation of harmonic sounds using linear models for the overtone series. In *ICASSP2002*, volume II, pages 1757–1760, 2002.

[6] Mark R. Every and John E. Szymanski. A spectral-filtering approach to music signal separation. In *DAFx-04*, pages 197–200, 2004.

[7] Matti Ryynänne, Tuomas Virtane, Jouni Paulu, and Anssi Klapur. Accompaniment separation and karaoke application based on automatic melody transcription. In *ICME2008*, pages 1417–1420, 2008.

[8] Zhiyao Duan, Yungang Zhang, Changshui Zhang, and Zhenwei Shi. Unsupervised single-channel music source separation by average harmonic structure modeling. *IEEE Trans. Audio, Speech and Lang. Process.*, 16(4):766–778, May 2008.

[9] Mathieu Lagrange, Luis Gustavo Martins, Jennifer Murdoch, and George Tzanetakis. Normalized cuts for predominant melodic source separation. *IEEE Trans. Audio, Speech and Lang. Process.*, 16(2):278–290, February 2008.

[10] Paris Smaragdis and Gautham J. Mysore. Separation by "humming": User-guided sound extraction from monophonic mixtures. In *WASPAA2009*, pages 69–72, 2009.

[11] Tetsuro Kitahara. *Computational Musical Instrument Recognition and Its Application to Content-based Music Information Retrieval*. PhD thesis, Kyoto University, 2007.

[12] Kunio Kashino. *Computational Auditory Scene Analysis for Music Signals*. PhD thesis, University of Tokyo, 1994.

[13] Masataka Goto. A real-time music-scene-analysis system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, September 2004.

[14] Anssi Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Trans. Audio, Speech and Lang. Process.*, 16(2):255–266, February 2008.

[15] Masataka Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *J. New Music Res.*, 30(2):159–171, June 2001.

[16] Simon Dixon. Automatic extraction of tempo and beat from expressive performances. *J. New Music Res.*, 30(1):39–58, 2001.

[17] Matthew E. P. Davies and Mark D. Plumbley. Context-dependent beat tracking of musical audio. *IEEE Trans. Audio, Speech and Lang. Process.*, 15(3):1009–1020, March 2007.

[18] Dan Barry, Derry Fitzgerald, Eugene Coyle, and Bob Lawlor. Drum source separation using percussive feature detection and spectral modulation. In *ISSC2005*, pages 13–17, 2005.

[19] H. Harb and L. Chen. A query by example music retrieval algorithm. In *WIAMIS '03*, pages 122–128, 2003.

[20] Daniel D. Lee and Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[21] Paris Smaragdis and Judith C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *WASPAA2003*, pages 177–180, 2003.

[22] Minje Kim, Jiho Yoo, Kyeongok Kang, and Seungjin Choi. Blind rhythmic source separation: Nonnegativity and repetability. In *ICASSP2010*, pages 2006–2009, 2010.

[23] Jiho Yoo, Minje Kim, Kyeongok Kang, and Seungjin Choi. Nonnegative matrix partial co-factorization for drum source separation. In *ICASSP2010*, pages 1942–1945, 2010.

[24] Marko Helén and Tuomas Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *EUSIPCO-2005*, 2005.

[25] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio, Speech and Lang. Process.*, 15(3):1066–1074, March 2007.

[26] Paris Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Trans. Audio, Speech and Lang. Process.*, 15(1):1–12, 2007. January.

[27] Mikkel N. Schmidt and Morten Mørup. Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In *ICA 2006*, pages 700–707, April 2006.

[28] Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Trans. Audio, Speech and Lang. Process.*, 15(3):982–994, March 2007.

[29] Kenichi Miyamoto, Hirokazu Kameoka, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. Harmonic-temporal-timbral clustering (httc) for the analysis of multi-instrument polyphonic music signals. In *ICASSP2008*, pages 113–116, 2008.

[30] Nobutaka Ono, Kenichi Miyamoto, Jonathan Le Roux, Hirokazu Kameoka, and Shigeki Sagayama. Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram. In *EUSIPCO2008*, 2008.

[31] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, April 1994.

[32] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[33] Hiroshi Saruwatari, Satoshi Kurita, Kazuya Takeda, Fumidata Itakura, Tsuyoshi Nishikawa, and Kiyohiro Shikano. Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, 2003(11):1135–1146, 2003.

[34] Michael A. Casey and Alex Westner. Separation of mixed audio sources by independent subspace analysis. In *ICMC2000*, pages 154–161, 2000.

[35] Dan Barry, Derry FitzGerald, Eugene Coyle, and Bob Lowlor. Single channel source separation using short-time independent component analysis. In *119th AES Convention*, 2005.

[36] Satoru Morita and Yasuhito Nanri. Sound source separation of trio using stereo musig sound signal based on independent component analysis. In *ICME2006*, pages 185–188, 2006.

[37] Shlomo Dubnov. Extracting sound objects by independent subspace analysis. In *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES22)*, 2002.

[38] Christian Uhle, Christian Dittmar, and Thomas Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 843–848, April 2003.

[39] Derry FitzGerald, Eugene Coyle, and Bob Lawlor. Independent subspace analysis using locally linear embedding. In *DAFx-03*, pages 13–17, 2003.

[40] Emmanuel Vincent. Musical source separation using time-frequency source priors. *IEEE Trans. Audio, Speech and Lang. Process.*, 14(1):91–98, January 2006.

[41] John Woodruff, Bryan Pardo, and Roger Dannenberg. Remixing stereo music with score-informed source separation. In *ISMIR2006*, pages 314–319, 2006.

[42] John Woodruff and Bryan Pardo. Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings. *EURASIP Journal on Applied Signal Processing*, 2007(1):1–10, January 2007.

[43] Wendong Huang and Ye Wang. A method for separating drum objects from polyphonic musical signals. In *WASPAA2005*, pages 307–310, 2005.

[44] Kazuyoshi Yoshii, Masataka Goto, and Hiroshi G. Okuno. Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression. *IEEE Trans. Audio, Speech and Lang. Process.*, 15(1):333–345, January 2007.

[45] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Drumix: An audio player with real-time drum-part rearrangement functions for active music listening. *IPSJ Journal*, 48(3):134–144, March 2007.

[46] Olivier Gillet and Gaël Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Trans. Audio, Speech and Lang. Process.*, 16(3):529–540, March 2008.

[47] Manuel Davy, Simon Godsill, and Jérôme Idier. Bayesian analysis of polyphonic western tonal music. *J. Acoust. Soc. Am.*, 119(4):2498–2517, April 2006.

[48] Emmanuel Vincent, Nancy Bertin, and Roland Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech and Lang. Process.*, 18(3):528–537, March 2010.

[49] A. Taylan Cemgil, Hilbert J. Kappen, and David Barver. A generative model for music transcription. *IEEE Trans. Audio, Speech and Lang. Process.*, 14(2):679–694, March 2006.

[50] Tero Tolonen and Matti Karjalainen. A computationally efficient multipitch analysis model. *IEEE Trans. Speech and Audio Process.*, 8(6):708–716, November 2000.

[51] Anssi Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Process.*, 11(6):804–816, November 2003.

[52] Eric D. Scheirer. Tempo and beat analysis of acoustic musical signals. *J. Acoust. Soc. Am.*, 103(1):588–601, January 1998.

[53] Hélène Laurent and Christian Doncarli. Stationarity index for abrupt changes detection in the time-frequency plane. *IEEE Signal Processing Letters*, 5(2):43–45, February 1998.

[54] Anssi P. Klapuri, Antti J. Eronen, and Jaakko T. Astola. Analysis of the meter of acoustic musical signals. *IEEE Trans. Audio, Speech and Lang. Process.*, 14(1):342–355, January 2006.

[55] Judith C. Brown. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *J. Acoust. Soc. Am.*, 103(3):1933–1941, 1999.

[56] K. D. Martin. *Sound-Source Recognition: A Theory and Computational Model*. PhD thesis, MIT, 1999.

[57] Antti Eronen and Anssi Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *ICASSP2000*, volume II, pages 753–756, 2000.

[58] Jonathan Le Roux, Nobutaka Ono, and Shigeki Sagayama. Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction. In *SAPA 2008*, pages 23–28, September 2008.

[59] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Music genre database and musical instrument sound database. In *ISMIR2003*, pages 229–230, 2003.

[60] Thomas P. Minka. Estimating a dirichlet distribution, 2000.

[61] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical, and jazz music databases. In *ISMIR2002*, pages 287–288, 2002.

[62] Andreas Rauber, Elias Pampalk, and Dieter Merkl. Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity. In *ISMIR2002*, pages 71–80, 2002.

[63] C. Cheng Yang. The MACSIS acoustic indexing framework for music retrieval: An experimental study. In *ISMIR2002*, pages 53–62, 2002.

[64] Eric Allamanche, Jürgen Herre, Oliver Hellmuth, Thorsten Kastner, and Christian Ertel. A multiple feature model for musical similarity retrieval. In *ISMIR2003*, pages 217–218, 2003.

[65] Yazhong Feng, Yueting Zhuang, and Yunhe Pan. Music information retrieval by detecting mood via computational media aesthetics. In *WI2003*, pages 235–241, 2003.

[66] Balaji Thoshkahna and K. R. Ramakrishnan. Projekt quebex: A query by example system for audio retrieval. In *ICME2005*, pages 265–268, 2005.

[67] Fabio Vignoli and Steffen Pauws. A music retrieval system based on user-driven similarity and its evaluation. In *ISMIR2005*, pages 272–279, 2005.

[68] Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Musical instrument recognizer "instrogram" and its application to music retrieval based on instrumentation similarity. In *ISM 2006*, pages 265–274, 2006.

[69] Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Trans. Audio, Speech and Lang. Process.*, 14(1):5–18, January 2006.

[70] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast features. In *ICME2002*, pages 113–116, 2002.

[71] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. A metric for distributions with applications to image databases. In *ICCV1998*, pages 59–66, 1998.

[72] Masataka Goto. AIST annotation for the RWC music database. In *ISMIR2006*, pages 359–360, 2006.

[73] Robert J. Turetsky and Daniel P. W. Ellis. Ground-truth transcriptions of real music from force-aligned MIDI synthesis. In *ISMIR2003*, 2003.

[74] Meinard Müller. *Information Retrieval for Music and Motion*, chapter 5. Springer, 2007.

[75] Naoki Yasuraoka, Takehiro Abe, Katsutoshi Itoyama, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Changing timbre and phrase in existing musical performances as you like. In *ACM Multimedia 2009*, pages 203–212, 2009.

[76] Masataka Goto. Active music listening intefaces based on signal processing. In *ICASSP2007*, volume IV, pages 1441–1444, 2007.

[77] Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Simultaneous processing of sound source separation and musical instrument identification using bayesian spectral modeling. In *ICASSP 2011*, 2011. accepted.

[78] Margaret J. Kartomi. *On Concepts and Classifications of Musical Instruments*. University Of Chicago Press, November 1990.

[79] Perfecto Herrera-Boyer, Geoffroy Peeters, and Shlomo Dubnov. Automatic classification of musical instrument sounds. *J. New Music Res.*, 32(1):3–21, March 2003.

[80] Slim Essid, Gaël Richard, and Bertrand David. Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Trans. Audio, Speech and Lang. Process.*, 14(1):68–80, January 2006.

[81] Perfecto Herrera-Boyer, Anssi Klapuri, and Manuel Davy. *Automatic Classification of Pitched Musical Instrument Sound*, chapter 6. Springer, 2006.

[82] Neville H. Fletcher and Thomas D. Rossing. *The Phisics of Musical Instruments*. Springer, second edition, 1998.

[83] Tetsuro Kitahara, Masataka Goto, and Hiroshi G. Okuno. Musical instrument identification considering pitch-dependent characteristics of timbre: A classifier based on F0-dependent multivariate normal distribution. *IPSJ Journal*, 44(10):2448–2458, October 2003.

[84] Harald Viste and Gianpaolo Evangelista. A method for separation of overlapping partials based on similarity of temporal envelopes in multichannel mixtures. *IEEE Trans. Audio, Speech and Lang. Process.*, 14(3):1051–1061, May 2006.

Relevant Publications

Chapters 3

1. Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Simultanious Realization of Score-Informed Sound Source Separation of Polyphonic Musical Signals and Constrained Parameter Estimation for Integrated Model of Harmonic and Inharmonic Structure," *IPSJ Journal*, Vol. 49, No. 3, pp. 1465–1479, March 2008 (in Japanese).

2. Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Integration and Adaptation of Harmonic and Inharmonic Models for Separating Polyphonic Musical Signals," in *Proc. of the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, Vol. I, pp. 57–60, April 2007.

Chapter 4

1. Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Parameter Estimation for Harmonic and Inharmonic Models by Using Timbre Feature Distributions," *IPSJ Journal*, Vol. 50, No. 7, pp. 1757–1767, July 2009.

2. Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Parameter Estimation for Harmonic and Inharmonic Models by Using Timbre Feature Distributions," *Journal of Information Processing*, Vol. 17, pp. 191–201, July 2009 (the same version as the above paper).

3. Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Instrument Equalizer for Query-by-Example Retrieval: Improving Sound Source Separation based on Integrated Harmonic and Inharmonic Models," in

Proc. of the 9th International Conference on Music Information Retrieval (ISMIR 2008), pp. 133–138, September 2008.

Chapter 5

1. Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Query-by-Example Music Information Retrieval by Score-Informed Source Separation and Remixing Technologies," *EURASIP Journal on Advances in Signal Processing*, accepted in January 2011.

2. Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Query-by-Example Music Retrieval Approach Based on Musical Genre Shift by Changing Instrument Volume," in *Proc. of the 12th International Conference on Digital Audio Effects (DAFx-09)*, September 2009.

Appendix 1

1. Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Simultaneous Processing of Sound Source Separation and Musical Instrument Identification Using Bayesian Spectral Modeling," in *Proc. of the 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, accepted.

All Publications by the Author

Major Publications

Journal Papers

1) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Simultanious Realization of Score-Informed Sound Source Separation of Polyphonic Musical Signals and Constrained Parameter Estimation for Integrated Model of Harmonic and Inharmonic Structure," *IPSJ Journal*, Vol. 49, No. 3, pp. 1465–1479, March 2008 (in Japanese).

2) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Parameter Estimation for Harmonic and Inharmonic Models by Using Timbre Feature Distributions," *IPSJ Journal*, Vol. 50, No. 7, pp. 1757–1767, July 2009.

3) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Query-by-Example Music Information Retrieval by Score-Informed Source Separation and Remixing Technologies," *EURASIP Journal on Advances in Signal Processing*, accepted in January 2011.

4) Takehiro Abe, Katsutoshi Itoyama, Kazuyoshi Yoshii, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "An Analysis-and-Synthesis Approach for Manipulating Pitch of a Musical Instrument Sound Considering Pitch-Dependency of Timbral Characteristics," *IPSJ Journal*, Vol. 50, No. 3, pp. 1054–1066, March 2009 (in Japanese).

5) Kouhei Sumi, Katsutoshi Itoyama, Kazuyoshi Yoshii, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Automatic Chord Sequence Recognition Based on Integration of Chord and Bass Pitch Features," *IPSJ Journal*, Vol. 52, No. 4, April

2011 (in Japanese).

International Conference Papers

6) Katsutoshi Itoyama, Tetsuro Kitahara, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Automatic Feature Weighting in Automatic Transcription of Specified Part in Polyphonic Music," in *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, pp. 172–175, October 2006.

7) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Integration and Adaptation of Harmonic and Inharmonic Models for Separating Polyphonic Musical Signals," in *Proc. of the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, Vol. I, pp. 57–60, April 2007.

8) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Instrument Equalizer for Query-by-Example Retrieval: Improving Sound Source Separation based on Integrated Harmonic and Inharmonic Models," in *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pp. 133–138, September 2008.

9) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Query-by-Example Music Retrieval Approach Based on Musical Genre Shift by Changing Instrument Volume," in *Proc. of the 12th International Conference on Digital Audio Effects (DAFx-09)*, September 2009.

10) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Simultaneous Processing of Sound Source Separation and Musical Instrument Identification Using Bayesian Spectral Modeling," in *Proc. of the 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, accepted.

11) Takehiro Abe, Katsutoshi Itoyama, Kazuyoshi Yoshii, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Synthesis approach for manipulating pitch of a musical instrument sound with considering timbral characteristics," in *Proc. of the 11th International Conference on Digital Audio Effects (DAFx-08)*, September 2008.

12) Kouhei Sumi, Katsutoshi Itoyama, Kazuyoshi Yoshii, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Automatic Chord Recognition based on Probabilistic Integration of Chord Transition and Bass Pitch Estimation," in *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pp. 39–44, September 2008.

13) Naoki Yasuraoka, Takehiro Abe, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno, "Changing Timbre and Phrase in Existing Musical Performances as You Like," in *Proc. of ACM International Conference on Multimedia (ACM Multimedia 2009)*, October 2009.

14) Akira Maezawa, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno, "Bowed String Sequence Estimation of a Violin Based on Adaptive Audio Signal Classification and Context-Dependent Error Correction," in *Proc. of the IEEE International Symposium on Multimedia (ISM2009)*, December 2009.

15) Akira Maezawa, Katsutoshi Itoyama, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Violin Fingering Estimation Based on Violin Pedagogical Fingering Model Constrained by Bowed Sequence Estimation from Audio Input," in *Proc. of the Twenty Third International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2010)*, pp. 249–259, June 2010.

16) Shimpei Aso, Takeshi Saitou, Masataka Goto, Katsutoshi Itoyama, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "SpeakBySinging: Converting Singing Voices to Speaking Voices While Retaining Voice Timbre," in *Proc. of the 13th International Conference on Digital Audio Effects (DAFx-10)*, September 2010.

Other Publications (All in Japanese)

Technical Reports

17) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Constrained Parameter Estimation of Harmonic and Inharmonic

Models for Separating Polyphonic Musical Audio Signals," in *IPSJ SIG Technical Report*, Vol. 2007, No. 37 (2007-MUS-70), pp. 81–88, May 2007.

18) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Parameter Estimation for Harmonic and Inharmonic Models by Using Timbre Feature Distributions," in *IPSJ SIG Technical Report*, Vol. 2007, No. 81 (2007-MUS-71), pp. 161–166, August 2007.

19) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "A Music Information Retrieval System Based on Timbre Similarity Using the Instrument Equalizer," in *IPSJ SIG Technical Report*, Vol. 2008, No. 78 (2008-MUS-76), pp. 143–148, August 2008.

20) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Musical Genre Shift of Polyphonic Musical Pieces by Changing Instrument Volume," in *IPSJ SIG Technical Report*, Vol. 2009-MUS-81, No. 3, pp. 1–6, July 2009.

21) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Parameter Estimation of Mixture Model of Multiple Instruments and Application to Musical Instrument Identification," in *IPSJ SIG Technical Report*, Vol. 2009-MUS-81, No. 13, pp. 1–6, July 2009.

22) Masatoshi Hamanaka, Yoshinari Takegawa, Kenichi Iwai, Naoya Takahashi, Tomoyasu Nakano, Yasunori Ohishi, Katsutoshi Itoyama, Tetsuro Kitahara, and Kazuyoshi Yoshii, "Demonstrations: Introduction of Research by Young Researchers IV," in *IPSJ SIG Technical Report*, Vol. 2006, No. 113 (2006-MUS-67), pp. 9–14, October 2006.

23) Takehiro Abe, Tetsuro Kitahara, Katsutoshi Itoyama, Masuzo Yanagida, "Parameter Estimation from Audio Signals Using Physical Models of Plucked Strings," in *Proc. of ASJ TCM on Musical Acoustics*, MA2006-91, March 2007.

24) Takehiro Abe, Katsutoshi Itoyama, Kazuyoshi Yoshii, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "A Method for Manipulating Pitch and Duration of Musical Instrument Sounds Dealing with Pitch-dependency of Timbre," in *IPSJ*

SIG Technical Report, Vol. 2008, No. 78 (2008-MUS-76), pp. 155-160, August 2008.

25) Akira Maezawa, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno, "Estimation of Bowed String Sequence of a Violin Performance Using Audio and Score-Based Anomaly Detection," in *IPSJ SIG Technical Report*, Vol. 2009-MUS-81, No. 5, pp. 1–6, July 2009.

26) Naoki Yasuraoka, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno, "Improvement of Performance Analysis-and-Symthesis Method by using Residual Spectrum Model for Reduction of Accompaniment or Sound Reverberation," in *IPSJ SIG Technical Report*, Vol. 2009-MUS-81, No. 10, pp. 1–6, July 2009.

27) Akira Maezawa, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno, "Bowed String Sequence Estimation of a Violin by Integrating String Identification Using Audio and Context-Dependent Error Correction," in *Proc. of the 2009 Autumn Meeting of ASJ*, 2-5-15, September 2009.

National Convention Papers

28) Katsutoshi Itoyama, Tetsuro Kitahara, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Automatic Feature Weighting in Automatic Transcription of Specified Part in Polyphonic Music," in *Proc. of the 68th IPSJ National Convention*, 2L-4, March 2006.

29) Katsutoshi Itoyama, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Sound Source Separation for Polyphonic Musical Signal Based on NMF Using Score Information," in *Proc. of the 69th IPSJ National Convention*, 2N-1, March 2007.

30) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Parameter Estimation for Harmonic and Inharmonic Models Using Prior Distributions from Multiple Instrument Bodies," in *Proc. of the 70th IPSJ National Convention*, 2X-6, March 2008.

31) Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Acoustic Feature Variation and Application to Similarity-based

Music Retrieval using Instrument Equalizer," in *Proc. of the 72nd IPSJ National Convention*, 6J-6, March 2010.

32) Kouhei Sumi, Katsutoshi Itoyama, Kazuyoshi Yoshii, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Automatic Chord Recognition Based on the Pitch of Bass Sound for Popular Music," in *Proc. of the 70th IPSJ National Convention*, 2X-5, March 2008.

33) Takehiro Abe, Katsutoshi Itoyama, Kazuyoshi Yoshii, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Synthesis Approach for Manipulating Pitch of a Musical Instrument Sound with Considering Timbral Characteristics," in *Proc. of the 70th IPSJ National Convention*, 2X-7, March 2008.

34) Hiroki Saito, Katsutoshi Itoyama, Kazuyoshi Yoshii, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Cross-media Retrieval Using a Congruency Model between Music and Video in Multimedia Content," in *Proc. of the 70th IPSJ National Convention*, 4X-4, March 2008.

35) Naoki Yasuraoka, Takehiro Abe, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno, "Performance Rendering and Sound Synthesis considering the Timbral Deviation within Note Sequence," in *Proc. of the 71st IPSJ National Convention*, 4R-1, March 2009.

36) Takehiro Abe, Katsutoshi Itoyama, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Musical Instrument Sound Morphing Based on Psychoacoustic Timbre Characteristics Using Harmonic and Inharmonic Models," in *Proc. of the 71st IPSJ National Convention*, 4R-2, March 2009.

37) Akira Maezawa, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno, "Probabilistic Classification of Monophonic Instrument Playing Techniques," in *Proc. of the 71st IPSJ National Convention*, 4R-3, March 2009.

38) Kaiping Wang, Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "A music retrieval approach from alternative genres of query by adjusting instrument volume," in *Proc. of the 71st IPSJ National Convention*, 5R-5, March 2009.

39) Hideki Takano, Kouhei Sumi, Katsutoshi Itoyama, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Automatic Chord Recognition Considering the Relation between Bass Pitch Probability and Chroma Vector," in *Proc. of the 71st IPSJ National Convention*, 5R-6, March 2009.

40) Shinpei Aso, Tsuyoshi Saitou, Masataka Goto, Katsutoshi Itoyama, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "SpeakBySinging: A Speaking Voice Synthesis System Converting Singing Voices to Speaking Voices By Controlling F0, Amplitude, and Duration," in *Proc. of the 72nd IPSJ National Convention*, 6U-1, March 2010.

41) Naoki Yasuraoka, Katsutoshi Itoyama, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Performance and Timbre Rendering for MIDI-Synthesized Audio Signal by using Harmonic Inharmonic GMM," in *Proc. of the 72nd IPSJ National Convention*, 5T-5, March 2010.

Awards

1) 24th TELECOM System Technology Award for Student, 2009.

2) IPSJ Funai Young Research Award, 2010.