博士論文

Source-Filter Representation and Phase Estimation in Continuous Wavelet Transform Domain for Monaural Music Audio Editing

（連続ウェーブレット変換領域における
ソースフィルタ表現と位相推定による
モノラル音楽音響信号加工の研究）

Tomohiko Nakamura

中村 友彦

連続ウェーブレット変換領域における ソースフィルタ表現と位相推定による モノラル音楽音響信号加工の研究

本研究では，モノラル音楽音響信号を音高や楽器などの単位に分解し，分解成分を個別に加工することを可能にする音楽信号分離および合成手法を提案する．これは，ユーザによる音楽制作や既存楽曲の加工支援システム，楽音それぞれ加工可能な音楽プレイヤー，計算機による自動編曲システムなどの音楽アプリケーションに応用できる．

音響信号加工の性能は音源分離の精度に大きく依存するため，加工の前段で高精度な分離を行うことが重要である．一般に高精度な分離のためには解を適切に限定するための手がかりが必要となる．時間周波数表現（スペクトログラム）領域では調波性などの音源分離に有用な手がかりを利用できるが，適切な周波数解像度のスペクトログラムを選択するためには対象がどのような音響信号であるかを考慮することが重要である．また，そのスペクトログラム上で対象の音響信号をどのように表現できるかという点も考慮する必要がある．したがって，時間周波数表現を意識したアプローチをとるべきである．

そこで，本研究では以下の 3 つの方針を考えこれらを同時に考慮した手法を提案する．まず第 1 の方針として（i）対数周波数解像度を与える連続ウェーブレット変換（continuous wavelet transform, CWT）によるスペクトログラムを利用する．これは音楽における各音高の基本周波数（F_0）は対数尺度で均等に並ぶ性質があるからである。次に，第 2 の方針として（ii）楽音の生成過程モデルを活用する．ソースフィルタ理論によると楽音の生成過程は楽器の振動体と共鳴体に分離して考えることができ，楽音のスペクトルに関する仮定が見通しよく立てられるためである．第 3 の方針として（iii）スペクトル漏れを考慮する．実際に観測されるスペクトルが取ることを許される形状には制約があり，もしスペクトル漏れの具体的な形状や関数が分かっていれば，近接した異なる音源の F_0 成分や高調波成分を分離する手がかりとなるからである．これら 3 つの方針を同時に考慮したアプローチを実現するためには，スペクトログラム領域とソースフィルタモデルなどの時間信号領域のモデルとの対応関係を得る必要がある．しかし，CWT の基底波形は直交しないためどのように対応関係

を得るべきかは必ずしも知られていない．

第 3 章では方針（i），(ii）を同時に考慮するために，CWT 領域へのソースフィルタモデルの導入に取り組む．対数周波数領域では F_0 と高調波周波数の間隔が F_0 によらず一定である性質を活かしつつ，ソースフィルタモデルを導入したスペクトログラムモデルを提案し，補助関数法と呼ばれる最適化原理を用いて提案モデルのパラメータを反復的に推定する収束性の保証されたアルゴリズムを導出する．ソースフィルタモデル導入による音源分離性能の向上を実験により確認した．

第 4 章では方針（i），(iii）を同時に考慮するために，CWT 領域でのスペクトル漏れの記述に取り組む．音源分離に有用な手がかりである時間周波数表現の局所的な構造と大局的な構造を同時に利用しつつ，スペクトル漏れを考慮したスペクトログラムモデルを導出し，前章と同様に補助関数法を用いて提案モデルのパラメータ推定アルゴリズムを導出する．この手法を調波時間因子分解（harmonic-temporal factor decomposition, HTFD）と呼ぶ．音源分離性能の評価実験によりスペクトル漏れの考慮と CWT 領域での分離の有効性を確認した．

第 5 章では全方針を考慮するため，第 4 章で提案した HTFD を拡張し CWT 領域でスペクトル漏れとソースフィルタモデルを同時に考慮したスペクトログラムモデルを導出する．HTFD の解析的な時間信号モデルを介し，離散時間信号領域で定義されるソースフィルタモデルと CWT 領域で定義された HTFD のスペクトログラムモデルのパラメータの対応関係が得られることを示し，新たなスペクトログラムモデルを提案する．前章と同様に，補助関数法を用いて閉形式の更新則からなるパラメータ推定アルゴリズムを導出する．スペクトル漏れの考慮に加えソースフィルタモデルを導入することの有効性を実験により確認した．

第 6 章では，分離後に加工された振幅スペクトログラムを時間信号に変換するために，振幅スペクトログラムからの高速位相推定法を提案する．時間周波数表現を意識したアプローチをとることで，時間領域信号に対応する複素スペクトログラムが満たす条件を導出し，その条件を元に位相推定問題が最適化問題として定式化できることを示す．前章までと同様に補助関数法に基づき収束性の保証されたアルゴリズムを導出し，提案法の有効性を実験により確認した．

第 7 章では，歌声の振幅スペクトログラムがスパース行列，伴奏の振幅スペクトログラムが低ランク行列とみなせることを利用し，L_p ノルム規準の非負値行列因子分解による歌声分離手法を提案する．歌声の振幅スペクトログラムのスパース性を表すパラメータを適切に定めることで，分離性能が向上することを確認した．

最後に，第 8 章では 2 つの音楽音響信号間で調波楽器音の周波数特性の置換および打楽器音の音色置換が可能なシステムを提案する．主観評価実験により，調波楽器音の周波数特性と打楽器音の音色どちらについても提案法の有効性を確認した．

Abstract

This thesis discusses monaural audio source separation and synthesis for decomposing music audio mixtures into musically meaningful components (e.g. pitches and notes) and the individual editing of the components. Such a process is applicable to a wide range of musical applications such as assistance systems for music composition and arrangement, music players that allow users to edit existing music pieces as per their preferences, and automated music arrangement systems.

The sound quality of edited audio signals greatly depends on the accuracy of source separation, and hence it is necessary to achieve accurate source separation followed by audio editing. Since monaural source separation is essentially an ill-posed problem, we generally require cues that allow us to adequately narrow down possible solutions. We can use various cues (e.g. harmonicity and repeating structures of music) in time-frequency representations (spectrograms). To select spectrograms having adequate frequency resolution, taking into account how music audio signals are characterized and how they are represented in the selected spectrogram domain are important. Therefore, we should take an approach that is aware of spectrograms.

To realize the approach, three principles are considered and methods are proposed in accordance with these principles. The first principle (i) is to use log-frequency spectrograms obtained with a continuous wavelet transform (CWT) since the fundamental frequencies (F_0) of musical pitches are geometrically spaced. The second principle (ii) is to utilize the source-filter model, which can describe the generation processes of instrument sounds fairly well with two components originating from vibrating objects and resonant structures. This enables us to make assumptions regarding the components individually. The third principle (iii) is to explicitly describe the spectral leakage effect. The spectral shape is not allowed to be arbitrary because of the redundancy of time-frequency transforms. Thus, identifying the shape of the spectral leakage can be valid for separating adjacent F_0 and harmonic com-

ponents of different audio sources. To simultaneously satisfy the three principles, clarifying the relation between spectrogram-domain and time-domain models, such as the source-filter model, is crucial. However, the method for obtaining this relation is unclear since the basis waveforms of CWT are usually non-orthogonal.

In Chapter 3, we develop a source separation method that simultaneously satisfies principles (i) and (ii). The source-filter model is incorporated into shifted non-negative matrix factorization (NMF), which takes into account the constant inter-harmonic spacing of the harmonic structure in log-frequency representations. Iterative parameter estimation algorithms with guaranteed convergence are derived based on an optimization principle called the auxiliary function approach. The incorporation of the source-filter model in the CWT domain was confirmed to be effective in monaural audio source separation through an experimental evaluation.

In Chapter 4, a source separation method called harmonic-temporal factor decomposition (HTFD) is proposed, which simultaneously satisfies principles (i) and (iii). HTFD uses a spectrogram model that specifically describes a mathematical form of the spectral leakage of a time domain signal model and takes into account local and global structures of spectrograms of harmonic sounds. A parameter estimation algorithm is derived based on the auxiliary function approach. Experimental results showed the effectiveness of using CWT and the specific description of the spectral leakage.

In Chapter 5, we extend HTFD to satisfy all the principles. The source-filter model is defined in the discrete-time domain and obtaining direct relation of parameters between the source-filter model and the spectrogram model of HTFD is not easy. However, the spectrogram model is derived from the analytic signal model and hence we can associate parameters of the source-filter model with the parameters of the spectrogram model via the signal model. Similarly to Chapter 4, a parameter estimation algorithm is derived based on the auxiliary function approach. We confirmed through an experimental evaluation the effectiveness of simultaneously incorporating the source-filter model and the spectral leakage in the CWT domain.

In Chapter 6, we address phase estimation from a modified magnitude part of a spectrogram (magnitude spectrogram) obtained with CWT to obtain its time domain signal. By taking the spectrogram-aware approach, we introduce a condition that complex spectrograms satisfy and formulate the phase estimation as the problem of minimizing a numerical criterion derived from the condition. Based on the auxiliary function approach, fast phase

estimation algorithms with guaranteed convergence are derived. An experimental evaluation demonstrated that the devised fast algorithms work 75 times faster than a conventional algorithm presented in previous literature while the reconstructed signals obtained with the audio quality of reconstructed signals obtained with the devised algorithms is almost the same as that of the original signals.

In Chapter 7, we present a method of enhancing singing voices in music audio signals using NMF with the L_p norm criterion by focusing on that spectrograms of singing voices can be seen as sparse matrices while spectrograms of accompaniment sounds can be seen as low-rank matrices. An experimental evaluation showed that reasonably good enhancement results were obtained with appropriate choices of p.

In Chapter 8, we develop a system that allows users to edit a music audio signal without using musical scores by replacing the timbres of drum sounds and the frequency characteristics of harmonic sounds with those of another music signal. The present system was confirmed to work well through a subjective experiment.

Acknowledgement (in Japanese)

本論文は，筆者が東京大学大学院情報理工学系研究科システム情報学専攻博士課程に在学中，システム第 5 研究室および連携講座守谷・亀岡研究室（2014 年度より亀岡研究室）で行った研究をまとめたものです．

最初に，博士課程期間中に指導していただき，学位審査員も引き受けていただいた亀岡弘和客員准教授に感謝致します．亀岡先生には，第 2 章から第 7 章の内容に渡り鋭い指摘やアイディアをご教授いただいただけでなく，筆者の言葉足らずな説明や議論に対しても辛抱強く聞いて意図を汲みとっていただきました．亀岡先生のご助力がなければ，本論文を完成させることはできなかったと思います．ミーティングできる時間がかなり限られていても親身に対応してくださり，筆者は大変有意義な 3 年間を過ごすことができました．

指導教官であり学位審査会で主査も務めて頂いた原辰次教授には，本研究を進めるにあたり適切な助言をいただきました．博士課程に入学する際に，修士課程で所属していた研究室が解散するため原先生に相談し，原先生のご専門である制御理論以外の分野の研究を行うことを許していただき，進学先として受け入れてくださったこと感謝しております．俯瞰的な視点に立った学術的な視点での助言は，大局的な視点をもつことの大切さを筆者に気づかせて下さり，本論文を執筆する際に非常に役立ちました．

お忙しい中学位審査員を引き受けて下さり，予備審査会と本審査会において数々の有用な助言や指導を下さった安藤繁教授，猿渡洋教授，真溪歩准教授，国立情報学研究所の小野順貴准教授に感謝致します．審査会での質疑応答などを通して自身の研究の貢献を明確にすることができました．特に第 6 章に関しては予備審査会での議論で問題に対する理解が大幅に進みました．

博士課程 1 年時に，筆者の夏季実習を受け入れていただいた産業技術総合研究所の後藤真考氏，吉井和佳氏（現在，京都大学大学院情報学研究科講師）に感謝致します．第 8 章はこの夏季実習での成果であり，1ヶ月半という短い間でしたが両氏の指導によって完成させることができました．また，この成果を国際会議 ICASSP2014 で発表する際に，後藤氏が代表研究者である CREST から発表の渡航費用を捻出していただき，経済的な面でも助けていた

だきました．

システム第５研究室の津村幸治准教授には，研究室輪講において筆者が思い込みで当然と考えていたことなどを何度も指摘して頂き，本研究をより理論的に強固なものにできる機会を与えてくださったことに感謝しております．また，守谷健弘氏（NTT コミュニケーション科学基礎研究所守谷特別研究室長）には，筆者が博士課程 1，2 年時にお忙しいにもかかわらず研究室輪講に足を運んで頂き，様々な助言をいただきました．東京大学工学系研究科峯松研究室の齋藤大輔助教には，亀岡研究室の計算機環境の整備や保守に関して手助けをしていただき，円滑に研究を進める手助けをしていただきました．

また，博士課程を全体を通して些細な疑問でも議論に付き合って下さった連携講座やシステム第５研究室の皆様に感謝申し上げます．卒業論文で来られた四方紘太郎氏や当時修士課程在学中であった高宗典玄氏には，第４章の研究において最初に動くものを作るなど短期間で筆者の研究の核となる部分の１つを組み上げていただきました．また，博士課程入学時から２年時までに日常的に議論をしてくださった樋口卓哉氏，杉浦亮介氏，門脇健人氏にも感謝申し上げます．亀岡研究室のメンバー全員での議論は博士課程での研究の進め方や方向性に強く影響を与え，筆者は本研究をより魅力的なものにすることができました．亀岡研究室で技術補佐員を勤め，現在システム第５研究室に秘書として勤務しておられる丹治尚子氏には，事務処理だけでなく実験準備などにも協力していただきました．丹治氏のおかげで，筆者は日々の生活や出張時の事務処理だけでなく研究に関しても円滑に進めることができました．この他にも学会で出会った方や研究室に訪問していただいた数多くの方々からの支援や議論なしには，本研究を成し遂げることができませんでした．

本研究の一部は日本学術振興会（JSPS）科研費 26730100，15J0992，科学技術振興機構（JST）CREST「コンテンツ共生社会のための類似度を可知化する情報環境の実現」，立石科学技術振興財団後期国際交流助成，原総合知的通信システム基金海外渡航旅費援助の支援を受けて行われました．

最後に，博士課程まで進学することを許してくれた両親に感謝の意を示して謝辞といたします．

Acknowledgement

I would like to express my best gratitude to all people who have supported me during my Ph.D. course at the University of Tokyo (UT). First of all, I would like to express the deepest appreciation to one of my supervisors Adjunct Associate Professor Hirokazu Kameoka from UT and Nippon Telegraph and Telephone Corporation (NTT), who is one of the members of my thesis committee. He not only gave me insightful ideas and suggestions on Chapter 2 to 7 but also tried understanding my insufficient explanations and statements patiently to read into what I want to say. Without his help, I would never have completed this thesis. Even when he had little time for discussion with me, he kindly spared time for me. His kind and patient attitude let my Ph.D student life be meaningful.

I am deeply grateful to Professor Shinji Hara from UT, who is the other supervisor and the chair of my thesis committee. I am grateful to him for allowing me to specialize in audio signal processing, which is different from his major control theory, to join his laboratory. His holistic and academic suggestions let me notice the necessity of a comprehensive understanding and were very useful for me to write up the thesis.

I express my gratitude to the other members of my thesis committee, Professor Shigeru Ando, Professor Hiroshi Saruwatari, Associate Professor Ayumu Matani from UT and Associate Professor Nobutaka Ono from National Institute of Informatics (NII). They gave me useful comments to clarify the contributions of the thesis and one of the comments let me deeply understand the problem addressed in Chapter 6.

I would like to thank Dr. Masataka Goto and Dr. Kazuyoshi Yoshii (currently a senior lecturer at Kyoto University) from National Institute of Advanced Industrial Science and Technology (AIST) for accepting the summer internship at the first year of my Ph.D. course. Chapter 8 is the result of the internship and without their help, I would never have completed the study in Chapter 8. To make a presentation of the result in ICASSP 2014, I was provided financial support by the CREST project in which the research director is Dr. Goto.

I thank Associate Professor Koji Tsumura for pointing out what I think natural again and again in the laboratory seminars, and giving me opportunities to fortify the theoretical foundation of the thesis. I would also like to thank Dr. Takehiro Moriya from NTT for showing up some of the laboratory seminars despite he being busy and providing me insightful comments in the laboratory seminars. I would like to express my gratitude to Assistant Professor Daisuke Saito from in Minematsu Laboratory of School of Engineering at UT. His help on maintenance of the computing environment of Kameoka Laboratory let the members of the laboratory concentrate on their own study smoothly.

I would like to offer my special thanks to the members of Kameoka Laboratory and Hara/Tsumura Laboratory for discussing many questions with me. Mr. Kotaro Shikata worked on his bachelor's degree study in Kameoka Laboratory to construct the core of Chapter 4, and Mr. Norihiro Takamune (currently a Ph.D. course student in Saruwatari Laboratory at UT) helped him positively. The discussion together with Mr. Takamune, Mr. Takuya Higuchi, Mr. Ryosuke Sugiura, Mr. Kento Kadowaki, who were master students in Kameoka Laboratory, in addition to Prof. Kameoka, affected my research interests strongly and let my study become more attractive. Ms. Naoko Tanji, who was an assistant technical staff in Kameoka Laboratory and is currently a secretary in Hara/Tsumura Laboratory, helped me with her excellent ability of paperwork and cooperated the preparation of the subjective experiment.

This work was supported by JSPS KAKENHI 26730100 and 15J0992, and CREST, Japan Science and Technology Agency. I received financial assistance for my presentations of Chapter 4, 5 and 6 at the international conferences by Tateisi Science and Technology Foundation and the Hara Research Foundation.

Finally I would like to thank my parents for allowing me to proceed to the Ph.D course.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

We discuss through the thesis separation and synthesis of music audio mixtures for audio editing. The schematic illustration of audio editing of our interest is depicted in Fig. 1.1: An audio signal performed by several harmonic musical instruments is separated into "musically meaningful components", listeners edited the components individually, and they can enjoy to listen to the edited signal. Here examples of "musically meaningful components" are sounds of individual pitches and musical instruments. Such a process is applicable to a wide range of music applications such as assistance systems for music composition and arrangement, music players that allow users to edit existing music pieces as per their preferences, and automatic music arrangement systems. To realize such systems, we need techniques for separating music audio signals into musically meaningful components, called audio source separation, with high accuracy because the performance of audio editing greatly depends on the accuracy of audio source separation. In this thesis, we mainly focus on audio source separation specifically for music audio signals, taking into account the characteristics of harmonic musical instruments.

The problem of audio source separation is not so simple to solve because there are many possible separation patterns for a given audio signal without prior information and the problem is inherently ill-posed. To adequately narrow down possible solutions, appropriate cues have continued to be sought in previous studies. If audio signals are recorded with multiple microphones, spatial cues of sources can be used. In contrast, in a situation where audio signals are recorded with a single microphone, we need other cues (e.g. statistical property

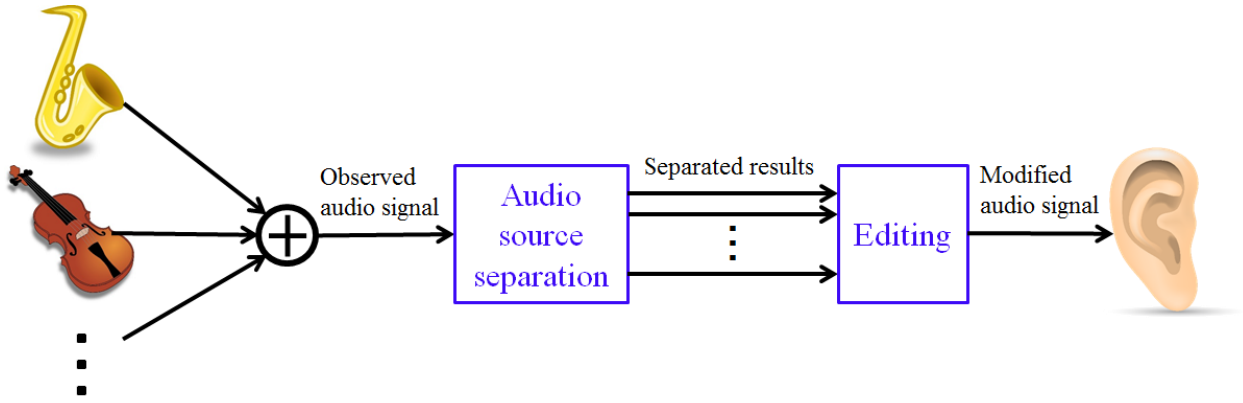


Figure 1.1: Schematic illustration of audio editing in the thesis.

of sources) instead of spatial cues. We will often be faced with such a situation, for example, when there are several different sources very close to each other even if multiple microphones are available and spatial cues are difficult to use, or when an observed signal is originally monaural. Although musical scores can also be used as cues, they are often unavailable and this makes source separation further difficult. In the following, we concentrate on monaural source separation without musical score information.

1.2 Monaural Audio Source Separation

Time-frequency representations, referred to as spectrograms, have been commonly used in many audio signal processing techniques containing monaural source separation because we can utilize various cues in spectrograms. A spectrogram of an audio signal is defined by an inner product of the signal and basis waveforms and how to select basis waveforms determines the frequency resolution of the spectrogram. A representative time-frequency transform is the short-time Fourier transform (STFT), which is frequently used and provides a time-frequency representation with a linearly uniform frequency resolution. However, the frequency resolution is not agreement with the human auditory system and the fundamental frequencies (F_0s) of musical pitches of equal temperament, which are geometrically spaced. Psychoacoustics studies have revealed that the critical bandwidths of the auditory filters increase approximately exponentially with increasing the center frequencies of the filters [3], and the difference threshold of frequency also increases approximately exponentially with the increase of frequency [4]. These suggest that a time-frequency representation with a logarithmic frequency (log-frequency) resolution would be more suitable for source separation

of music audio signals than a time-frequency representation with linear frequency resolution. Such a time-frequency transform is the continuous wavelet transform (CWT) [5], also known as constant-Q transform [6]. CWT provides a time-frequency representation of a time domain signal with a logarithmically uniform frequency resolution. Indeed, recent studies have reported that using the CWT instead of the STFT significantly improves the performances of source separation with multi-channel input [7], multiple F_0 estimation [1,8,9] and singing voice separation [10].

For monaural source separation two main approaches have thus far been adopted, which focus on different structure of spectrograms. One approach is based on the concept of computational auditory scene analysis (CASA). Humans have an excellent ability to concentrate on listening to a specific sound in a situation where there are multiple sound sources. The significant ability to recognize the external environment is referred to as the auditory scene analysis and Bregman investigated its psychological evidences through experiments [11]. The auditory scene analysis process consists of two stages. In the first stage, an incoming audio signal is separated into spectrogram-like segments, each of which should originate from a single source. In the second stage, the segments that are likely to have originated from the same source are grouped into a perceptual structure called an auditory stream. The aim of CASA is to imitate the auditory scene analysis process with computers as our ears do. Some studies tried to formulate the CASA problem as an optimization problem using the grouping cues [1, 12–17]. For example, in [1, 17], an attempt has been made to imitate the auditory scene analysis process by clustering time-frequency components based on a constraint designed according to the auditory grouping cues (such as the harmonicity and the coherences and continuities of amplitude and frequency modulations). This method is called "harmonic-temporal clustering (HTC)." Many other conventional methods can be found in [18, 19].

While the above approach uses strong assumptions on local spectral structures of sources, the other approach instead focuses on global structure in time-frequency representations. In the approach, an observed magnitude spectrogram is interpreted as a non-negative matrix and non-negative matrix factorization (NMF) [20] is applied to it [21]. The idea behind this approach is that the spectrum at each frame is assumed to be represented as a weighted sum of a limited number of common spectral templates. Since the spectral templates and the mixing weights should both be non-negative, this implies that an observed spectrogram is modeled as the product of two non-negative matrices. Thus, factorizing an observed

spectrogram into the product of two non-negative matrices allows us to estimate the unknown spectral templates constituting the observed spectra and decompose the observed spectra into components associated with the estimated spectral templates. Since the introduction to monaural source separation [21], many NMF variants have been presented [8, 22–45]. Some of NMF-based methods have been developed for spectrograms having a logarithmic frequency resolution and utilized a fact that the inter-harmonic spacings of a harmonic structure are constant in log-frequency representations, which was also exploited in [46, 47]. Shifted NMF [37], a.k.a shift-invariant probabilistic latent component analysis (PLCA) [38], is particularly unique in that it takes account of the fact and uses a shifted copy of a spectrum template to represent the spectra of different F_0s. The extension of shifted NMF outperformed the state-of-the-art methods in terms of multipitch F_0 estimation, whose aim is to estimate F_0s of individual sources in a polyphonic music signal, in an international contest of music information retrieval named music information retrieval evaluation exchange (MIREX) [48] in 2013.

The generating processes of music instrument sounds are also very important cues for monaural source separation. The processes can be explained fairly well by the source-filter theory. According to the source-filter theory, an instrument signal is assumed to consist of an excitation signal and a linear filter. The excitation signal is associated with a vibrating object (e.g. a violin string) and varies with pitch. In contrast, the filter represents the resonance structure of the instrument and varies with timbre. Thus, the theory enables us to represent pitch and timbre of an instrument signal separately. The source-filter model was incorporated into NMF in [49], and its variants have been presented in the STFT spectrogram domain with considerable successes [41–44, 50]. If we assume the independence of an excitation signal and a filter as with studies of the source-filter model (e.g. [51–56]), the spectrum of an instrument sound can be described as a product of an excitation spectrum and a filter spectrum in the fast Fourier transform (FFT) domain due to the convolution theorem. However, the CWT is not always orthogonal and so it is unclear how to describe the source-filter representation in the CWT domain.

1.3 Phase Estimation

Many monaural source separation methods mentioned in the above work in the magnitude or power spectrogram domain, and we must be able to construct a time domain signal from

an estimated or modified magnitude spectrogram, in which phase information is missing. To this end, we also address the problem of constructing a time-domain signal by estimating an appropriate phase from a magnitude spectrogram, which we call phase estimation. For STFT spectrograms, a well-known phase estimation algorithm has been presented in [57]. The algorithm consists in iteratively performing the STFT and the inverse STFT and at each iteration, the magnitude part of the updated STFT spectrogram while leaving the phase part unchanged. After two decades, Le Roux *et al.* have thus far proposed a fast algorithm for estimating the phase from a magnitude STFT spectrogram [58,59] by using the fact that the waveforms in the overlapping part of consecutive frames must be consistent. This implies the fact that an STFT spectrogram is a redundant representation when the hop-size is shorter than the frame length and thus it satisfies a certain condition that it corresponds to a time domain signal. We have referred to this condition as *the consistency condition*. The problem of estimating the phase from a magnitude STFT spectrogram can be formulated as an optimization problem of minimizing the consistency criterion that describes how far an arbitrary complex array deviates from this condition. This formulation has provided a new insight into the well-known Griffin's algorithm, allowing us to derive a fast approximate algorithm and give a very intuitive proof of its convergence.

An algorithm for estimating the phase from a magnitude CWT spectrogram has been proposed by Irino *et al.* [60], which consists in iteratively performing the CWT and the inverse CWT. At each iteration, the magnitude part of the updated CWT spectrogram is replaced by the given magnitude CWT spectrogram while leaving the phase part unchanged. However, since the CWT has a large computational cost, Irino's algorithm requires a long processing time for computation, which has been a serious obstacle for its practical uses. Thus, we consider it necessary to develop a faster algorithm. The convergence of the algorithm as well as the computational cost is an important issue. Efficient methods for computing the CWT and the inverse CWT have been recently proposed [61–64]. It may appear that simply carrying out one of these methods for the CWT and inverse CWT steps in Irino's algorithm would reduce the computational cost. However, it is not clear whether the convergence of such an algorithm is guaranteed.

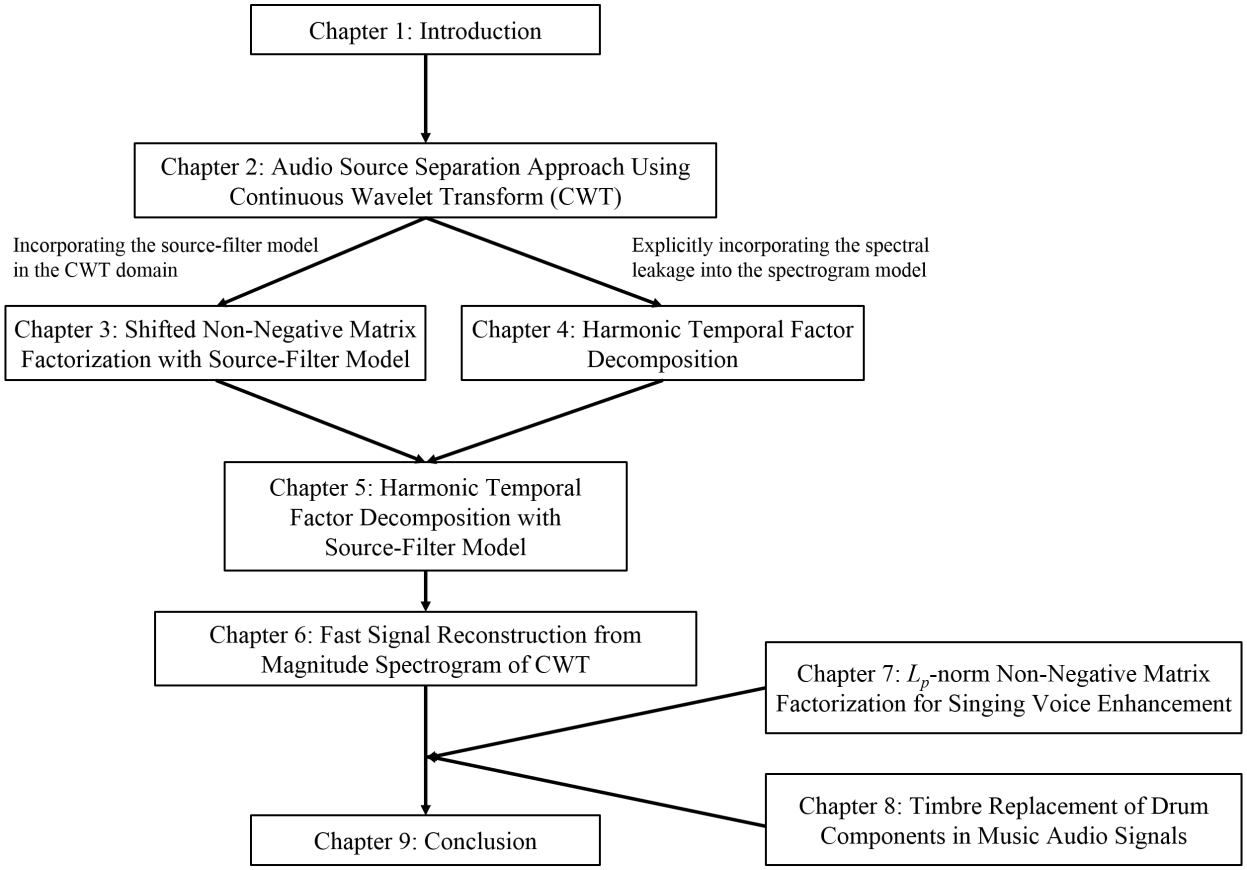


Figure 1.2: Outline of the thesis.

1.4 Objectives and Outline

As the above mentioned, the definition of basis waveforms determines the frequency resolution of spectrograms. To select spectrograms having adequate frequency resolution, it is important to take into account how music audio signals are characterized and how they are represented in the selected spectrogram domain. Therefore, we should take an approach for monaural audio source separation that is aware of spectrograms. To realize the approach, we consider three principles listed in the following.

[P1] Use spectrograms having a log-frequency resolution obtained with the CWT.

[P2] Utilize the source-filter model.

[P3] Take into account the spectral leakage.

On the basis of these principles, we present monaural source separation algorithms.

The organization of the thesis is shown in Fig. 1.2. In Chapter 2, we first discuss a spectrogram-aware approach for monaural audio source separation, describe the reasons

why the three principles are considered, and show technical issues that should be solved to realize the approach. In Chapter 3, we present a monaural audio source separation method that simultaneously satisfies [P1] and [P2] by exploiting the constant inter-harmonic spacings of a harmonic structure in the log-frequency domain and approximately incorporating the source-filter model. In Chapter 4, we propose a monaural audio source separation method simultaneously satisfies [P1] and [P3]. The method uses a CWT spectrogram model that can associate parameters in the CWT domain with parameters in the time domain and derive an efficient algorithm to estimate the parameters from an observed CWT spectrogram. We call the method harmonic-temporal factor decomposition (HTFD). In Chapter 5, we present a monaural source separation method that simultaneously satisfies all the principles by incorporating the parameters of the source-filter model into the CWT spectrogram model of HTFD via the signal model in the time domain. For converting modified magnitude spectrograms obtained with the CWT into time domain signals, we attempt the problem of estimating the phase from a modified magnitude CWT spectrogram, specifically focusing on the convergence and computation speed of algorithms, in Chapter 6. Moreover, we present a method to enhance a singing voice in a monaural music signal in Chapter 7, and develop a system that allows users to replace drum components in a monaural music signal with those in another music signal in Chapter 8. Finally, we give a conclusion of the thesis in Chapter 9.

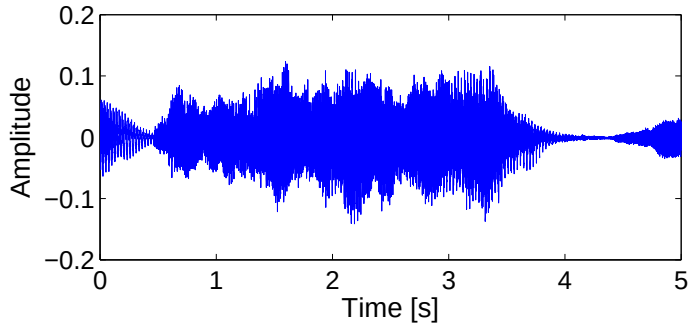
Chapter 2

Spectrogram-Aware Approach Using Continuous Wavelet Transform Representations

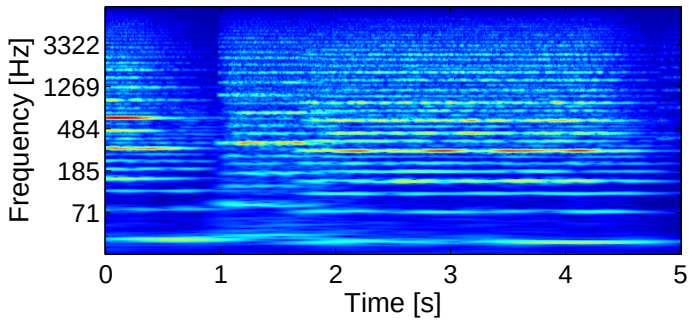
2.1 Time Domain Representation and Time-Frequency Representation

Generally, popular music audio signals contain percussive sounds and harmonic sounds. Separating the two components from music audio signals has been attempted [65–67] and the methods work well. Thus, we here consider audio signals performed by harmonic musical instruments and explore an adequate separation domain that matches for harmonic sounds.

Let us first compare a time domain representation and a time-frequency representation. A time-domain representation is straightforward to compute and additivity of source signals holds. However, the time domain representation allows the sources to cancel each other out, which causes to extremely increase possible separation patterns and makes monaural source separation difficult. Especially for music signals, multiple notes are often performed at the same time. On the other hand, the time-frequency representation such as the magnitude spectrogram obtained with CWT (Fig. 2.1 (b)) is more sparse than the time domain representation (Fig. 2.1 (a)). This is because a time-frequency transform such as STFT and CWT can decompose an audio signal into individual frequency components in each frame even if they are overlapped in the time domain representation. Hence the cancellation of source



(a) Waveform



(b) Spectrogram obtained with CWT

Figure 2.1: Examples of time-domain representation and time-frequency representation.

spectra occur more rarely than that of source signals in the time domain representation. One may think that the effect of the cancellation would be reduced by preparing waveform templates of sources. However, the signals of instrument sounds in real world generally vary at different occurrences due to the irregular behavior of the phases. In contrast, the magnitude or power spectra of the signals are known to be often relatively identical to each other at different occurrences.

Furthermore, if we discard phases from time-frequency representations and assume the additivity of magnitude or power spectra, we do not need to consider the cancellation problem. This assumption has empirically been confirmed to be a good approximation in considerable successes of NMF. These suggest that the magnitude spectrogram would be more suitable for monaural source separation of harmonic audio signals.

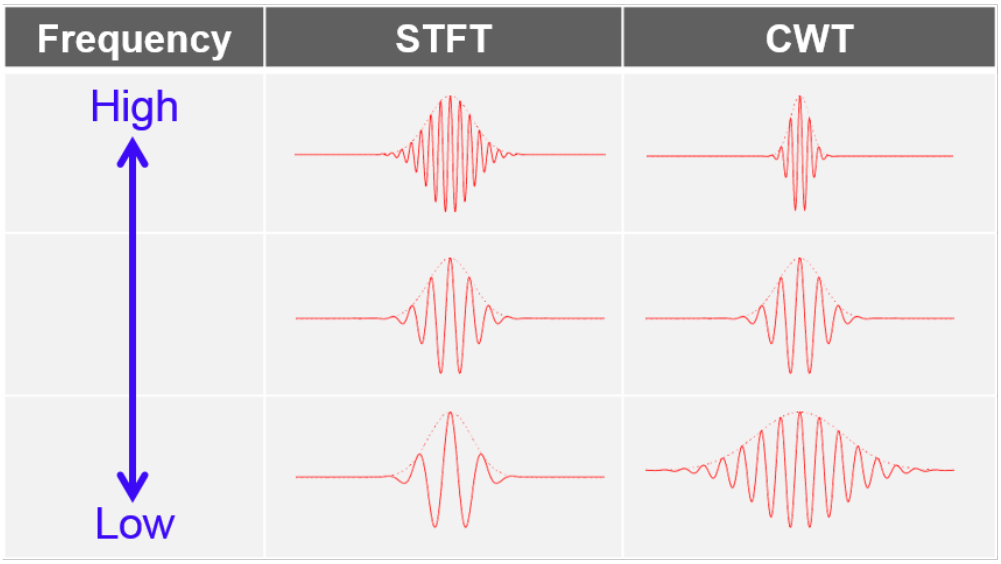


Figure 2.2: Basis waveforms of STFT and CWT with respect to frequency.

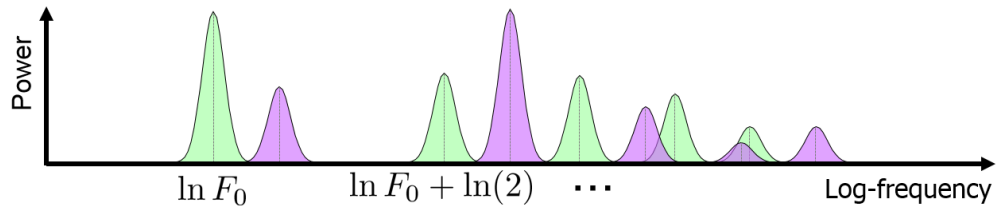
2.2 Spectrogram-Aware Approach for Monaural Audio Source Separation

In this section, we show the reasons why the three principles listed in Sec. 1.4 should be considered.

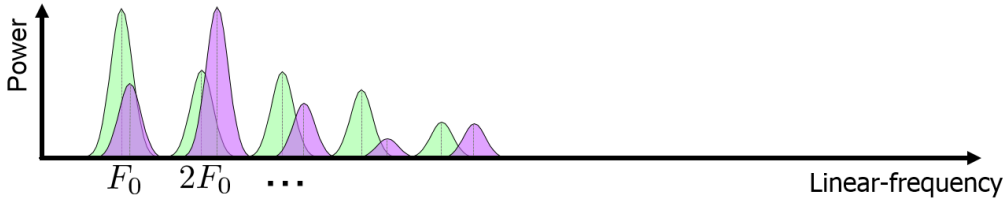
2.2.1 Time-Frequency Transform and Characteristics of Harmonic Audio Signals

Spectrograms are defined by an inner product of audio signal and basis waveforms. How to select basis waveforms determines the characteristics of the time-frequency representation. The schematic illustrations of basis waveforms are shown in Fig. 2.2. For instance, if windowed sinusoids are used as basis waveforms, where the window length is fixed, the resulting spectrograms have a linear frequency resolution, which corresponds to STFT. In contrast, if we use windows whose length is in proportion to periods as basis waveforms, the resulting spectrograms have a log-frequency resolution, which corresponds to CWT.

Now, let us compare a time-frequency representation with linear frequency resolution and one with logarithmic frequency resolution. Spectra of periodic signals have harmonic structure as shown in Fig. 2.3. Here we consider a situation where we would like to separate two signals of different low pitches. Musical pitches in equal temperament are distributed



(a) CWT spectrum.

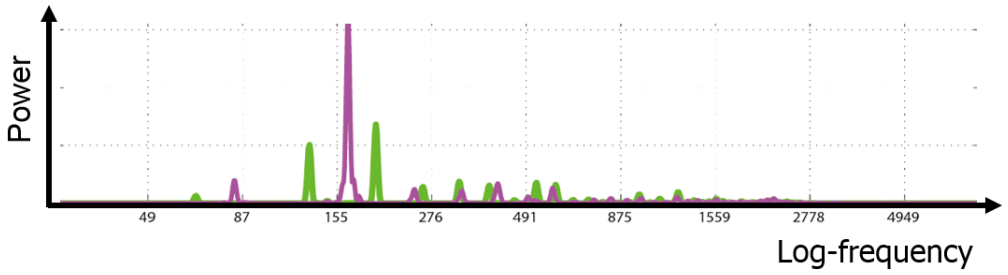


(b) STFT spectrum.

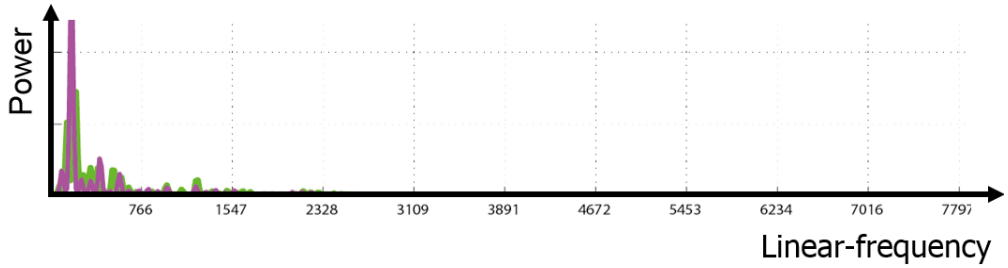
Figure 2.3: Schematic comparison of spectra of a harmonic audio signal obtained with CWT and STFT.

geometrically, and the inter-harmonic spacings of a harmonic structure is constant in the log-frequency domain. This nature is unique in the log-frequency domain, and some methods that takes account of the nature have been presented [37,38,46,47]. Hence the overlap of the F_0 and harmonic components of the two signals is independent of octaves and constant in the CWT domain. On the other hand, the spacings depend on a F_0 in the linear frequency domain, and the lower pitches the two signals have, the more complex the overlap becomes. This makes it difficult to separate the signals using their F_0 and low harmonic components. Although the frequency difference of F_0s is expanded in high harmonics in the STFT domain and the high harmonics can be a useful cue for the separation, music signals in real world often have large energy in low harmonics and small energy in high harmonics. Actual spectra performed by the clarinet are shown in Fig. 2.4 and we can confirm that the high harmonics have small energy. Due to the fact, the cue in the STFT domain is difficult to use. These results suggest that CWT should be more suited than STFT for harmonic signals.

Furthermore, the log-frequency resolution also appears in the human auditory system, particularly in pitch perception [3,4,68,69]. For the above reasons, we consider the principle [P1].



(a) CWT spectrum.



(b) STFT spectrum.

Figure 2.4: Spectra of audio signals performed by the clarinet.

2.2.2 Continuous Wavelet Transform

CWT provides a time-frequency representation with a logarithmic frequency resolution and has been originally presented by [6] and is essentially the same as constant-Q transform [5]. The CWT has a large computational cost, which has been a serious obstacle for its practical uses. However, efficient methods for computing the CWT and the inverse CWT have been recently proposed [61–64], one of which will be described in Section 6.5.1.

The CWT represents a time domain signal as a summation of wavelet basis waveforms, also known as analyzing wavelets, whose periods (the reciprocals of the center frequencies) correspond to a scale parameter. We here consider discretizing the scale parameter such that the center frequencies of the wavelet basis waveforms are equally spaced on a log-frequency scale. Let $l = 0, 1, \cdots, L-1$ and $m = 0, 1, \cdots, M-1$ be the indices of scale and time shift parameters, respectively, where L is the number of the discretized scale parameters and M is the length of an input signal. Given a discrete time domain signal $\boldsymbol{f} = [f_0, f_1, \cdots, f_{M-1}]^\top \in \mathcal{F} := \{\boldsymbol{f}; \boldsymbol{f} \in \mathbb{C}^M, \sum_t f_t = 0\}$, the component of a CWT spectrogram associated with scale

$a_l > 0$, arranged as $\boldsymbol{s}_l = [s_{l,0}, s_{l,1}, \cdots, s_{l,M-1}]^\top$, is defined as

$$\boldsymbol{s}_l = W_l \boldsymbol{f}, \tag{2.1}$$

$$W_l := \begin{bmatrix} \psi^*_{l,0} & \psi^*_{l,M-1} & \cdots & \psi^*_{l,1} \\ \psi^*_{l,1} & \psi^*_{l,0} & \cdots & \psi^*_{l,2} \\ \vdots & \vdots & \ddots & \vdots \\ \psi^*_{l,M-1} & \psi^*_{l,M-2} & \cdots & \psi^*_{l,0} \end{bmatrix}. \tag{2.2}$$

Here $\psi^*_{l,m}$ is the complex conjugate of the wavelet basis waveform $\psi_{l,m} := \psi(t\Delta/a_l)/a_l$, where Δ denotes the sampling period of the input signal, $\psi(t\Delta)$ is a mother wavelet satisfying the admissibility condition. Each row of W_l contains the wavelet basis waveform of scale a_l with a different time shift parameter. Then, the CWT spectrogram $\boldsymbol{s} = [\boldsymbol{s}_0^\top, \boldsymbol{s}_1^\top, \cdots, \boldsymbol{s}_{L-1}^\top]^\top$ is given as

$$\boldsymbol{s} = W \boldsymbol{f}, \tag{2.3}$$

where W denotes the CWT matrix, defined as

$$W = [W_0^\top, W_1^\top, \cdots, W_{L-1}^\top]^\top. \tag{2.4}$$

Whether the inverse CWT of $W\boldsymbol{f}$ equals to $\boldsymbol{f}$ for all $\boldsymbol{f} \in \mathcal{F}$ depends on W. For simplicity, we hereafter assume that the equality holds. It is important to note that the following discussion is valid if the equality does not hold.

The inverse CWT can be defined by the pseudo inverse of W, defined as W^+, and the inverse of $\boldsymbol{s}$ is given as $W^+\boldsymbol{s}$. This implicitly means that the inverse CWT of $\boldsymbol{s}$ is the solution to the following minimization problem:

$$\underset{\tilde{\boldsymbol{f}} \in \mathcal{F}}{\operatorname{argmin}} \|\boldsymbol{s} - W\tilde{\boldsymbol{f}}\|_2^2, \tag{2.5}$$

where $\|\boldsymbol{s}\|_2$ denotes the L^2 norm of $\boldsymbol{s}$.

2.2.3 Low-Rank Approximation of Spectrograms

Same notes and instruments tend to appear in a music audio signal again and again. From this tendency, the spectrograms of music audio signals tend to have low-rank structures. NMF has introduced the structures explicitly with considerable success in monaural audio

source separation. The idea behind NMF is that the spectrum at each frame is assumed to be represented as a weighted sum of a limited number of common spectral templates. Since the spectral templates and the mixing weights should both be non-negative, this implies that an observed spectrogram is modeled as the product of two non-negative matrices. Thus, factorizing an observed spectrogram into the product of two non-negative matrices allows us to estimate the unknown spectral templates constituting the observed spectra and decompose the observed spectra into components associated with the estimated spectral templates.

To examine how different the low-rankness changes with different time-frequency representations and music genres, we conducted an experiment using the RWC music genre database [2]. We used a measure to evaluate the rank of spectrograms as a nuclear norm of a magnitude spectrogram normalized with a Frobenius norm. CWT spectrograms were computed with the fast approximate CWT algorithm [61, 62] using the log-normal wavelet [1], which has a Gaussian shape in the log-frequency domain. We set a parameter corresponding to a standard deviation of the Gaussian as one fifth of a semitone interval, and the center frequencies of the CWT ranged 27.5 to 7902 Hz with 100/3 cent interval. STFT spectrograms were computed with a Gaussian window of 64 ms and a hopsize of 10 ms. We randomly extract parts of each music signal with a duration of 10 s and compute the measures. We repeated the operation ten times and calculated the average measures and standard errors.

The results are summarized with respect to main categories defined in the RWC music genre database in Fig. 2.5. (The results with respect to subcategories defined in the database are displayed in Fig. A.1.) We can confirm that the ranks of the CWT spectrograms were more on average than that of the STFT spectrograms for Latin songs while the CWT and STFT spectrograms have similar ranks averagely for classical musical pieces. This result suggests that percussive sounds deteriorate the accuracy of audio source separation in the CWT domain. For Jazz musical pieces, the ranks of the CWT spectrograms are higher than the STFT spectrograms. This may be because the musical pieces contain the bass solo part and the CWT spectrograms captured the spectral dynamics at very low F_0s.

2.2.4 Generating Processes of Musical Instrument Sounds

The generating processes of many musical instrument sounds in real world can be explained fairly well by the source-filter theory. With the theory, an instrument signal is assumed to consist of an excitation signal and a linear filter. The excitation signal is associated with a

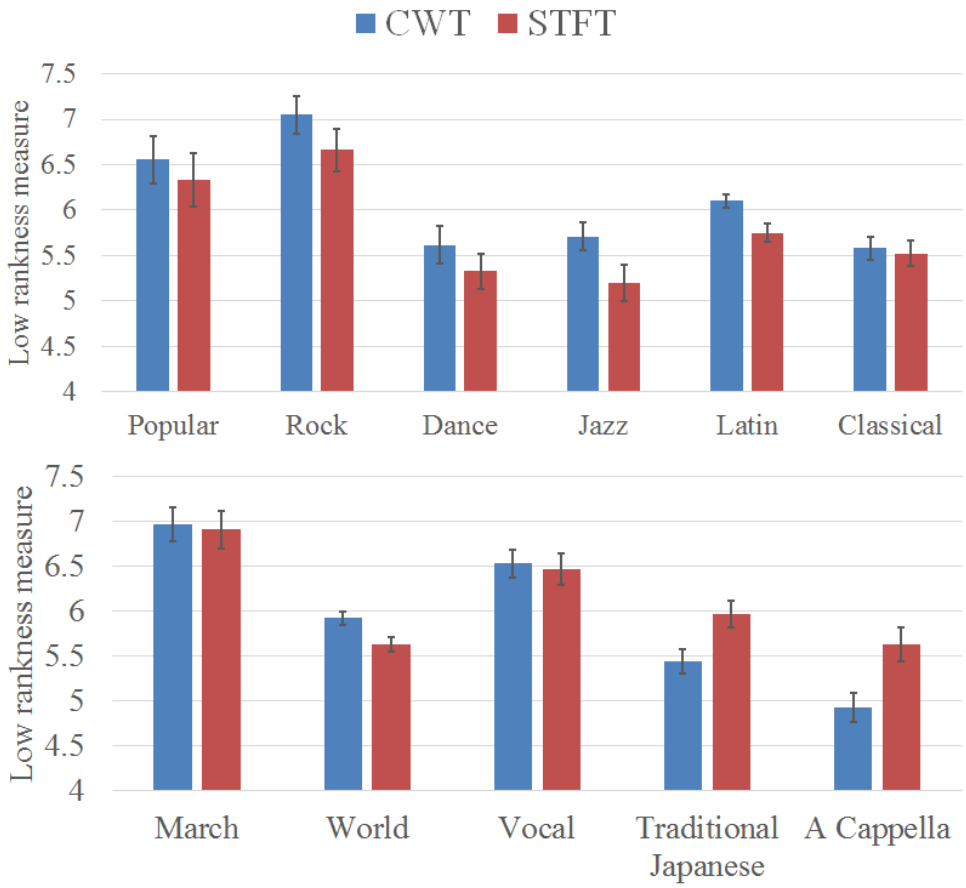


Figure 2.5: Comparison of STFT and CWT spectrograms in low-rankness for main categories of music genres.

vibrating object (e.g. a violin string) and varies with pitch. In contrast, the filter represents the resonance structure of the instrument and varies with timbre. Thus, the theory enables us to represent pitch and timbre components of an instrument signal separately and to make assumptions regarding the components individually. If we can obtain adequate parameters of the source-filter model, estimated spectral shapes may be suppressed to be far different from spectra of real musical instruments. For the above reasons, we consider the principle [P2].

Let us assume that a discrete-time audio signal produced by a musical instrument within a short-time segment is an output of a Pth order autoregressive (AR) process, i.e. an Pth order all-pole system. That is, if we denote the signal by $f[i]$ for $i = 0, 1, \cdots, I-1$, $f[i]$ can be described as

$$\beta[p]f[i] = \sum_{p=1}^{P} \beta[p]f[i-p] + \epsilon[i] \tag{2.6}$$

where i, $\epsilon[i]$, and $\beta[p]$ $(p = 0, 1, \cdots, P)$ denote the discrete-time index, an excitation signal

and the AR coefficients, respectively. By abuse of notation, we understand $f[i] = 0$ for $i \neq 0, 1, \cdots, I-1$. As can be seen from Eq. (2.6), each $f[i]$ can be predicted by a linear combination of the P latest samples, and thus an audio compression method using this representation is called linear predictive coding (LPC) [51].

2.2.5 Spectral Leakage

Finally, we describe reasons of the principle [P3]. Any time-frequency representation has spectral leakage, which means that energy of an input signal spreads in the frequency direction even if the input signal is a sinusoid of infinite length. The choice of the basis waveforms determines the shape of the spectral leakage and thus there are certain constraints of the neighboring time-frequency components. Here, let us consider the case where harmonic components of different audio sources are close to each other. In this case, if the concrete shape and functions of the individual components is known in advance, they can be useful cues for the separation. Thus, identifying the shape of the spectral leakage can be valid for separating adjacent F_0 and harmonic components of different sources.

2.3 Issues to Realize the Spectrogram-Aware Approach

All the principles are important for monaural audio source separation with high accuracy, but methods that satisfy any one of the principles have been presented in previous literature as mentioned in the above. One of the reasons why methods satisfying all the principles have yet been presented is that source separation in the CWT domain have different difficulty from that in the STFT domain. For example, if we assume the independence of an excitation signal and a filter as with many studies of the source-filter model [51–56], the spectrum of an instrument sound can be described as a product of an excitation spectrum and a filter spectrum in the FFT domain due to the convolution theorem. Furthermore, if each time slice of STFT spectrograms is assumed to be independent, the source-filter model can be incorporated in the STFT domain as in the FFT domain. In contrast, since basis waveforms of CWT are not always orthogonal rather non-orthogonal, the exact source-filter representation in the CWT domain is unclear and would be different from that in the STFT domain. This makes it not easy to incorporate useful time-domain models and cues in

the time domain into CWT-domain representations. The non-orthogonality of the basis waveforms is the main obstacle to develop a monaural audio source separation approach in the CWT domain.

On the basis of the above discussion, we attempt the following issues to realize the spectrogram-aware approach:

[I1] How can we incorporate the source-filter model in the CWT domain ? (Chapter 3)

[I2] How can we describe the spectral leakage in the CWT domain ? (Chapter 4)

[I3] How can we simultaneously incorporate the source-filter model and the spectral leakage in the CWT domain ? (Chapter 5)

Chapter 3

Shifted Non-Negative Matrix Factorization with Source-Filter Model

3.1 Chapter Overview

This chapter proposes an extension of NMF, which combines the shifted NMF model with the source-filter model. Shifted NMF was proposed as a powerful approach for monaural source separation and multiple F_0 estimation, which is particularly unique in that it takes account of the constant inter-harmonic spacings of a harmonic structure in log-frequency representations and uses a shifted copy of a spectrum template to represent the spectra of different F_0s. However, for those sounds that follow the source-filter model, this assumption does not hold in reality, since the filter spectra are usually invariant under F_0 changes. A more reasonable way to represent the spectrum of a different F_0 is to use a shifted copy of a harmonic structure template as the excitation spectrum and keep the filter spectrum fixed. Thus, we can describe the spectrogram of a mixture signal as the sum of the products between the shifted copies of excitation spectrum templates and filter spectrum templates. Furthermore, the time course of filter spectra represents the dynamics of the timbre, which is important for characterizing the feature of an instrument sound. Thus, we further incorporate the non-negative matrix factor deconvolution (NMFD) model into the above model to describe the filter spectrogram. We derive a computationally efficient and convergence-guaranteed algorithm for estimating the unknown parameters of the constructed model based

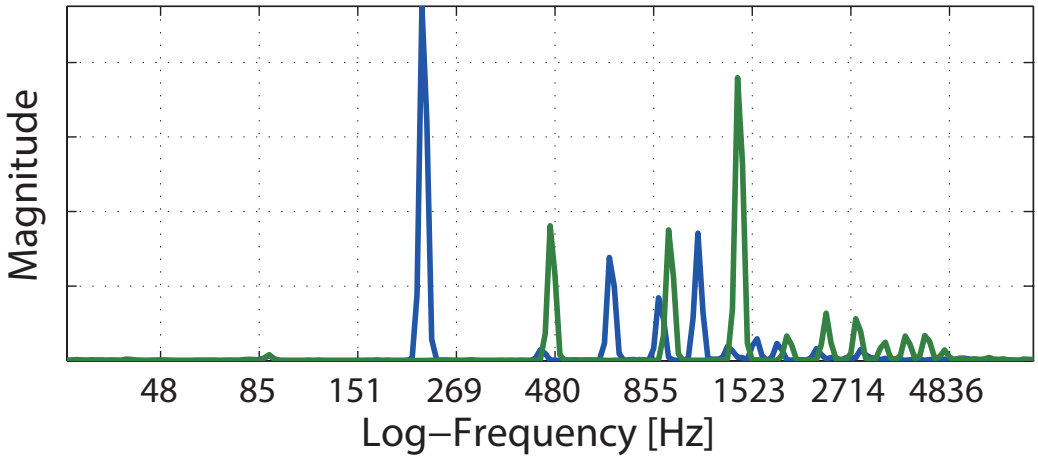


Figure 3.1: Two spectra of clarinet sounds at different pitches.

on the auxiliary function approach. Experimental results revealed that the proposed method outperformed shifted NMF in terms of the source separation accuracy.

3.2 Introduction

One major approach to monaural source separation involves applying NMF to an observed magnitude (or power) spectrogram interpreted as a non-negative matrix [21]. While many variants and extensions of NMF were developed based on spectrograms with linear frequency resolution such as the STFT, spectrograms with log-frequency resolution such as the CWT were chosen to utilize a fact that F_0s of pitches in music are geometrically spaced in some studies [8, 32, 33, 36–38]. Specifically, shifted NMF [37], a.k.a shift-invariant PLCA [38], has been often used in monaural source separation and multiple F_0 estimation with considerable success [39, 40, 70]. In this method, all notes performed by a specific instrument are assumed to have a same harmonic structure in the entire musical piece. With this assumption, each basis spectrum is associated with an individual instrument, and all notes performed by the instrument can be represented by translating the basis spectrum up or down on the log-frequency axis. The shift-invariant property has been also utilized in [47].

However, the above assumption is not always a valid approximation in real situations, specifically for sounds at separate pitches. In fact, as we can see from Fig. 3.1, the two spectra are apparently far different in relative energy of partials with each other. Although previous studies limited a possible translation range [37] and associated multiple basis spectra with each instrument [39, 40, 70] to cope with the problem, these increase the number of

parameters excessively and makes the estimation less reliable.

To explore more compact representations, we focus on a fact that the generating processes of instrument sounds can be explained fairly well by the source-filter theory. According to the theory, an audio signal of an instrument is modeled by an excitation signal, produced by a vibrating object (e.g. a violin string), and a linear filter, representing the resonance structure of the instrument. The excitation signal varies with pitch, whereas the filter varies with timbre. This suggests that the shift-invariant property should be imposed only on excitation spectra of instrument sounds.

Motivated by the above, we propose a new method of separating individual instrument sounds from a mixture audio signal by incorporating the source-filter model into shifted NMF. While many NMF variants containing the source-filter [41–44, 50] used STFT spectrograms, the proposed method is developed based on spectrograms with log-frequency resolution to utilize the shift-invariant property. We hereafter consider spectrograms obtained with the CWT such that its center frequencies are geometrically spaced. We first describe spectrograms of individual instrument sounds separately, assuming that the excitation spectra are shift-invariant. Furthermore, we introduce time-extended filter models to represent temporal dynamics of timbre, which would be useful for monaural source separation. Second, we formulate the NMF as a minimization problem and derive a convergence-guaranteed algorithm that consists of multiplicative update equations based on an optimization principle called the auxiliary function approach [71–73]. Finally, we evaluate the impact of the incorporation of the source-filter model in source separation accuracy using recorded music signals.

3.3 Incorporating Source-Filter Model into Shifted Non-Negative Matrix Factorization

3.3.1 Spectrogram Model of Single Instrument Sound

Let us define indexes of log-frequency and time by $l = 0, \cdots, L-1$ and $m = 0, \cdots, M-1$, respectively. We consider a CWT spectrogram of an audio signal that follows the source-filter model. If we can assume that the frequency response of a filter (a filter spectrum) is constant at each subband, the spectrum obtained with CWT can be described by the product of an excitation spectrum and a filter spectrum as in the STFT domain. Since the inter-harmonic

spacings of a harmonic structure are constant in the log-frequency domain, a shifted copy of a excitation spectrum template can be used to represent the excitation spectra of different F_0s, as with shifted NMF [37] and shift-invariant PLCA [38]. The spectrogram $\tilde{X}^{(\mathrm{ex})}_{l,m,k} \geq 0$ of source excitation $k(= 0, \cdots, K-1)$ is modeled as a convolution of a excitation spectrum template $S_{k,l} \geq 0$ with time-varying gains $U^{(\mathrm{ex})}_{k,p,m} \geq 0$, i.e. $\tilde{X}^{(\mathrm{ex})}_{l,m,k} = \sum_{p \in \mathcal{P}} S_{k,l-p} U^{(\mathrm{ex})}_{k,p,m}$, where p is the frequency shift index and $\mathcal{P}$ is the set of possible frequency shifts. By abuse of notation, we understand that $S_{k,l-p} = 0$ unless $0 \leq l - p \leq L - 1$.

On the other hand, we describe the filter spectrogram in a similar manner to NMFD [27] and NMF-2D [8] to capture the dynamics of the timbre, which is important for characterizing the feature of an instrument sound. The spectrogram $\tilde{X}^{(\mathrm{filt})}_{l,m,r} \geq 0$ of filter $r(= 0, \cdots, R-1)$ is represented by a time convolution of a time-frequency profile $F_{r,l,\tau} \geq 0$ with time-varying gains $U^{(\mathrm{filt})}_{r,m} \geq 0$, i.e. $\tilde{X}^{(\mathrm{filt})}_{l,m,r} = \sum_{\tau=0}^{M^{(\mathrm{tap})}-1} F_{r,l,\tau} U^{(\mathrm{filt})}_{r,m-\tau}$, where $\tau = 0, \cdots, M^{(\mathrm{tap})} - 1$ is the time shift index and $M^{(\mathrm{tap})}$ is the tap size of the time-frequency profiles. By abuse of notation, we understand that $U^{(\mathrm{filt})}_{r,m-\tau} = 0$ unless $0 \leq m - \tau \leq M - 1$.

As we want the magnitude spectra of filters to be smooth and non-negative in the log-frequency domain, we parameterize $F_{r,l,\tau}$ by N envelope kernels $G_{l,n} \geq 0$ and their mixture weights $W_{r,n,\tau} \geq 0$ that satisfies $\sum_{n,\tau} W_{r,n,\tau} = 1$:

$$F_{r,l,\tau} = \sum_{n} W_{r,n,\tau} G_{l,n}, \tag{3.1}$$

$$G_{l,n} := \frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(\omega_l - \rho_n)^2}{2\nu^2}} \tag{3.2}$$

where $n = 0, \cdots, N-1$ is the index of envelope kernel and $\omega_l \in (0, \pi]$ is the normalized angular frequency corresponding to the lth log-frequency. The kernel $G_{l,n}$ for $n \geq 1$ is identical to a normal distribution of a normalized angular frequency with mean ρ_n and variance ν^2.

Multiple excitation and filter spectra can be used for an instrument to describe complex spectral changes, but we hereafter assign an excitation spectrum and a filter spectrum to each instrument for the simplicity. By putting $U_{k,r,p,m-\tau} = U^{(\mathrm{ex})}_{k,p,m} U^{(\mathrm{filt})}_{k,r,p,m-\tau}$ and treating $U_{k,r,p,m-\tau}$ itself as a parameter, the spectrogram of an instrument sound associated with source excitation k and filter r can be written as

$$\tilde{X}_{l,m,k,r} = \sum_{p,\tau} F_{r,l,\tau} S_{k,l-p} U_{k,r,p,m-\tau}. \tag{3.3}$$

Assuming the additivity of magnitude spectrograms as with conventional NMFs, the ob-

Table 3.1: List of indices for the proposed model.

Notation	Meaning
$l = 0, \cdots, L-1$	Log-frequency index
$m = 0, \cdots, M-1$	Time index
$k = 0, \cdots, K-1$	Source excitation index
$r = 0, \cdots, R-1$	Filter index
$p \in \mathcal{P}$	Frequency shift index ($\mathcal{P}$ is the set of possible frequency shifts.)
$n = 0, \cdots, N-1$	Index of kernels constituting filter spectrograms
$\tau = 0, \cdots, M^{(\mathrm{tap})}-1$	Time shift index of filter spectrograms

served spectrogram can be represented as

$$X_{l,m} = \sum_{k,r} \tilde{X}_{l,m,k,r}. \tag{3.4}$$

To avoid the indeterminacy in scaling, we put $\sum_l S_{k,l} = 1$ for all k. The indexes used in this model is summarized in Tab. 3.1.

Although a model similar to the above has been mentioned in [74], the temporal dynamics was not incorporated into the source-filter model in the literature. Any experimental evaluation was not given and the incorporation of the source-filter model into shifted NMF has yet been validated. We will thus confirm the efficacy of the incorporation of the source-filter model in Sec. 3.5.

3.3.2 Formulation

For a given magnitude spectrogram $Y := \{Y_{l,m}\}_{l,m}$, we would like to find the parameters $S := \{S_{k,p}\}_{k,p}$, $W := \{W_{r,n,\tau}\}_{r,n,\tau}$ and $U := \{U_{k,r,l,m}\}_{k,r,l,m}$ of the proposed model such that minimizes

$$\mathcal{L}_*(S, W, U) = \sum_{l,m} D_*(Y_{l,m}||X_{l,m}) + \mathcal{R}_*(U). \tag{3.5}$$

The first term of Eq. (3.5) is a goodness-of-fit measure between Y and $X := \{X_{l,m}\}_{l,m}$. How to define the measure is very important since it corresponds to an assumption to the

statistical nature of observed data. If we define D_{I} as the generalized Kullback-Leibler divergence (a.k.a I divergence), it implicitly assumes that $Y_{l,m}$ follows a Poisson distribution with mean $X_{l,m}$:

$$D_{\mathrm{I}}(Y_{l,m}||X_{l,m}) = Y_{l,m} \ln \frac{Y_{l,m}}{X_{l,m}} - Y_{l,m} + X_{l,m}. \tag{3.6}$$

From this fact, it is known that minimizing $\sum_{l,m} D_{\mathrm{I}}(Y_{l,m}||X_{l,m})$ with respect to $X_{l,m}$ amounts to the maximum likelihood estimation of $X_{l,m}$. This measure is frequently used in conventional NMF algorithms and has been confirmed to work well for audio source separation empirically. Another commonly-used measure is the Itakura-Saito divergence D_{IS}:

$$D_{\mathrm{IS}}(Y^2_{l,m}||X_{l,m}) = \frac{Y^2_{l,m}}{X_{l,m}} - \ln \frac{Y^2_{l,m}}{X_{l,m}} - 1. \tag{3.7}$$

This corresponds to the assumption that an observed complex spectrogram follows a circularly-symmetric complex normal distribution with mean zero and variance $X_{l,m}$, in which $X_{l,m}$ can be interpreted as a model of a power spectral density of the observed signal.

The second term $\mathcal{R}_*(U)$ is a regularizer for U. In popular and classical western music, the number of pitches occurred in a musical piece and the number of times each note is performed are usually limited, and so inducing the sparsity of U would facilitate the source separation. To reflect it, we can design the regularizer in analogy to the Bayesian modeling. The conjugate prior of the Poisson distribution is a gamma distribution $\mathrm{Gam}(x;\alpha,\beta) \propto x^{\alpha-1}e^{-\beta x}$ and thus we design the regularizer $\mathcal{R}_{\mathrm{I}}(U)$ for D_{I} as

$$\mathcal{R}_{\mathrm{I}}(U) = \sum_{k,r,p,m} \left\{ -(\alpha^{(\mathrm{I})} - 1) \ln U_{k,r,p,m} + \beta^{(\mathrm{I})} U_{k,r,p,m} \right\}, \tag{3.8}$$

where $\alpha^{(\mathrm{I})} > 0$ and $\beta^{(\mathrm{I})} > 0$ are associated with the shape and rate parameters of a gamma distribution, respectively. Similarly, the conjugate prior of a circularly-symmetric complex normal distribution with known mean and unknown variance is an inverse gamma distribution $\mathrm{InvGam}(y;\alpha,\beta) \propto x^{-\alpha-1}e^{-\beta/x}$, and we design the regularizer for D_{IS} as

$$\mathcal{R}_{\mathrm{IS}}(U) = \sum_{k,r,p,m} \left\{ (\alpha^{(\mathrm{IS})} + 1) \ln U_{k,r,p,m} + \frac{\beta^{(\mathrm{IS})}}{U_{k,r,p,m}} \right\}, \tag{3.9}$$

where $\alpha^{(\mathrm{IS})} > 0$ and $\beta^{(\mathrm{IS})} > 0$ are associated with the shape and scale parameters of an inverse gamma distribution, respectively. The less $\alpha^{(\mathrm{I})}$ ($\alpha^{(\mathrm{IS})}$), the more sparse U tends to become.

3.4 Parameter Estimation Algorithms Based on Auxiliary Function Approach

3.4.1 Parameter Estimation Algorithm for Proposed Model with I Divergence Criterion

We first derive a parameter estimation algorithm for the I divergence. Since $\mathcal{L}_{\mathrm{I}}(S,W,U)$ involves summations over k,r,p,τ and n in the logarithmic function, the current minimization problem is difficult to solve analytically. However, we can develop a computationally efficient algorithm for finding a locally optimal solution based on the auxiliary function approach [71–73]. The first step to apply the auxiliary function approach, is to define an upper bound function for the objective function $\mathcal{L}(S,W,U)$, arranged as $\mathcal{L}^{+}(S,W,U,\Lambda)$, such that $\mathcal{L}(S,W,U)=\min_{\Lambda}\mathcal{L}^{+}(S,W,U,\Lambda)$. We call Λ an auxiliary variable and $\mathcal{L}^{+}(S,W,U,\Lambda)$ an auxiliary function. If we can construct $\mathcal{L}^{+}(S,W,U,\Lambda)$, $\mathcal{L}(S,W,U)$ is non-increasing under the updates $\{S,W,U\}\leftarrow\operatorname*{argmin}_{S,W,U}\mathcal{L}^{+}(S,W,U,\Lambda)$ and $\Lambda\leftarrow\operatorname*{argmin}_{\Lambda}\mathcal{L}^{+}(S,W,U,\Lambda)$.

Since the logarithmic function is a concave function, we can obtain an upper bound function by invoking the Jensen's inequality:

$$\begin{aligned}-Y_{l,m}\ln X_{l,m}\leq&-Y_{l,m}\sum_{k,r,p,\tau,n}\lambda_{l,m,k,r,p,\tau,n}\left(\ln S_{k,l-p}+\ln W_{r,n,\tau}\right.\\&\left.+\ln G_{l,n}+\ln U_{k,r,p,m-\tau}-\ln\lambda_{l,m,k,r,p,\tau,n}\right)\end{aligned}\tag{3.10}$$

where $\lambda_{l,m,k,r,p,\tau,n}\geq 0$ is an auxiliary variable such that $\sum_{k,r,p,\tau,n}\lambda_{l,m,k,r,p,\tau,n}=1$ for all l and m. The equality holds if and only if

$$\lambda_{l,m,k,r,p,\tau,n}=\frac{S_{k,l-p}W_{r,n,\tau}G_{l,n}U_{k,r,p,m-\tau}}{X_{l,m}}.\tag{3.11}$$

The auxiliary function can thus be written as

$$\begin{aligned}\mathcal{L}_{\mathrm{I}}^{+}(S,W,U,\Lambda)\underset{c}{=}&-\sum_{l,m}Y_{l,m}\sum_{k,r,p,\tau,n}\lambda_{l,m,k,r,p,\tau,n}\left(\ln S_{k,l-p}+\ln W_{r,n,\tau}+\ln U_{k,r,p,m-\tau}\right.\\&\left.-\ln\lambda_{l,m,k,r,p,\tau,n}\right)+\sum_{l,m}X_{l,m}+\sum_{k,r,p,m}\left\{(\alpha^{(\mathrm{I})}-1)\ln U_{k,r,p,m}-\beta^{(\mathrm{I})}U_{k,r,p,m}\right\}\end{aligned}\tag{3.12}$$

where $\Lambda:=\{\lambda_{l,m,k,r,p,\tau,n}\}_{l,m,k,r,p,\tau,n}$. By setting the partial derivatives of $\mathcal{L}^{+}(S,W,U,\Lambda)$ with respect to S, W and U at zeros and substituting Eqs. (3.11) into Λ, we can derive the

following update equations:

$$S_{k,l'} \leftarrow S_{k,l'} \frac{\sum_{l,m} \frac{Y_{l,m}}{X_{l,m}} \sum_{r,\tau} F_{r,l,\tau} U_{k,r,l-l',m-\tau}}{\sum_{l,m,r,\tau} F_{r,l,\tau} U_{k,r,l-l',m-\tau}}, \quad S_{k,l'} \leftarrow \frac{S_{k,l'}}{\sum_{l} S_{k,l}}, \tag{3.13}$$

$$W_{r,n,\tau} \leftarrow W_{r,n,\tau} \frac{\sum_{l,m} \frac{Y_{l,m}}{X_{l,m}} \sum_{p,k} G_{l,n} S_{k,l-p} U_{k,r,p,m-\tau}}{\sum_{l,m,p,k} G_{l,n} S_{k,l-p} U_{k,r,p,m-\tau}}, \quad W_{r,n,\tau} \leftarrow \frac{W_{r,n,\tau}}{\sum_{n',\tau'} W_{r,n',\tau'}}, \tag{3.14}$$

$$U_{k,r,p,m'} \leftarrow \frac{U_{k,r,p,m'} \sum_{l,m} \frac{Y_{l,m}}{X_{l,m}} \sum_{p,k} F_{r,l,m-m'} S_{k,l-p} + \alpha^{(\mathrm{I})} - 1}{\sum_{l,m,p,k} F_{r,l,m-m'} S_{k,l-p} + \beta^{(\mathrm{I})}}, \tag{3.15}$$

Both second update rules of Eqs. (3.13) and (3.14) are to normalize S and U. It is important to note that once the initial values of W and S are set to be non-negative, the multiplicative update equations ensures the non-negativity of the entries of W and S. Since the non-negativity of U does not hold, we can ensure it by simply performing $U_{k,r,p,m} \leftarrow \max\{0, U_{k,r,p,m}\}$ at each update.

One may think that the update equations contain time-consuming convolutions and correlations and would require a long computation time. However, we can invoke the fast Fourier transform (FFT) to calculate the convolutions and correlations, and they are not time-consuming in practice. For example, Eq (3.13) contains a convolution in τ (correlation in l) and a naive calculation of the convolution is of $\mathcal{O}(M^{(\mathrm{tap})}T)$ ($\mathcal{O}(LP)$), whereas the FFT reduces the complexity to $\mathcal{O}((T+M^{(\mathrm{tap})})\ln(T+M^{(\mathrm{tap})}))$ ($\mathcal{O}((L+P)\ln(L+P))$, respectively).

3.4.2 Parameter Estimation Algorithm for Proposed Model with IS-Divergence Criterion

Similarly to the above, we can construct an auxiliary function for the IS divergence, using two inequalities. The logarithm function is a concave function and we can derive

$$\ln X_{l,m} \leq \frac{X_{l,m} - C_{l,m}}{C_{l,m}} + \ln C_{l,m}, \tag{3.16}$$

where $C_{l,m}$ is an auxiliary variable and the equality holds if and only if $C_{l,m} = X_{l,m}$. Since $1/x$ is a convex function of x, we can invoke Jensen's inequality as

$$\frac{Y_{l,m}^2}{X_{l,m}} \leq \sum_{k,r,p,\tau,n} \frac{Y_{l,m}^2 \phi_{l,m,k,r,p,\tau,n}^2}{S_{l-p,k} G_{l,n} W_{r,n,\tau} U_{k,r,p,m-\tau}}, \tag{3.17}$$

where $\phi_{l,m,k,r,p,\tau,n} \geq 0$ is another auxiliary variable satisfying $\sum_{k,r,p,\tau,n} \phi_{l,m,k,r,p,\tau,n} = 1$ for all l and m. The equality of this inequality holds if and only if

$$\phi_{l,m,k,r,p,\tau,n} = \frac{S_{k,l-p} W_{r,n,\tau} G_{l,n} U_{k,r,p,m-\tau}}{X_{l,m}} \tag{3.18}$$

Hence the auxiliary function can be derived as

$$\begin{aligned}\mathcal{L}^{+}_{\mathrm{IS}}(S,W,U,C,\Phi) \underset{c}{=} &\sum_{l,m} \left(\sum_{k,r,p,\tau,n} \frac{Y^2_{l,m} \phi^2_{l,m,k,r,p,\tau,n}}{S_{l-p,k} G_{l,n} W_{r,n,\tau} U_{k,r,p,m-\tau}} + \frac{X_{l,m} - C_{l,m}}{C_{l,m}} + \ln C_{l,m} \right) \\ &+ \sum_{k,r,p,m} \left\{ (\alpha^{(\mathrm{IS})} + 1) \ln U_{k,r,p,m} + \frac{\beta^{(\mathrm{IS})}}{U_{k,r,p,m}} \right\}\end{aligned} \tag{3.19}$$

The update equations can be derived similarly as

$$S_{k,l'} \leftarrow S_{k,l'} \sqrt{\frac{\sum_{l,m,r,\tau} \frac{Y^2_{l,m}}{X^2_{l,m}} F_{r,l,\tau} U_{k,r,p,m}}{\sum_{l,m,r,\tau} \frac{F_{r,l,\tau} U_{k,r,p,m}}{X_{l,m}}}}, \quad S_{k,l'} \leftarrow \frac{S_{k,l'}}{\sum_l S_{k,l}}, \tag{3.20}$$

$$W_{r,n,\tau} \leftarrow W_{r,n,\tau} \sqrt{\frac{\sum_{l,m,k,p} \frac{Y^2_{l,m}}{X^2_{l,m}} S_{k,l-p} G_{l,n} U_{k,r,p,m}}{\sum_{l,m,k,p} \frac{S_{k,l-p} G_{l,n} U_{k,r,p,m}}{X_{l,m}}}}, \quad W_{r,n,\tau} \leftarrow \frac{W_{r,n,\tau}}{\sum_{n',\tau'} W_{r,n',\tau'}}, \tag{3.21}$$

$$U_{k,r,p,m'} = \frac{A_{k,r,p,m'}}{\sqrt{\left(\frac{\alpha^{(\mathrm{IS})} + 1}{2}\right)^2 + \left(\sum_{l,m} \frac{S_{k,l-p} F_{r,l,m-m'}}{X_{l,m}}\right) A_{k,r,p,m'}} + \frac{\alpha^{(\mathrm{IS})} + 1}{2}} \tag{3.22}$$

where

$$A_{k,r,p,m'} = \sum_{l,m} \frac{Y^2_{l,m}}{X^2_{l,m}} F_{r,l,m-m'} S_{k,l-p} U^2_{k,r,p,m'} + \beta^{(\mathrm{IS})}. \tag{3.23}$$

3.5 Experiments

3.5.1 Experimental Conditions

To evaluate the proposed algorithms in signal-to-distortion ratio (SDR), we conducted a supervised source separation experiment. SDRs were computed with the BSSEval toolbox [75]. For the convenience, we call the proposed algorithm with the I divergence criterion (the IS divergence criterion) *I-SNMFwSF* (*IS-SNMFwSF*, respectively). For comparison,

we employed shifted NMF with the I divergence criterion (*I-SNMF*) and that with the IS divergence criterion (*IS-SNMF*). While the original shifted NMF [37] does not contain any terms inducing the sparsity of parameters, the use of $\mathcal{R}_*(U)$ improved SDRs and we here used $\mathcal{R}_*(U)$.

The experimental data was the Bach10 dataset [76], which consists of audio recordings of ten four-part chorales by J. S. Bach. Each recording is a mixture of violin, clarinet saxophone and bassoon performances, which correspond to the soprano, alto, tenor and bass parts of each musical piece, respectively. Audio recordings of individual parts are also contained in the dataset. All recordings were monaural and downsampled to 16 kHz. Magnitude spectrograms were computed with the fast approximate CWT algorithm [61,62]. The center frequencies ranged from 27.5 to 7902 Hz with 100/3 cent interval and the log-normal wavelet [1] was used as an analyzing wavelet. The wavelet has a Gaussian shape with a common variance in the log-frequency domain, and we set a parameter corresponding to the standard deviation of the Gaussian as a one fifth of a semitone interval.

We first trained S and W of the proposed models and basis spectra of the shifted NMFs with the audio recordings of individual parts of the five musical pieces (training data), and then performed source separation on the audio recordings of the other five musical pieces (test data). With the proposed algorithms, a pair of a source excitation and a filter was trained for each instrument, and a total of four pairs of a source and a filter were used for the separation. With the shifted NMFs, one basis spectrum was assigned to each instrument and a total of four basis spectra were used for the separation. For each test data, we designed a soft time-frequency mask as $\tilde{X}_{l,m,k,r}/X_{l,m}$ to obtain separated audio signals of the sources. The proposed methods and the shifted NMFs ran for 100 iterations both in the training and test stages. As $\alpha^{(\mathrm{I})}$ or $\alpha^{(\mathrm{IS})}$, we use $\alpha^{(\mathrm{train})} = 1.0 \times 10^{-10}, 0.2, 0.4, 0.6, 0.8, 1.0$ for the training data and $\alpha^{(\mathrm{test})} = 1.0 \times 10^{-10}, 0.2, 0.4, 0.6, 0.8, 1.0$ for the test data. The other parameters were set as follows: $\beta^{(\mathrm{I})} = \beta^{(\mathrm{IS})} = 1.0 \times 10^{-10}$, $M^{(\mathrm{tap})} = 1$, $N = 140$, $\nu = \pi/2(N-1)$ and $\rho_n = \pi n/(N-1)$ for $n = 0, \cdots, N-1$.

3.5.2 Results

Table 3.2 summarizes average SDR improvements with standard errors obtained with all algorithms for each musical piece, and Fig. 3.2 compares all algorithms in average SDR improvements, signal-to-interferences ratio (SIR) improvements and signal-to-artifacts ra-

Table 3.2: Average SDR improvements with standard errors [dB] obtained with the proposed algorithms (I-SNMFwSF and IS-SNMFwSF) and the shifted NMFs (I-SNMF and IS-SNMF). The displayed results were the highest in overall average SDR improvement of all combinations $(\alpha^{(\text{train})}, \alpha^{(\text{test})})$ for each algorithm, and the pairs of two values below the algorithm names are $(\alpha^{(\text{train})}, \alpha^{(\text{test})})$ for the highest results.

Musical Piece	I-SNMFwSF $(0.6, 1.0 \times 10^{-10})$	IS-SNMFwSF $(1.0, 0.6)$	I-SNMF $(1.0, 0.4)$	IS-SNMF $(0.4, 1.0)$
No. 1	$\mathbf{6.35 \pm 1.60}$	5.59 ± 0.45	4.00 ± 1.34	2.39 ± 1.35
No. 2	$\mathbf{6.39 \pm 1.67}$	4.81 ± 0.69	4.74 ± 1.06	2.91 ± 1.49
No. 3	$\mathbf{5.67 \pm 1.61}$	3.84 ± 0.82	3.53 ± 1.08	2.88 ± 1.51
No. 4	$\mathbf{5.02 \pm 0.57}$	4.58 ± 0.69	3.23 ± 0.44	2.24 ± 0.52
No. 5	$\mathbf{6.60 \pm 1.40}$	5.01 ± 0.54	4.32 ± 1.08	3.64 ± 1.07
Overall	$\mathbf{6.01 \pm 0.58}$	4.77 ± 0.29	3.97 ± 0.44	2.81 ± 0.51

tio (SAR) for all data. In both divergences, the proposed algorithms provided around 2 dB higher SDR improvements on average compared to the shifted NMFs. Thus, we can confirm that the incorporation of the source-filter model improves the source separation accuracy in the CWT domain. The results of *IS-SNMFwSF* were less on average than those of *I-SNMFwSF*, and the tendency of the difference in separation accuracy between the I divergence and the IS divergence is consistent with the results on conventional NMFs (for example, see [77]).

Fig. 3.3 displays average SDR improvements and standard errors for individual musical instruments. *I-SNMFwSF* with $N \geq 100$ provided significantly higher SDR improvements for all musical instruments compared to *I-SNMF*. This shows that the incorporation of the source-filter model is valid for the four musical instruments. We found that the best N was different for each musical instrument, and exploring the best Ns for other musical instruments and classifying them is one of the future works.

3.6 Summary

This chapter has developed a new source separation method by incorporating the source-filter model into shifted NMF. With the proposed model, the observed spectrogram is represented by a product of excitation and filter spectrograms. The excitation spectrogram is described with shifted NMF to exploit the constant inter-harmonic spacings of a harmonic

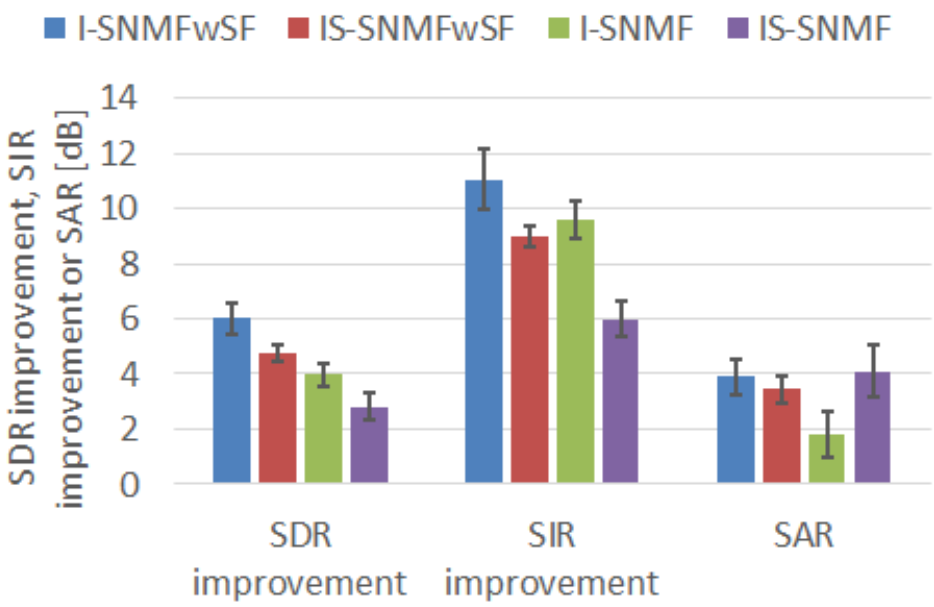


Figure 3.2: Average SDR improvements, SIR improvements and SARs with standard errors for overall data. The parameters and algorithms are the same as Table 3.2.

structure in the log-frequency domain, and the filter spectrogram is modeled by NMFD to represent temporal dynamics of timbre. We have derived iterative algorithms of estimating parameters for the I divergence and IS divergence criterions based on the auxiliary function approach. We have experimentally confirmed that the proposed algorithm outperformed shifted NMF in the accuracy of source separation. In future, we will examine the effect of setting $M^{(\mathrm{tap})} > 1$ to the source separation accuracy.

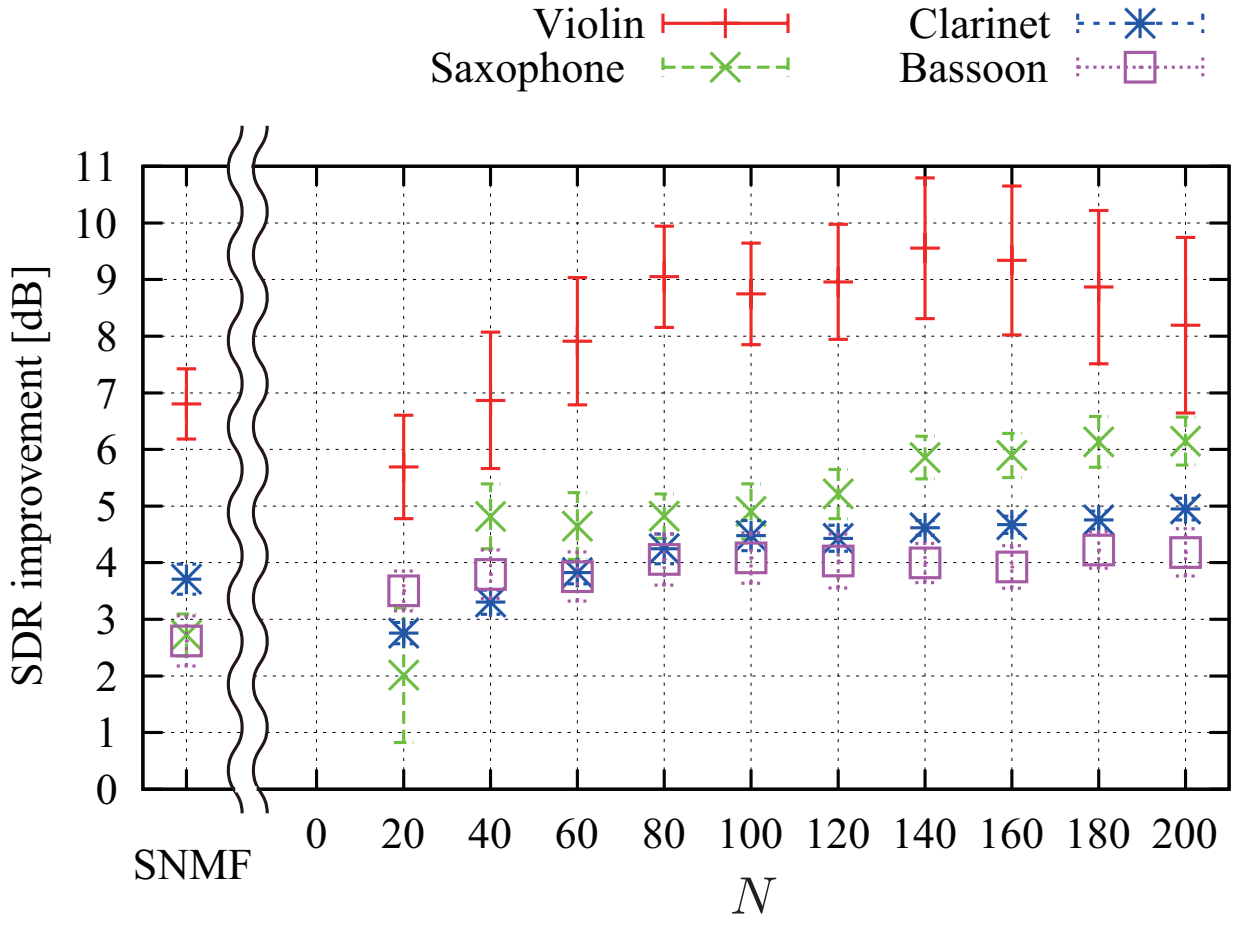


Figure 3.3: Average SDR improvements and standard errors obtained with the proposed algorithm (*I-SNMFwSF*) and *I-SNMF* for each musical instrument. "SNMF" corresponds to *I-SNMF*.

Chapter 4

Harmonic Temporal Factor Decomposition

4.1 Chapter Overview

For monaural source separation two main approaches have thus far been adopted. One approach involves applying NMF to an observed magnitude spectrogram, interpreted as a non-negative matrix. The other approach is based on the concept of CASA. A CASA-based approach called the "harmonic-temporal clustering (HTC)" aims to cluster the time-frequency components of an observed signal based on a constraint designed according to the local time-frequency structure common in many sound sources (such as harmonicity and the continuity of frequency and amplitude modulations). This chapter proposes a new approach for monaural source separation called the "Harmonic-Temporal Factor Decomposition (HTFD)" by introducing a spectrogram model that combines the features of the models employed in the NMF and HTC approaches. We further describe some ideas how to design the prior distributions for the present model to incorporate musically relevant information into the separation scheme.

4.2 Introduction

Monaural source separation is a process in which the signals of concurrent sources are estimated from a monaural polyphonic signal and is one of fundamental objectives offering a wide range of applications such as music information retrieval, music transcription and

audio editing.

While we can use spatial cues for blind source separation with multichannel inputs, for monaural source separation we need other cues instead of the spatial cues. For monaural source separation two main approaches have thus far been adopted. One approach is based on the concept of computational auditory scene analysis (e.g., [78]). The auditory scene analysis process described by Bregman [11] involves grouping elements that are likely to have originated from the same source into a perceptual structure called an auditory stream. In [1,17], an attempt has been made to imitate this process by clustering time-frequency components based on a constraint designed according to the auditory grouping cues (such as the harmonicity and the coherences and continuities of amplitude and frequency modulations). This method is called "HTC."

The other approach involves applying NMF to an observed magnitude spectrogram interpreted as a non-negative matrix [21]. The idea behind this approach is that the spectrum at each frame is assumed to be represented as a weighted sum of a limited number of common spectral templates. Since the spectral templates and the mixing weights should both be non-negative, this implies that an observed spectrogram is modeled as the product of two non-negative matrices. Thus, factorizing an observed spectrogram into the product of two non-negative matrices allows us to estimate the unknown spectral templates constituting the observed spectra and decompose the observed spectra into components associated with the estimated spectral templates.

The two approaches described above rely on different clues for making separation possible. Roughly speaking, the former approach focuses on the local time-frequency structure of each source, while the latter approach focuses on a relatively global structure of music spectrograms (such a property that a music signal typically consists of a limited number of recurring note events). Rather than discussing which clues are more useful, we believe that both of these clues can be useful for achieving a reliable monaural source separation algorithm. This belief has led us to develop a new model and method for monaural source separation that combine the features of both HTC and NMF. We call the present method "HTFD."

The present model is formulated as a probabilistic generative model in such a way that musically relevant information can be flexibly incorporated into the prior distributions of the model parameters. Given the recent progress of state-of-the-art methods for a variety of music information retrieval (MIR)-related tasks such as audio key detection, audio chord

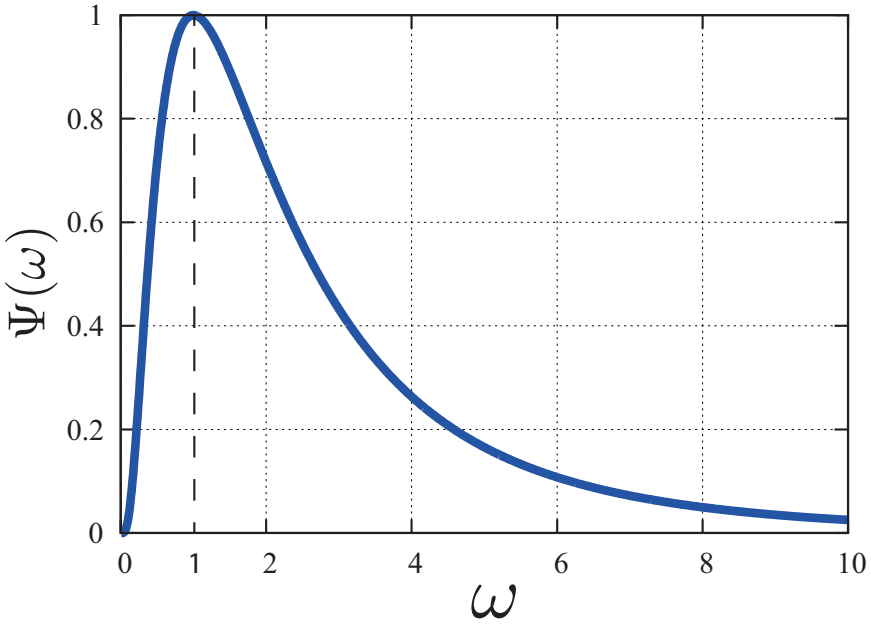


Figure 4.1: The Fourier transform of the log-normal wavelet defined in [1].

detection, and audio beat tracking, information such as key, chord and beat extracted from the given signal can potentially be utilized as reliable and useful prior information for source separation. The inclusion of auxiliary information in the separation scheme is referred to as informed source separation and is gaining increasing momentum in recent years (see e.g., among others, [35, 42, 79, 80]). This chapter further describes some ideas how to design the prior distributions for the present model to incorporate musically relevant information.

We henceforth denote the normal, Dirichlet and Poisson distributions by $\mathcal{N}$, Dir and Pois, respectively.

4.3 Spectrogram Model of Music Signal

4.3.1 Continuous Wavelet Transform of Source Signal Model

As in [1], this section derives the CWT of a source signal. Let us first consider as a signal model for the sound of the kth pitch the analytic signal representation of a pseudo-periodic signal given by

$$f_k(u) = \sum_{n=1}^{N} a_{k,n}(u) e^{\mathrm{j}(n\theta_k(u)+\varphi_{k,n})}, \tag{4.1}$$

where $u \in (-\infty, \infty)$ denotes the continuous time, $n\theta_k(u) + \varphi_{k,n}$ the instantaneous phase of the nth harmonic and $a_{k,n}(u)$ the instantaneous amplitude. This signal model implicitly

ensures not to violate the 'harmonicity' and 'coherent frequency modulation' constraints of the auditory grouping cues. Now, let the wavelet basis function be defined by

$$\psi_{\alpha,t}(u) = \frac{1}{\sqrt{2\pi}\alpha}\psi\left(\frac{u-t}{\alpha}\right), \tag{4.2}$$

where α is the scale parameter such that $\alpha > 0$, t the shift parameter and $\psi(u)$ the mother wavelet with the center frequency of 1 satisfying the admissibility condition. $\psi_{\alpha,t}(u)$ can thus be used to measure the component of period α at time t. The CWT of $f_k(u)$ is then defined by

$$W_k(\ln\tfrac{1}{\alpha}, t) = \int_{-\infty}^{\infty}\sum_{n=1}^{N} a_{k,n}(u)e^{\mathrm{j}(n\theta_k(u)+\varphi_{k,n})}\psi^*_{\alpha,t}(u)\mathrm{d}u. \tag{4.3}$$

Since the dominant part of $\psi^*_{\alpha,t}(u)$ is typically localized around time t, the result of the integral in Eq. (4.3) shall depend only on the values of $\theta_k(u)$ and $a_{k,n}(u)$ near t. By taking this into account, we replace $\theta_k(t)$ and $a_{k,n}(t)$ with zero-and first-order approximations around time t:

$$\begin{aligned} a_{k,n}(u) =& a_{k,n}(t) + \frac{\mathrm{d}a_{k,n}(u)}{\mathrm{d}u}\Big|_{u=t}(u-t) + \cdots && (4.4)\\ \simeq& a_{k,n}(t) && (4.5)\\ \theta_k(u) =& \theta_k(t)\frac{\mathrm{d}\theta_k(u)}{\mathrm{d}u}\Big|_{u=t}(u-t) + \frac{1}{2}\frac{\mathrm{d}^2\theta_k(u)}{\mathrm{d}u^2}\Big|_{u=t}(u-t)^2 + \cdots && (4.6)\\ \simeq& \theta_k(t) + \dot{\theta}_k(t)(u-t). && (4.7) \end{aligned}$$

Note that the variable $\dot{\theta}_k(u)$ corresponds to the instantaneous fundamental frequency (F_0). By undertaking the above approximations, applying the Parseval's theorem, and putting $x = \ln(1/\alpha)$ and $\Omega_k(t) = \ln\dot{\theta}_k(t)$, we can further write Eq. (4.3) as

$$W_k(x,t) = \sum_{n=1}^{N} a_{k,n}(t)\Psi^*(ne^{-x+\Omega_k(t)})e^{j(n\theta_k(t)+\varphi_{k,n})}, \tag{4.8}$$

where x denotes log-frequency and Ψ the Fourier transform of ψ. Since the function Ψ can be chosen arbitrarily, as with [1], we employ the following unimodal real function whose maximum is taken at $\omega = 1$:

$$\Psi(\omega) = \begin{cases} e^{-\frac{(\ln\omega)^2}{2\sigma^2}} & (\omega > 0) \\ 0 & (\omega \le 0) \end{cases}. \tag{4.9}$$

The illustration of $\Psi(\omega)$ is shown in Fig. 4.1. Eq. (4.8) can then be written as

$$W_k(x,t) = \sum_{n=1}^{N} a_{k,n}(t)e^{-\frac{(x-\Omega_k(t)-\ln n)^2}{2\sigma^2}}e^{j(n\theta_k(t)+\varphi_{k,n})}. \tag{4.10}$$

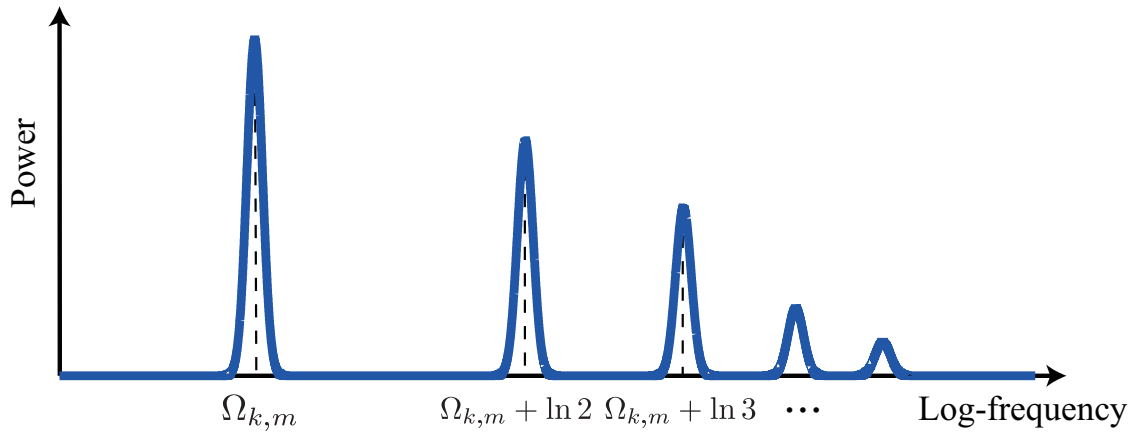


Figure 4.2: Spectral model of the pseudo-periodic signal at time t_m in the CWT domain.

If we now assume that the time-frequency components are sparsely distributed so that the partials rarely overlap each other, $|W_k(x,t)|$ is given approximately as

$$|W_k(x,t)| \simeq \sum_{n=1}^{N} |a_{k,n}(t)| e^{-\frac{(x-\Omega_k(t)-\ln n)^2}{2\sigma^2}}. \tag{4.11}$$

This assumption means that the magnitude spectra of the partials can approximately be considered additive. Note that a cutting plane of the spectrogram model given by Eq. (4.11) at time t is expressed as a harmonically-spaced Gaussian mixture function as depicted in Fig. 4.2. It should be noted that this model is identical to the one employed in the HTC approach [1]. It is worthwhile noting that the spectrogram model approximately describes the spectral leakage of the pseudo-periodic audio signal and enables us to develop a method that takes account of the spectral leakage effect in the CWT domain.

Although we have defined the spectrogram model above in continuous time and continuous log-frequency, we actually obtain observed spectrograms as a discrete time-frequency representation through computer implementations. Thus, we henceforth use $Y_{l,m} := Y(x_l, t_m)$ to denote an observed spectrogram where x_l $(l = 0, \ldots, L-1)$ and t_m $(m = 0, \ldots, M-1)$ stand for the uniformly-quantized log-frequency points and time points, respectively. We will also use the notation $\Omega_{k,m}$ and $a_{k,n,m}$ to indicate $\Omega_k(t_m)$ and $a_{k,n}(t_m)$.

4.3.2 Observed Spectrogram Model

The key assumption behind the NMF model is that the spectra of the sound of a particular pitch is expressed as a multiplication of time-independent and time-dependent factors. In order to extend the NMF model to a more reasonable one, we consider it important to clarify

which factors involved in the spectra should be assumed to be time-dependent and which factors should not. For example, the F_0 must be assumed to vary in time during vibrato or portamento. As with the NMF model, the scale of the spectrum should also be assumed to be time-varying, whereas the spectral shape of each pitch can be relatively static.

These assumptions can be reflected to the present model in the following way. We factorize $|a_{k,n,m}|$ into the product of time-independent and time-dependent variables, $w_{k,n}$ and $U_{k,m}$:

$$|a_{k,n,m}| = w_{k,n} U_{k,m} \tag{4.12}$$

$w_{k,n}$ can be interpreted as the relative magnitude of harmonic n and $U_{k,m}$ as the time-varying magnitude of the sound of pitch k. To avoid an indeterminacy in scaling, we introduce $\sum_n w_{k,n} = 1$ for all k.

If we assume the additivity of magnitude spectra, the magnitude spectrogram of a superposition of K pitched sounds is given by the sum of Eq. (4.11) over k. In equation, we can write a spectrogram model $X_{l,m}$ as

$$X_{l,m} = \sum_{k=0}^{K-1} \tilde{X}_{k,l,m}, \tag{4.13}$$

$$\tilde{X}_{k,l,m} = \underbrace{\sum_{n=1}^{N} w_{k,n} e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma}}}_{H_{k,l,m}} U_{k,m}. \tag{4.14}$$

If we denote the term inside the parenthesis by $H_{k,l,m}$, $X_{l,m}$ can be rewritten as $X_{l,m} = \sum_k H_{k,l,m} U_{k,m}$ and so the relation to the NMF model may become much clearer. It should be noted that the change of $H_{k,l,m}$ with the F_0 can be represented by shifting a specific time slice of $H_{k,l,m}$ up or down since the inter-harmonic spacings of a harmonic structure in the log-frequency domain are constant. This characteristics unique in the log-frequency domain has been also utilized in shifted NMF [37], shift-invariant probabilistic latent component analysis [38] and specmurt analysis [47].

4.3.3 Formulating Probabilistic Model

Since the assumptions and approximations we made so far do not always hold exactly in reality, an observed spectrogram $Y_{l,m}$ may diverge from $X_{l,m}$ even though the parameters are optimally determined. One way to simplify the process by which this kind of deviation occurs would be to assume a probability distribution of $Y_{l,m}$ with the expected value of $X_{l,m}$.

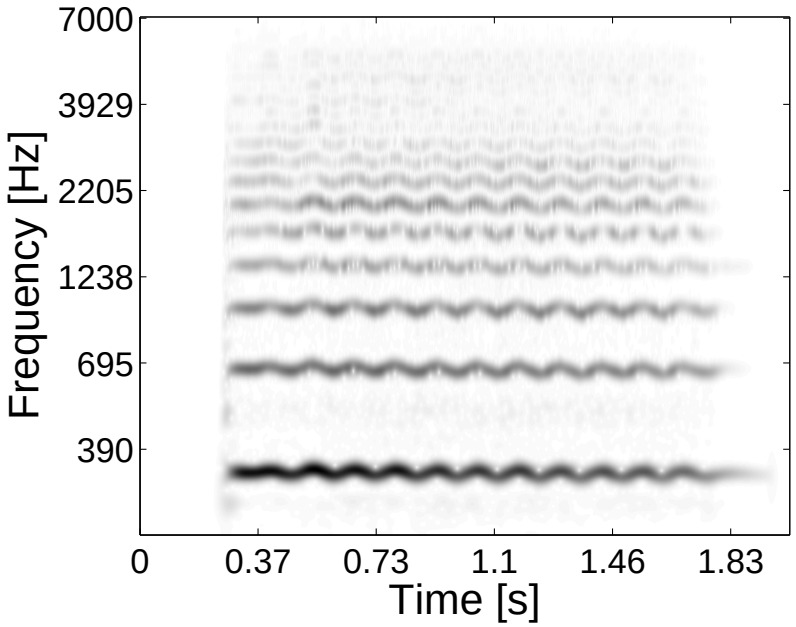


Figure 4.3: Spectrogram of a violin vibrato sound recorded in RWC music instrument database [2].

Here, we assume that $Y_{l,m}$ follows a Poisson distribution with mean $X_{l,m}$

$$Y_{l,m} \sim \text{Pois}(Y_{l,m}; X_{l,m}), \tag{4.15}$$

where

$$\text{Pois}(z;\xi) = \frac{\xi^z e^{-\xi}}{\Gamma(z)}. \tag{4.16}$$

This defines our likelihood function

$$p(Y|\Theta) = \prod_{l,m} \text{Pois}(Y_{l,m}; X_{l,m}), \tag{4.17}$$

where Y denotes the set consisting of $Y_{l,m}$ and Θ the entire set consisting of the unknown model parameters. It should be noted that the maximization of the Poisson likelihood with respect to $X_{l,m}$ amounts to optimally fitting $X_{l,m}$ to $Y_{l,m}$ by using the generalized Kullback-Leibler divergence (a.k.a I-divergence) as the fitting criterion. The choice of the Poisson likelihood is made for the convenience of deriving the optimization algorithm, which we will show in Sec. 4.5.

The F_0 of stringed and wind instruments often varies continuously over time with musical expressions such as vibrato and portamento. For example, the F_0 of a violin sound varies periodically around the note frequency during vibrato, as depicted in Fig. 4.3. Let us denote the standard log-F_0 corresponding to the kth note by μ_k. To appropriately describe the

variability of an F_0 contour in both the global and local time scales, we design a prior distribution for $\mathbf{\Omega}_k := (\Omega_{k,0}, \Omega_{k,1}, \ldots, \Omega_{k,M-1})^\mathsf{T}$ by employing the product-of-experts (PoE) [81] concept using two probability distributions. First, we design a distribution $q_\mathrm{g}(\mathbf{\Omega}_k)$ describing how likely $\Omega_{k,0}, \ldots, \Omega_{k,L-1}$ stay near μ_k. Second, we design another distribution $q_\mathrm{l}(\mathbf{\Omega}_k)$ describing how likely $\Omega_{k,0}, \ldots, \Omega_{k,L-1}$ are locally continuous along time. Here we define $q_\mathrm{g}(\mathbf{\Omega}_k)$ and $q_\mathrm{l}(\mathbf{\Omega}_k)$ as

$$q_\mathrm{g}(\mathbf{\Omega}_k) = \mathcal{N}(\mathbf{\Omega}_k; \mu_k \mathbf{1}_M, \upsilon_k^2 I_M), \tag{4.18}$$

$$q_\mathrm{l}(\mathbf{\Omega}_k) = \mathcal{N}(\mathbf{\Omega_k}; \mathbf{0}_M, \tau_k^2 D^{-1}), \tag{4.19}$$

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & 0 & 0 & -1 & 1 \end{bmatrix}, \tag{4.20}$$

where I_M denotes an $M \times M$ identity matrix, D an $M \times M$ band matrix, $\mathbf{1}_M$ an M-dimensional all-one vector, and $\mathbf{0}_M$ an M-dimensional all-zero vector, respectively. υ_k denotes the standard deviation from mean μ_k, and τ_k the standard deviation of the F_0 jumps between adjacent frames. The prior distribution of $\mathbf{\Omega}_k$ is then derived as

$$p(\mathbf{\Omega}_k) \propto q_\mathrm{g}(\mathbf{\Omega}_k)^{\alpha_\mathrm{g}} q_\mathrm{l}(\mathbf{\Omega}_k)^{\alpha_\mathrm{l}} \tag{4.21}$$

where α_g and α_l are the hyperparameters that weigh the contributions of $q_g(\mathbf{\Omega}_k)$ and $q_\mathrm{l}(\mathbf{\Omega}_k)$ to the prior distribution.

4.3.4 Relation to Other Models

It should be noted that the present model is related to other models proposed previously. If we do not assume a parametric model for $H_{k,l,m}$ and treat each $H_{k,l,m}$ itself as the parameter, the spectrogram model $X_{l,m}$ can be seen as an NMF model with time-varying basis spectra, as in [25]. In addition to this assumption, if we assume that $H_{k,l,m}$ is time-invariant (i.e., $H_{k,l,m} = H_{k,l}$), $X_{l,m}$ reduces to the regular NMF model [21]. Furthermore, if we assume each basis spectrum to have a harmonic structure, $X_{l,m}$ becomes equivalent to the harmonic

NMF model [32, 33]. If we assume that $\Omega_{k,m}$ is equal over time m, $X_{l,m}$ reduces to a model similar to the ones described in [82, 83]. Furthermore, if we describe $U_{k,m}$ using a parametric function of m, $X_{l,m}$ becomes equivalent to the HTC model [1, 17]. With a similar motivation, Hennequin *et al.* developed an extension to the NMF model defined in the short-time Fourier transform (STFT) domain to allow the F_0 of each basis spectrum to be time-varying [84].

4.4 Incorporation of Auxiliary Information

We consider using side-information obtained with the state-of-the-art methods for MIR-related tasks including key detection, chord detection and beat tracking to assist source separation.

When multiple types of side-information are obtained for a specific parameter, we can combine the use of the mixture-of-experts and PoE [81] concepts according to the "AND" and "OR" conditions we design similarly in the previous section. For example, pitch occurrences typically depend on both the chord and key of a piece of music. Thus, when the chord and key information are obtained, we may use the product-of-experts concept to define a prior distribution for the parameters governing the likeliness of the occurrences of the pitches. In the next subsection, we describe specifically how to design the prior distributions.

4.4.1 Designing Prior Distributions

The likeliness of the pitch occurrences in popular and classical western music usually depend on the key or the chord used in that piece. The likeliness of the pitch occurrences can be described as a probability distribution over the relative energies of the sounds of the individual pitches.

Since the number of times each note is activated is usually limited, inducing sparsity to the temporal activation of each note event would facilitate the source separation. The likeliness of the number of times each note is activated can be described as well as a probability distribution over the temporal activations of the sound of each pitch.

To allow for designing such prior distributions, we decompose $U_{k,m}$ as the product of three variables: the total energy C, the pitch-wise relative energy $E_k = \sum_m U_{k,m}/V$ (i.e. $\sum_k E_k = 1$), and the pitch-wise normalized amplitude $A_{k,m} = U_{k,m}/(CE_k)$ (i.e. $\sum_m A_{k,m} = 1$

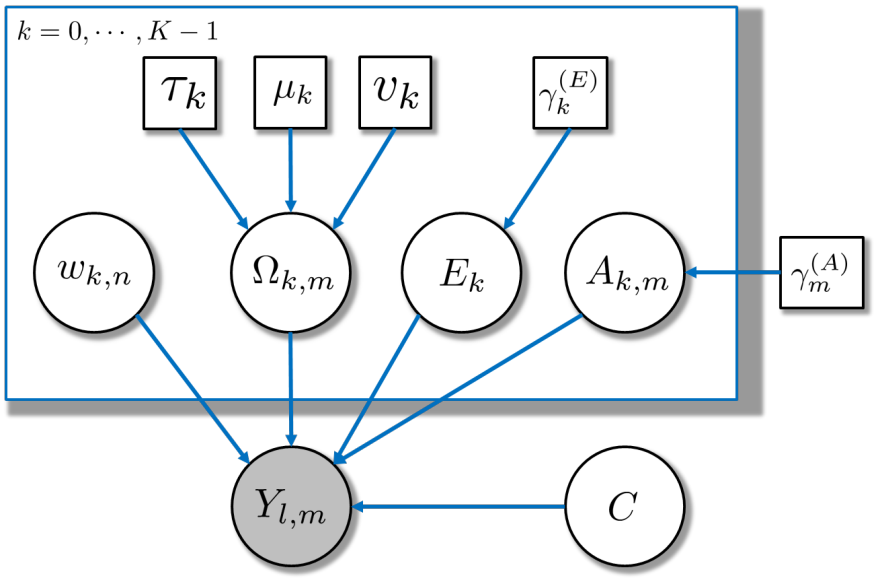


Figure 4.4: Plate notation of the proposed overall generative model.

for all k). Hence we can write

$$U_{k,m} = CE_kA_{k,m} \tag{4.22}$$

This decomposition allows us to incorporate different kinds of prior information into our model by separately defining prior distributions over $\boldsymbol{E} = (E_0, \ldots, E_{K-1})^\top$ and $\boldsymbol{A}_k = (A_{k,0}, \ldots, A_{k,M-1})^\top$. Here we introduce Dirichlet distributions:

$$\boldsymbol{A}_k \sim \mathrm{Dir}(\boldsymbol{A}_k; \boldsymbol{\gamma}_k^{(A)}), \tag{4.23}$$

$$\boldsymbol{E} \sim \mathrm{Dir}(\boldsymbol{E}; \boldsymbol{\gamma}^{(E)}), \tag{4.24}$$

where $\mathrm{Dir}(\boldsymbol{z}; \boldsymbol{\xi}) \propto \prod_i z_i^{\xi_i - 1}$, $\boldsymbol{\gamma}^{(A)} := (\gamma_0^{(A)}, \ldots, \gamma_{M-1}^{(A)})^\top$, and $\boldsymbol{\gamma}^{(E)} := (\gamma_0^{(E)}, \ldots, \gamma_{K-1}^{(E)})^\top$. For $p(\boldsymbol{E})$, we set $\gamma_k^{(E)}$ at a reasonably high value if the kth pitch is contained in the musical scale and vice versa. For $p(\boldsymbol{A}_k)$, we set $\gamma_m^{(A)} < 1$ so that the Dirichlet distribution becomes a sparsity inducing distribution.

In summary, the overall generative model is depicted in plate notation in Fig. 4.4.

4.5 Parameter Estimation Algorithm

The random variables of interest in the present model are

$\Omega_{k,m}$: the logarithm of F_0 for pitch k at time m,

$w_{k,n}$: Relative energy of harmonic n of pitch k,

C: Total energy,

E_k: Relative energy of pitch k,

$A_{k,m}$: Pitch-wise normalized activation of pitch k at time m.

We denote the set of the above random variables as Θ. Given an observed magnitude spectrogram Y, we would like to find the estimates of Θ that maximizes the posterior density $p(\Theta|Y) \propto p(Y|\Theta)p(\Theta)$. We therefore consider the problem of maximizing

$$\mathcal{L}(\Theta) := \ln p(Y|\Theta) + \ln p(\Theta), \tag{4.25}$$

with respect to Θ where

$$\ln p(Y|\Theta) \underset{c}{=} \sum_{l,m} \left(Y_{l,m} \ln X_{l,m} - X_{l,m} \right) \tag{4.26}$$

$$\ln p(\Theta) = \sum_{k} \ln p(\boldsymbol{\Omega}_k) + \sum_{k} \ln p(\boldsymbol{A}_k) + \ln p(\boldsymbol{E}). \tag{4.27}$$

$\underset{c}{=}$ denotes equality up to constant terms. Since the first term of Eq. (4.26) involves summation over k and n, analytically solving the current maximization problem is intractable. However, we can develop a computationally efficient algorithm for finding a locally optimal solution based on the auxiliary function approach [71–73], by using a similar idea described in [1].

When applying an auxiliary function approach to a certain maximization problem, the first step is to define a lower bound function for the objective function. As mentioned earlier, the difficulty with the current maximization problem lies in the first term in Eq. (4.26). By using the fact that the logarithm function is a concave function, we can invoke the Jensen's inequality

$$Y_{l,m} \ln X_{l,m} \geq Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \ln \frac{w_{k,n} e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{\lambda_{k,n,l,m}} \tag{4.28}$$

$$= Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \left(\ln w_{k,n} + \ln U_{k,m} - \frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2} - \ln \lambda_{k,n,l,m} \right) \tag{4.29}$$

to obtain a lower bound function, where $\lambda_{k,n,l,m}$ is a positive variable that sums to unity: $\sum_{k,n} \lambda_{k,n,l,m} = 1$. The equality of the inequality (4.28) holds if and only if

$$\lambda_{k,n,l,m} = \frac{w_{k,n}^2 e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{X_{l,m}}. \tag{4.30}$$

Although one may notice that the second term in Eq. (4.26) is nonlinear in $\Omega_{k,m}$, the summation of $X_{l,m}$ over l can be approximated fairly well using the integral $\int_{-\infty}^{\infty} X(x, t_m)\mathrm{d}x$ since $\sum_l X_{l,m}$ is the sum of the values at the sampled points $X(x_1, t_m), \ldots, X(x_L, t_m)$ with an equal interval, say Δ_x. Hence,

$$\begin{aligned}
\sum_l X_{l,m} &\simeq \frac{1}{\Delta_x} \int_{-\infty}^{\infty} X(x, t_m)\mathrm{d}x \\
&= \frac{1}{\Delta_x} \sum_{k,n} w_{k,n} U_{k,m} \int_{-\infty}^{\infty} e^{-\frac{(x - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} \mathrm{d}x \\
&= \frac{\sqrt{2\pi}\sigma}{\Delta_x} \sum_k U_{k,m} \sum_n w_{k,n}.
\end{aligned} \tag{4.31}$$

This approximation implies that the second term in Eq. (4.26) depends little on $\Omega_{k,m}$. The choice of the Poisson likelihood enables us to use the approximation, which leads that update equations can be derived in closed form.

An auxiliary function can thus be written as

$$\begin{aligned}
\mathcal{L}^+(\Theta, \lambda) \underset{c}{=} & Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \left(\ln w_{k,n} + \ln U_{k,m} - \frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2} - \ln \lambda_{k,n,l,m} \right) \\
& - \frac{\sqrt{2\pi}\sigma}{\Delta_x} \sum_m \sum_k U_{k,m} \sum_n w_{k,n} + \ln p(\Theta)
\end{aligned} \tag{4.32}$$

where λ denotes the set consisting of $\lambda_{k,n,l,m}$. We can derive update equations for the model parameters, using the above auxiliary function. By setting at zero the partial derivative of $\mathcal{L}^+(\Theta, \lambda)$ with respect to each of the model parameters, we obtain

$$w_{k,n} \leftarrow \frac{\sum_{l,m} Y_{l,m} \lambda_{k,n,l,m}}{C E_k}, \quad w_{k,n} \leftarrow \frac{w_{k,n}}{\sum_n w_{k,n}} \tag{4.33}$$

$$\boldsymbol{\Omega}_k \leftarrow \left(\frac{\alpha_\mathrm{l}}{\tau^2} D + \frac{\alpha_\mathrm{g}}{\upsilon_k^2} \boldsymbol{I}_M + \sum_{n,l} \mathrm{diag}(\boldsymbol{p}_{k,n,l}) \right)^{-1} \left(\mu_k \frac{\alpha_\mathrm{g}}{\upsilon_k^2} \mathbf{1}_M + \sum_{n,l} (x_l - \ln n) \boldsymbol{p}_{k,n,l} \right), \tag{4.34}$$

$$C \leftarrow \frac{\sum_{l,m} Y_{l,m}}{\sqrt{2\pi}\sigma / \Delta_x}, \tag{4.35}$$

$$E_k \leftarrow \sum_{l,m} Y_{l,m} \sum_n \lambda_{k,n,l,m} + \gamma_k^{(E)} - 1, \quad E_k \leftarrow \frac{E_k}{\sum_{k'} E_{k'}} \tag{4.36}$$

$$A_{k,m} \leftarrow \frac{\sum_l Y_{l,m} \sum_n \lambda_{k,n,l,m} + \gamma_m^{(A)} - 1}{R_k}, \quad A_{k,m} \leftarrow \frac{A_{k,m}}{\sum_{m'} A_{k,m'}}, \tag{4.37}$$

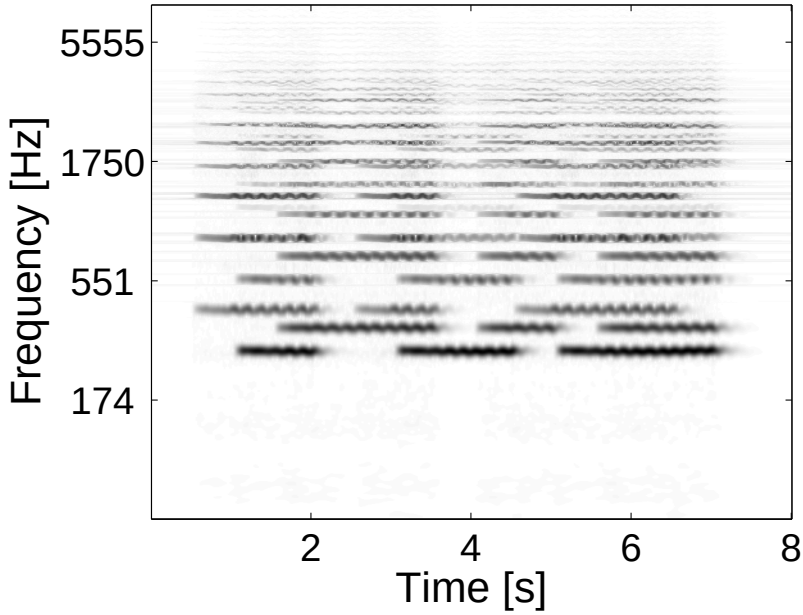


Figure 4.5: Spectrogram of a mixed audio signal of three violin vibrato sounds (D♭4, F4 and A♭4).

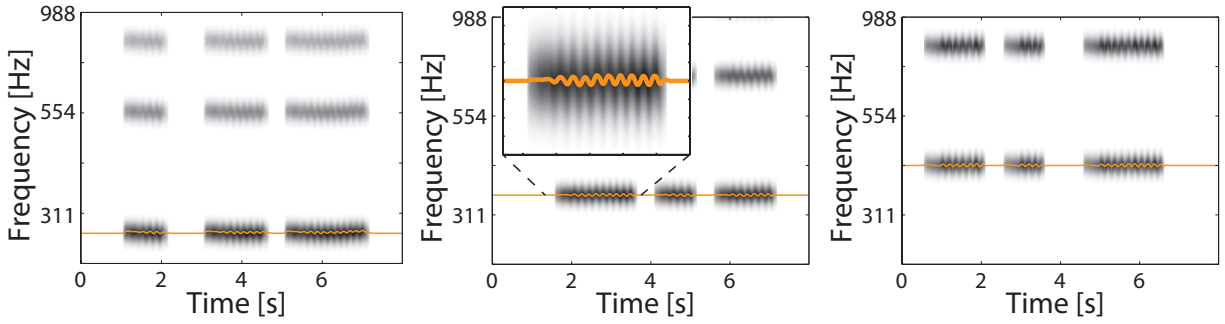
$$\boldsymbol{p}_{k,n,l} := \frac{1}{\sigma^2}\left[Y_{l,1}\lambda_{k,n,l,1}, Y_{l,2}\lambda_{k,n,l,2}, \cdots, Y_{l,M}\lambda_{k,n,l,M}\right]^\top, \tag{4.38}$$

where $\mathrm{diag}(\boldsymbol{p})$ converts a vector $\boldsymbol{p}$ into a diagonal matrix with the elements of $\boldsymbol{p}$ on the main diagonal. The second equations in the update rules of $w_{k,n}$, E_k and $A_{k,m}$ are for the normalization. With $\gamma_k^{(E)} < 1$ or $\gamma_m^{(A)} < 1$, $E_{k,m}$ and $A_{k,m}$ after the update rules may be negative. To keep E_k ($A_{k,m}$) to be non-negative, we set $E_k = 0$ ($A_{k,m} = 0$) if $E_{k,m}$ ($A_{k,m}$, respectively) becomes negative.

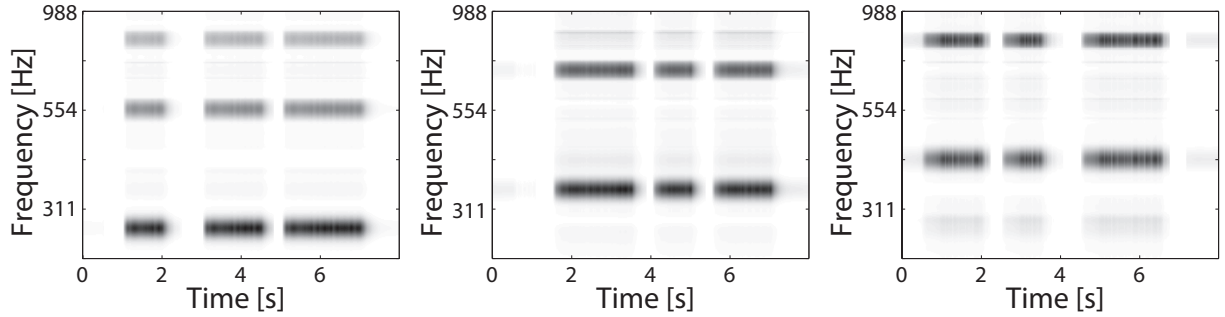
4.6 Objective Experiments

4.6.1 F_0 Tracking of Violin Sound

To confirm whether HTFD can track the F_0 contour of a sound, we compared HTFD with NMF with the I-divergence, by using a 16 kHz-sampled audio signal which were artificially made by mixing D♭4, F4 and A♭4 violin vibrato sounds from the RWC instrument database [2]. The F_0 of the pitch name A4 was set at 440 Hz. The spectrogram of the mixed signal is shown in Fig. 4.5. To convert the signal into a spectrogram, we employed the fast approximate CWT algorithm [61] with a 16 ms time-shift interval. As an analyzing wavelet, we used the log-normal wavelet with $\sigma = 0.02$. $\{x_l\}_l$ ranged 55 to 7040 Hz per 10 cent. The parameters of HTFD were set as follows: $N = 8$, $(\alpha_{\mathrm{g}}, \alpha_{\mathrm{l}}) = (1, 1)$, $\boldsymbol{\gamma}^{(E)} = (1 - 2.4 \times 10^{-3})\mathbf{1}_K$,



(a) Estimates of spectrograms and F_0 contours (orange lines) obtained with HTFD



(b) Estimates of spectrograms obtained with NMF

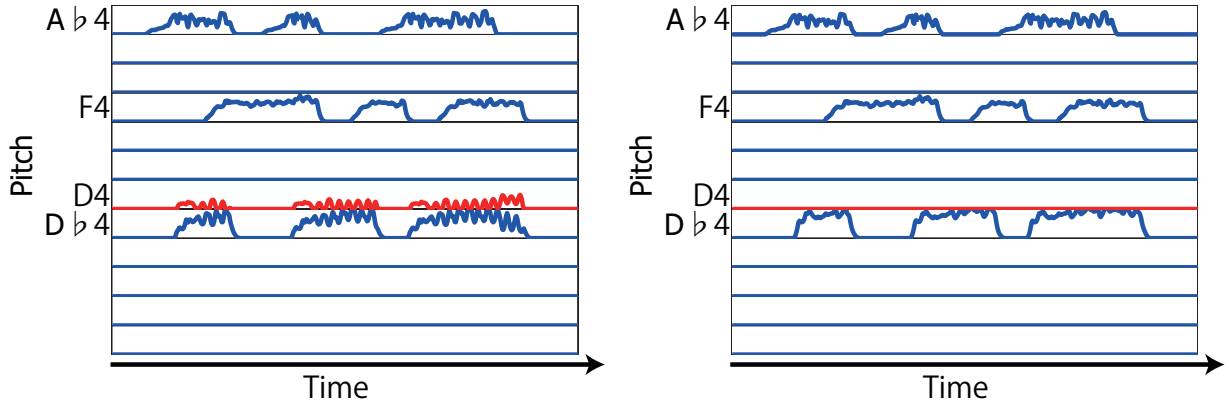
Figure 4.6: Estimated spectrogram models by HTFD and NMF. In left-to-right fashion, the spectrogram models are for D♭4, F4 and A♭4.

$\boldsymbol{\gamma}^{(A)} = (1 - 3.96 \times 10^{-6})\mathbf{1}_M$, $(\tau_k, \upsilon_k) = (0.83, 1.25)$, and $\mu_k = \ln(55) + (k-1) \times \ln(2)/12$ (A1 to A♯7, i.e. $K = 73$) for all k. The initial values of Θ were set as follows: $U_{k,m} = 1$ and $\Omega_{k,m} = \mu_k$ for all k and m, and $w_{k,n} \propto e^{-n}$ for all k and n. Such an initialization of $w_{k,n}$ corresponds to ideal harmonic structure and is known to be effective in multiple F_0 estimation [32]. The number of NMF bases were set at three. The parameter updates of both HTFD and NMF were stopped at 100 iterations.

While the estimates of spectrograms obtained with NMF were flat and the vibrato spectra seemed to be averaged (Fig. 4.6 (a)), those obtained with HTFD tracked the F_0 contours of the vibrato sounds appropriately (Fig. 4.6 (b)), and clear vibrato sounds were contained in the separated audio signals by HTFD.

4.6.2 Separation Using Key Information

We next examined whether the prior information of a sound improve source separation accuracy. The key of the sound used in 4.6.1, was assumed as D♭ major. The key information was incorporated in the estimation scheme by setting $\gamma_k^{(E)} = 1 - 2.4 \times 10^{-3}$ for the pitch



(a) Without key information (b) With key information

Figure 4.7: Temporal activations of A3–A♭4 estimated with HTFD using and without using prior information of the key. The red curves represent the temporal activations of D4.

indices that are not contained in the D♭ major scale and $\gamma_k^{(E)} = 1 - 3.0 \times 10^{-3}$ for the pitch indices contained in that scale. The other conditions were the same as 4.6.1.

With HTFD without using the key information, the estimated activations of the pitch indices that were not contained in the scale, in particular D4, were high as illustrated in Fig. 4.7 (a). In contrast, those estimated activations with HTFD using the key information were suppressed as shown in Fig. 4.7 (b). These results thus support strongly that incorporating prior information improve the source separation accuracy.

4.6.3 Transposing from One Key to Another

Here we show some results of an experiment on automatic key transposition [85] using HTFD. The aim of key transposition is to change the key of a musical piece to another key. We separated the spectrogram of a polyphonic sound into spectrograms of individual pitches using HFTD, transposed the pitches of the subset of the separated components, added all the spectrograms together to construct a pitch-modified polyphonic spectrogram, and constructed a time-domain signal from the modified spectrogram using the method described in [62]. For the key transposition, we adopted a simple way: To transpose, for example, from A *major* scale to A *natural minor* scale, we changed the pitches of the separated spectrograms corresponding to C♯, F♯ and G♯ to C, F and G, respectively. Some results are demonstrated in `http://tomohikonakamura.github.io/Tomohiko-Nakamura/demo/HTFD.html`.

4.6.4 Source Separation Accuracy

To evaluate the efficacy of HTFD, we then measured signal-to-distortion ratios (SDRs), signal-to-interferences ratios (SIRs) and signal-to-artifacts ratios (SARs), which were calculated with the BSSEval toolbox [75]. To examine whether the explicit incorporation of the spectral leakage into the model improves source separation accuracy, we compared HTFD (*CWT-HTFD*) with Harmonic NMF (*CWT-HNMF*) [32], which does not take account of the spectral leakage. Furthermore, to examine whether using CWT instead of STFT improves source separation accuracy, we implemented HTFD in the STFT domain with a Gaussian window (*STFT-HTFD*), which corresponds to [84], and Harmonic NMF in the STFT domain (*STFT-HNMF*), and compared *CWT-HTFD* and *CWT-HNMF* with *STFT-HTFD* and *STFT-HNMF*. We used the same prior distributions of activation matrices defined by Eqs. (4.23) and (4.24) for all algorithms.

It was difficult to prepare real performances played at each pitch, and so audio signals for the experiment were obtained by synthesizing the first 30 seconds of No. 1 to No. 5 from the RWC classic music database [2]. For the synthesis, we used a MIDI synthesizer called FluidSynth [86] and a high-quality GeneralUser GS 1.4 soundfont to maximize realistic synthesis as possible. In the synthesis, all control messages contained in the MIDI files were preserved. The sampling rate was set at 16 kHz. To compute CWT spectrograms of the signals, the fast approximate CWT algorithm was used with center frequencies ranging from 27.5 to 7040 Hz per 100/3 cent and a time shift of 0.916 ms. As an analyzing wavelet, we used the log-normal wavelet with $\sigma = \ln(2)/48, \ln(2)/60, \ln(2)/72$, which correspond to one fourth, one fifth and one sixth of a semitone interval. For the computation of STFT, we used $64, 32$ and 128 ms Gaussian windows with a hopsize of 10 ms.

The parameter of *CWT-HTFD* were set as follows: $N = 20$, $\alpha_{\mathrm{g}} = \alpha_{\mathrm{l}} = 1$, $\tau_k = \ln(2)/72 \times 9.16$ (a vibrato frequency of 6 Hz), $v_k = \ln(2)/36$ (a one third of semitone interval) and $\mu_k = \ln(27.5) + (k-1)\ln(2)/12$ for all k. The number of spectral bases was set at 88 for Harmonic NMFs, and the spectral bases correspond to pitch A0 to A♯7 with a semitone interval. The initial values of the parameters of HTFD were set as in Sec. 4.6.1, and a spectral basis for pitch k of Harmonic NMF was initialized by placing peaks at positions corresponding to F_0s and their harmonics, with magnitude decaying exponentially with increasing frequency. All algorithms were run for 100 iterations. To reduce computation time, the magnitude CWT spectrograms were sampled every ten time points such that

Table 4.1: Average SDR improvements, SIR improvements and SARs [dB] with standard errors for overall data. "CWT-HTFD" and "CWT-HNMF" represent HTFD and Harmonic NMF in the CWT domain, and "STFT-HTFD" and "STFT-HNMF" represent HTFD and Harmonic NMF in the STFT domain.

Algorithm	SDR improvement	SIR improvement	SAR
CWT-HTFD $(10^{-3} \times \mathbf{1}_K, 10^{-1} \times \mathbf{1}_M)$	$\mathbf{16.42 \pm 0.62}$	$\mathbf{26.07 \pm 0.88}$	$\mathbf{0.27 \pm 0.39}$
CWT-HNMF $(10^{-2} \times \mathbf{1}_K, \mathbf{1}_M)$	15.34 ± 0.75	25.50 ± 0.95	-0.32 ± 0.52
STFT-HTFD $(10^{-3} \times \mathbf{1}_K, 10^{-3} \times \mathbf{1}_M)$	14.02 ± 0.69	24.06 ± 0.87	-1.25 ± 0.45
STFT-HNMF $(10^{-3} \times \mathbf{1}_K, 10^{-3} \times \mathbf{1}_M)$	12.86 ± 0.80	23.51 ± 0.90	-2.01 ± 0.51

$t_m - t_{m-1} \simeq 10$ ms and then used for the estimation. For the separation, we designed a time-frequency mask as $\tilde{X}_{k,l,m}/X_{l,m}$ followed by linearly interpolating it.

Table 4.1 displays the results obtained with the individual algorithms for all data. The results of *CWT-HTFD* and *CWT-HNMF* were for $\sigma = \ln(2)/60$ and the results of *STFT-HTFD* and *STFT-HNMF* were for a 128 ms frame since the algorithms provided the highest SDR improvements of all σ or all frame lengths. The pairs of two values below the algorithm names are $(\boldsymbol{\gamma}^{(E)}, \boldsymbol{\gamma}^{(A)})$, which scored the highest SDR improvement for each algorithm in all combinations $\gamma_k^{(E)} = 1.0 \times 10^{-3,-2,-1,0}$ for all k and $\gamma_m^{(A)} = 1.0 \times 10^{-3,-2,-1,0}$ for all m. The table shows that the *CWT-HTFD* outperformed the other algorithms in all measures on average. The comparison of *CWT-HTFD* (*CWT-HNMF*) with *STFT-HTFD* (*STFT-HNMF*, respectively) shows that the use of CWT is valid in SDR. The difference between *CWT-HTFD* and *CWT-HNMF* lies in the explicit incorporation of the spectral leakage into the model, we can confirm that this incorporation is effective for unsupervised monaural source separation by comparing the results of *CWT-HTFD* with those of *CWT-HNMF*.

4.7 Subjective Evaluation in Audio Quality of Separated Signals

Finally, we conducted a XAB test on the audio quality of separated audio signals to examine whether HTFD is also effective for human listening. Test signals were prepared as

Table 4.2: Average preference scores over all subjectives. The values in the parentheses are 95% confidence intervals.

Less similar / Similar	CWT-HTFD ($\sigma = \ln(2)/60$)	CWT-HTFD ($\sigma = \ln(2)/48$)	STFT-HNMF (64 ms frame)	STFT-HNMF (32 ms frame)
CWT-HTFD ($\sigma = \ln(2)/60$)	-	$0.6([0.39, 0.79])$	$0.96([0.80, 1.00])$	$0.95([0.80, 1.00])$
CWT-HTFD ($\sigma = \ln(2)/48$)	-	-	$1.0([0.86, 1.0])$	$1.0([0.86, 1.0])$
STFT-HNMF (64 ms frame)	-	-	-	$0.9([0.69, 0.97])$
STFT-HNMF (32 ms frame)	-	-	-	-

follows: We mixed a part of all pitch-wise signals (original signals or separated results) and used the mixed signal as a test signal for each musical piece. The chosen number of pitches for mixing was 30 % of the number of pitches contained in the corresponding musical piece, and the separated signals were chosen in descending order of the signal energy. The mixed signals of original pitch-wise signals were presented as X, and those of separated pitch-wise signals were presented as A and B. As the separated algorithms, we used CWT-HTFD with $\sigma = \ln(2)/48, \ln(2)/60$ and STFT-HNMF with 32 and 64 ms frames. The other parameters were the same as in Table 4.1. Each pair of four types of the test signals was presented to six listeners in random order. The listeners were asked to choose A or B such that is more similar to X. During the XAB test, the listeners were able to listen to all signals again and again.

Table 4.2 shows average preference scores and 95 % confidence interval for all pairs of the algorithms. The confidence interval was computed with the binomial test. These results demonstrate that CWT-HTFD yields significant improvements in the audio quality of separated signals subjectively, and the incorporation of the spectral leakage and CWT are also confirmed to be effective in audio quality of separated signals while increasing the number of subjectives is one of future works to increase the reliability of the result.

4.8 Summary

This chapter has proposed a new approach for monaural source separation called the "Harmonic-Temporal Factor Decomposition (HTFD)" by introducing a spectrogram model that combines the features of the models employed in the NMF and HTC approaches. We have further described some ideas how to design the prior distributions for the present model to incorporate musically relevant information into the separation scheme. We have experimentally confirmed the F_0 tracking ability of HTFD, the reduction of estimation errors with the key prior information. From a source separation experiment and a subjective experiment, we have confirmed that the explicit incorporation of the spectral leakage into the model and the use of CWT are valid for monaural source separation.

Chapter 5

Harmonic Temporal Factor Decomposition with Source-Filter Model

5.1 Chapter Overview

In the previous chapter, we proposed a new approach for monaural source separation called HTFD, which combines the features of the spectrogram models employed in two main approaches for monaural source separation. One approach involves applying NMF to an observed magnitude spectrogram interpreted as a non-negative matrix. The other approach is based on the concept of CASA. A CASA-based approach called HTC aims to cluster the time-frequency components of an observed signal based on a constraint designed according to the local time-frequency structure common in many sound sources (such as harmonicity and the continuity of frequency and amplitude modulations). In addition, taking into account the generative processes of many sound sources is also important to improve the accuracy of source separation. In this chapter, we incorporate the source-filter model defined in the discrete time domain, into the spectrogram model of HTFD, which defined in the CWT domain. To do this, we focus on that parameters of an analytic signal model are associated with those of the spectrogram model in HTFD and derive the explicit relationship between parameters of the source-filter model and the spectrogram model. Experimental evaluations show that the incorporation of the source-filter model is effective in monaural source separation.

5.2 Introduction

For monaural source separation two main approaches have thus far been adopted. One approach is based on the concept of computational auditory scene analysis, and the other approach involves applying NMF to an observed magnitude spectrogram interpreted as a non-negative matrix. The two approaches rely on different clues for making separation possible. Roughly speaking, the former approach focuses on the local time-frequency structure of each source, while the latter approach focuses on a relatively global structure of music spectrograms (such a property that a music signal typically consists of a limited number of recurring note events). Rather than discussing which clues are more useful, we believe that both of these clues can be useful for achieving a reliable monaural source separation algorithm. On the basis of this belief, we developed a new approach, called HTFD, for monaural source separation that combine the features of both HTC and NMF in Chapter 4.

To further improve the accuracy of source separation, the generative processes of musical instrument sounds is also important. With the source-filter theory, an instrument signal is considered to be an excitation signal filtered by a linear filter. The excitation signal corresponds to a vibrating object and varies with pitch. In contrast, the filter corresponds to resonance structure of the instrument and varies with timbre. F_0 varies with time during vibrato and portamento, whereas the spectral envelope associated with timbre is relatively static. This suggests that, incorporating the static property of the spectral envelope into the spectrogram model of HTFD suppress unnatural spectral shapes.

In this chapter, we incorporate the source-filter model defined in the discrete time domain into the CWT spectrogram model of HTFD. To do this, we focus on that parameters of an analytic signal model are associated with those of the spectrogram model in HTFD and derive the explicit relationship between parameters of the source-filter model and the spectrogram model. Based on the relationship, we reconstruct the generative model of an observed spectrogram and present an efficient algorithm consisting of closed-form update equations. Experiments examine the effect of the incorporation of the source-filter model to the accuracy of source separation.

5.3 Spectrogram Model of Music Signal

5.3.1 Continuous Wavelet Transform of Source Signal Model

As in the previous chapter, this section derives the CWT of a source signal. We first consider as a signal model for the sound of pitch k the analytic signal representation of a pseudo-periodic signal given by (4.1) where u denotes the time, $n\theta_k(u) + \varphi_{k,n}$ the instantaneous phase of the n-th harmonic and $a_{k,n}(u)$ the instantaneous amplitude. To derive the CWT of the analytic signal model, we define the wavelet basis function as Eq. (4.2) where α is the scale parameter such that $\alpha > 0$, t the shift parameter and $\psi(u)$ the mother wavelet with the center frequency of 1 satisfying the admissibility condition. $\psi_{\alpha,t}(u)$ can thus be used to measure the component of period α at time t. The CWT of $f_k(u)$ is then given by Eq. (4.3). Since the dominant part of $\psi^*_{\alpha,t}(u)$ is typically localized around time t, the result of the integral in Eq. (4.3) shall depend only on the values of $\theta_k(u)$ and $a_{k,n}(u)$ near t. This justifies to replace $\theta_k(t)$ and $a_{k,n}(t)$ with zero-and first-order approximations around time t (see Eqs. (4.5) and (4.7)). By using the above approximations and applying the Parseval's theorem, we can further write Eq. (4.3) as Eq. (4.8) where $x := \ln(1/\alpha)$ denotes log-frequency, $\Omega_k(t) := \ln \dot{\theta}_k(t)$ the logarithm of the instantaneous fundamental frequency (F_0), and Ψ the Fourier transform of ψ. The function Ψ can be chosen arbitrarily and thus as with [1], we employ the log-normal wavelet, defined by Eq. (4.9). Although we omit details of the derivation to avoid the duplicate description, the magnitude spectrogram of the analytic signal model $|W_k(x,t)|$ can be obtained approximately as Eq. (4.11) with the assumption that the time-frequency components are sparsely distributed so that the partials rarely overlap each other, which means that the magnitude spectra of the partials can approximately be considered additive. If we assume the additivity of magnitude spectra, the magnitude spectrogram of a superposition of K pitched sounds is given by the sum of Eq. (4.11) over k. It should be noted that this model is identical to the one employed in the HTC approach [1].

Although we have so far defined the spectrogram model above in continuous time and continuous log-frequency, we actually obtain observed spectrograms as a discrete time-frequency representation through computer implementations. Thus, we henceforth use $Y_{l,m} := Y(x_l, t_m)$ to denote an observed spectrogram where x_l ($l = 0, \ldots, L-1$) and t_m ($m = 0, \ldots, M-1$) stand for the uniformly-quantized log-frequency points and time

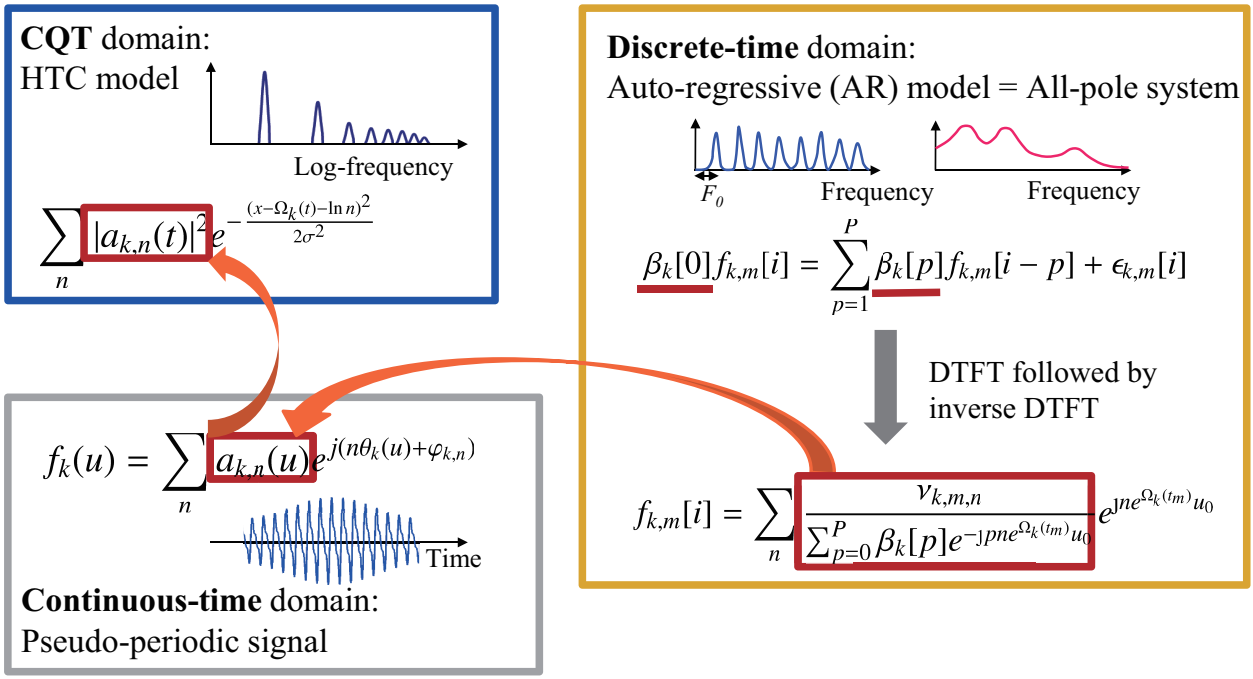


Figure 5.1: Schematic illustration of the incorporation of the source-filter model into the spectrogram model of HTFD.

points, respectively. We will also use the notation $\Omega_{k,m}$ and $a_{k,n,m}$ to indicate $\Omega_k(t_m)$ and $a_{k,n}(t_m)$.

5.3.2 Incorporating Source-Filter Model

The generating processes of many sound sources in real world can be explained fairly well by the source-filter theory. In this section, we follow the idea described in [73] to incorporate the source-filter model into the above model. The schematic illustration of the incorporation is shown in Fig. 5.1. Let us assume that each signal $f_k(u)$ within a short-time segment is an output of an all-pole system. That is, if we use $f_{k,m}[i]$ to denote the discrete-time representation of $f_k(u)$ within a short-time segment centered at time t_m, $f_{k,m}[i]$ can be described as

$$\beta_{k,m}[0] f_{k,m}[i] = \sum_{p=1}^{P} \beta_{k,m}[p] f_{k,m}[i-p] + \epsilon_{k,m}[i], \tag{5.1}$$

where i, $\epsilon_{k,m}[i]$, and $\beta_{k,m}[p]$ $(p=0,\ldots,P)$ denote the discrete-time index, an excitation signal, and the autoregressive (AR) coefficients, respectively. As we have already assumed in 5.3.1 that the F_0 of $f_{k,m}[i]$ is $e^{\Omega_{k,m}}$, to make the assumption consistent, the F_0 of the excitation

signal $\epsilon_{k,m}[i]$ must also be $e^{\Omega_{k,m}}$. We thus define $\epsilon_{k,m}[i]$ as

$$\epsilon_{k,m}[i] = \sum_{n=1}^{N} v_{k,n,m} e^{jne^{\Omega_{k,m}} iu_0}, \tag{5.2}$$

where u_0 denotes the sampling period of the discrete-time representation and $v_{k,n,m}$ denotes the complex amplitude of the nth partial. By applying the discrete-time Fourier transform (DTFT) to Eq. (5.1) and putting $B_{k,m}(z) := \beta_{k,m}[0] - \beta_{k,m}[1]z^{-1} \cdots - \beta_{k,m}[P]z^{-P}$, we obtain

$$F_{k,m}(\omega) = \frac{\sqrt{2\pi}}{B_{k,m}(e^{j\omega})} \sum_{n=1}^{N} v_{k,n,m} \delta(\omega - ne^{\Omega_{k,m}} u_0), \tag{5.3}$$

where $F_{k,m}$ denotes the DTFT of $f_{k,m}$, ω the normalized angular frequency, and δ the Dirac delta function. The inverse DTFT of Eq. (5.3) gives us another expression of $f_{k,m}[i]$:

$$f_{k,m}[i] = \sum_{n=1}^{N} \frac{v_{k,n,m}}{B_{k,m}(e^{jne^{\Omega_{k,m}} u_0})} e^{jne^{\Omega_{k,m}} iu_0}. \tag{5.4}$$

By comparing Eq. (5.4) and the discrete-time representation of Eq. (4.1), we can associate the parameters of the source filter model defined above with the parameters introduced in Sec. 5.3.1 through the explicit relationship:

$$|a_{k,n,m}| = \left| \frac{v_{k,n,m}}{B_{k,m}(e^{jne^{\Omega_{k,m}} u_0})} \right|. \tag{5.5}$$

5.3.3 Constraining Model Parameters

The key assumption behind the NMF model is that the spectra of the sound of a particular pitch is expressed as a multiplication of time-independent and time-dependent factors. In order to extend the NMF model to a more reasonable one, we consider it important to clarify which factors involved in the spectra should be assumed to be time-dependent and which factors should not. For example, the F_0 must be assumed to vary in time during vibrato or portamento. Of course, the scale of the spectrum should also be assumed to be time-varying (as with the NMF model). On the other hand, the timbre of an instrument can be considered relatively static throughout an entire piece of music.

We can reflect these assumptions in the present model in the following way. For convenience of the following analysis, we factorize $|a_{k,n,m}|$ into the product of two variables, $w_{k,n,m}$ and $U_{k,m}$

$$|a_{k,n,m}| = w_{k,n,m} U_{k,m}. \tag{5.6}$$

$w_{k,n,m}$ can be interpreted as the relative energy of the nth harmonic and $U_{k,m}$ as the time-varying normalized amplitude of the sound of the kth pitch such that $\sum_{k,m} U_{k,m} = 1$. In the same way, let us put $v_{k,n,m}$ as

$$v_{k,n,m} = \tilde{w}_{k,n,m} U_{k,m}. \tag{5.7}$$

Since the all-pole spectrum $1/|B_{k,m}(e^{j\omega})|$ is related to the timbre of the sound of the kth pitch, we want to constrain it to be time-invariant. This can be done simply by eliminating the subscript m. Eq. (5.5) can thus be rewritten as

$$w_{k,n,m} = \left| \frac{\tilde{w}_{k,n,m}}{B_k(e^{jne^{\Omega_{k,m}} u_0})} \right|. \tag{5.8}$$

We can use $\Omega_{k,m}$ as is since it is already dependent on m.

To sum up, we obtain a spectrogram model $X_{l,m}$ as

$$X_{l,m} = \sum_k \tilde{X}_{k,l,m} \tag{5.9}$$

$$\tilde{X}_{k,l,m} = \left(\sum_{n=1}^{N} w_{k,n,m} e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} \right) U_{k,m}, \tag{5.10}$$

where $\tilde{X}_{k,l,m}$ represents the spectrogram of pitch k. Note that the term inside the parenthesis, which corresponds to a spectral template in the NMF model, varies in time together with not only $\Omega_{k,m}$ but also $w_{k,m,n}$ compared to the spectrogram model of HTFD defined by Eq. (4.13).

5.3.4 Formulating Probabilistic Model

Since the assumptions and approximations we made so far do not always hold exactly in reality, an observed spectrogram $Y_{l,m}$ may diverge from $X_{l,m}$ even though the parameters are optimally determined. One way to simplify the process by which this kind of deviation occurs would be to assume a probability distribution of $Y_{l,m}$ with the expected value of $X_{l,m}$. Here, we assume that $Y_{l,m}$ follows a Poisson distribution with mean $X_{l,m}$

$$Y_{l,m} \sim \mathrm{Pois}(Y_{l,m}; X_{l,m}), \tag{5.11}$$

where $\mathrm{Pois}(z;\xi) = \xi^z e^{-\xi}/\Gamma(z)$. This defines our likelihood function

$$p(Y|\Theta) = \prod_{l,m} \mathrm{Pois}(Y_{l,m}; X_{l,m}), \tag{5.12}$$

where Y denotes the set consisting of $Y_{l,m}$ and Θ the entire set consisting of the unknown model parameters. It should be noted that the maximization of the Poisson likelihood with respect to $X_{l,m}$ amounts to optimally fitting $X_{l,m}$ to $Y_{l,m}$ by using the I-divergence as the fitting criterion.

Eq. (5.8) implicitly defines the conditional distribution $p(w|\tilde{w}, \beta, \Omega)$ expressed by the Dirac delta function

$$p(w|\tilde{w}, \beta, \Omega) = \prod_{k,n,m} \delta\left(w_{k,n,m} - \left|\frac{\tilde{w}_{k,n,m}}{B_k(e^{jne^{\Omega_{k,m}}u_0})}\right|\right). \tag{5.13}$$

The conditional distribution $p(w|\beta, \Omega)$ can thus be obtained by defining the distribution $p(\tilde{w})$ and marginalizing over $\tilde{w}$. If we now assume that the complex amplitude $\tilde{w}_{k,n,m}$ follows a circular complex normal distribution

$$\tilde{w}_{k,n,m} \sim \mathcal{N}_{\mathbb{C}}(\tilde{w}_{k,n,m}; 0, \nu^2),\ n = 1, \ldots, N, \tag{5.14}$$

where $\mathcal{N}_{\mathbb{C}}(z; 0, \xi^2) = e^{-|z|^2/\xi^2}/(\pi\xi^2)$, we can show, as in [73], that $w_{k,n,m}$ follows a Rayleigh distribution:

$$w_{k,n,m} \sim \text{Rayleigh}\left(w_{k,n,m}; \frac{\nu}{|B_k(e^{jne^{\Omega_{k,m}}u_0})|}\right), \tag{5.15}$$

where $\text{Rayleigh}(z; \xi) = (z/\xi^2)e^{-z^2/(2\xi^2)}$. This defines the conditional distribution $p(w|\beta, \Omega)$.

While the prior distributions defined in Eqs. (4.21), (4.24) and (4.23) can be used for prior distributions of $U_{k,m}$, we instead use a gamma distribution for the simplicity:

$$U_{k,m} \sim \text{Gam}(U_{k,m}; \alpha^{(U)}, \beta^{(U)}), \tag{5.16}$$

where $\alpha^{(U)} > 0$ and $\beta^{(U)} > 0$ are shape and rate parameters. Overall, the entire generative model is depicted in plate notation in Fig. 5.2.

After our conference paper [36], a similar spectrogram model that incorporates the source-filter model in the CWT domain has been presented [87]. The model can be derived from our spectrogram model by assuming that the magnitude and phase of $\tilde{w}_{k,n,m}$ follow the delta distribution and a uniform distribution over $[0, 2\pi)$, respectively, instead of Eq. (5.14).

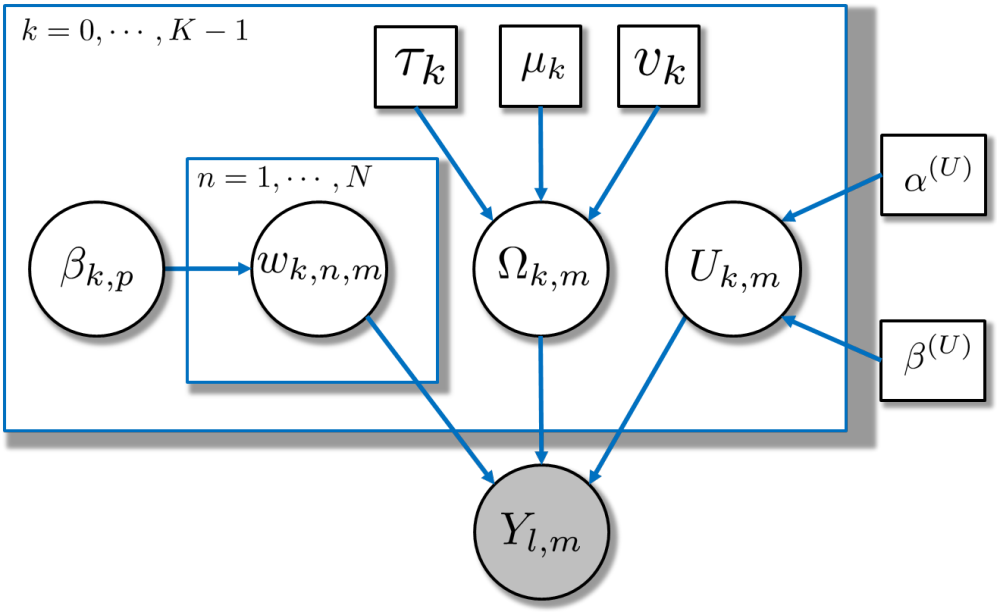


Figure 5.2: Plate notation of the proposed overall generative model.

5.4 Parameter Estimation Algorithm

The random variables in the present model are listed below:

$\Omega_{k,m}$: the logarithm of F_0 for pitch k at time m,

$w_{k,m,n}$: Relative energy of harmonic n of pitch k at time m,

$\beta_{k,p}$: the pth AR coefficient of pitch k,

$U_{k,m}$: Temporal activation of pitch k at time m.

We here denote the set of the random variables except for $w_{k,m,n}$ by Θ. For an observed magnitude spectrogram Y, we would like to find the estimates of Θ that maximizes the posterior density $p(\Theta|Y)$. The problem can be written as the problem of maximizing

$$\mathcal{J}(\Theta) := \ln p(Y|\Theta) + \ln p(\Theta), \tag{5.17}$$

with respect to Θ where

$$\ln p(\boldsymbol{Y}|\boldsymbol{\Theta}) = \ln \int_{\mathcal{W}} \prod_{l,m} \mathrm{Pois}(Y_{l,m}; X_{l,m}) p(w|\beta, \Omega) \mathrm{d}w \Big) \tag{5.18}$$

$$\ln p(\Theta) = \sum_{k} \ln p(\boldsymbol{\Omega}_k) + \ln p(U) \tag{5.19}$$

where $\mathcal{W}$ is the domain of w. Since the first term of Eq. (5.18) involves the integration of w, conventional smooth optimization techniques are difficult to use. However, we can

derive a computationally efficient algorithm consisting of closed-form equations for updating parameters by using a similar idea described in [1, 73].

When applying an auxiliary function approach to a certain maximization problem, the first step is to define a lower bound function for the objective function. As the abovementioned, the integration of w makes it difficult to derive closed-form update equations. The logarithm function is a concave function, and so we can invoke the Jensen's inequality to derive a lower bound function of Eq. (5.18) by introducing auxiliary variables $q(w)$ such that $\int_{\mathcal{W}} q(w)\mathrm{d}w = 1$ and $q(w) \geq 0$ for all w:

$$\ln p(Y|\Theta) \geq \int_{\mathcal{W}} q(w) \left(\ln \frac{\prod_{l,m} \mathrm{Pois}(Y_{l,m}; X_{l,m}) p(w|\beta, \Omega)}{q(w)} \right) \mathrm{d}w \tag{5.20}$$

$$= \int_{\mathcal{W}} q(w) \left(\sum_{l,m} \ln \mathrm{Pois}(Y_{l,m}; X_{l,m}) + \ln p(w|\beta, \Omega) - \ln q(w) \right) \mathrm{d}w \tag{5.21}$$

$$\underset{c}{=} \int_{\mathcal{W}} q(w) \left\{ \sum_{l,m} (Y_{l,m} \ln X_{l,m} - X_{l,m}) + \ln p(w|\beta, \Omega) - \ln q(w) \right\} \mathrm{d}w \tag{5.22}$$

where $\underset{c}{=}$ denotes equality up to constant terms. The equality of the inequality (5.20) holds if and only if

$$q(w) = p(w|Y, \Theta). \tag{5.23}$$

It should be noted that the derived lower bound can be seen as the Q function in the Expectation-Maximization (EM) algorithm. For the simplicity of notation, we hereafter denote the expectation of $w_{k,m,n}$, $w_{k,m,n}^2$ and $\ln w_{k,m,n}$over the auxiliary variables $q(w)$ by

$$\mathrm{E}[w_{k,m,n}] := \int_{\mathcal{W}} q(w) w_{k,m,n} \mathrm{d}w, \tag{5.24}$$

$$\mathrm{E}[w_{k,m,n}^2] := \int_{\mathcal{W}} q(w) w_{k,m,n}^2 \mathrm{d}w, \tag{5.25}$$

$$\mathrm{E}[\ln w_{k,m,n}] := \int_{\mathcal{W}} q(w) \ln w_{k,m,n} \mathrm{d}w. \tag{5.26}$$

Next, by using the fact that the logarithm function is a concave function, we can invoke the Jensen's inequality

$$\int_{\mathcal{W}} q(w) Y_{l,m} \ln X_{l,m} \mathrm{d}w \geq \int_{\mathcal{W}} q(w) Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \ln \frac{w_{k,n,m} e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{\lambda_{k,n,l,m}} \mathrm{d}w, \tag{5.27}$$

to obtain a lower bound function, where $\lambda_{k,n,l,m}$ is an auxiliary variable such that $\sum_{k,n} \lambda_{k,n,l,m} = 1$ for all l, m and $\lambda_{k,n,l,m} \geq 0$ for all k, n, l and m. The equality of the inequality (5.27) holds

if and only if

$$\lambda_{k,n,l,m} = \frac{e^{\mathrm{E}[\ln w_{k,n,m}]} e^{-\frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}}{X_{l,m}}. \tag{5.28}$$

Although one may notice that the second term in Eq. (5.18) is nonlinear in $\Omega_{k,m}$, the summation of $X_{l,m}$ over l can be approximated fairly well using the integral $\int_{-\infty}^{\infty} X(x, t_m)\mathrm{d}x$, as in Sec. 4.5, since $\sum_l X_{l,m}$ is the sum of the values at the sampled points $X(x_1, t_m), \ldots, X(x_L, t_m)$ with an equal interval, say Δ_x. Hence,

$$\begin{aligned}\sum_l X_{l,m} &\simeq \frac{1}{\Delta_x} \int_{-\infty}^{\infty} X(x, t_m)\mathrm{d}x \\ &= \frac{\sqrt{2\pi}\sigma}{\Delta_x} \sum_k U_{k,m} \sum_n w_{k,n,m}.\end{aligned} \tag{5.29}$$

This approximation implies that the second term in Eq. (5.18) depends little on $\Omega_{k,m}$. Furthermore, we focus on the fact that a quadratic function tangent to the absolute value function is an upper bound of the absolute value function. By writing the tangent point as $\xi_{k,n,m} \in \mathbb{R}_{\geq 0}$, the lower bound can be specifically written as

$$-\frac{\sqrt{2\pi}\sigma}{\Delta_x} \sum_k U_{k,m} \sum_n w_{k,n,m} \geq -\frac{\sqrt{2\pi}\sigma}{\Delta_x} \sum_k U_{k,m} \sum_n \frac{\xi_{k,n,m}^{-1} w_{k,n,m}^2 + \xi_{k,n,m}}{2}. \tag{5.30}$$

The equality holds if and only if $\xi_{k,n,m} = w_{k,n,m}$.

The auxiliary function can thus be approximately written as

$$\begin{aligned}&\mathcal{L}^+(\Theta, q(w), \lambda) \\ &\underset{c}{=} \sum_{l,m} Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \left(\mathrm{E}[\ln w_{k,n,m}] + \ln U_{k,m} - \frac{(x_l - \Omega_{k,m} - \ln n)^2}{2\sigma^2} - \ln \lambda_{k,n,l,m} \right) \\ &\quad - \frac{\sqrt{2\pi}\sigma}{\Delta_x} \sum_m \sum_k U_{k,m} \sum_n \frac{\xi_{k,n,m}^{-1} \mathrm{E}[w_{k,n,m}^2] + \xi_{k,n,m}}{2} + \sum_{k,n,m} \left(\mathrm{E}[\ln w_{k,n,m}] - \ln \frac{\nu^2}{|B_k(e^{\mathrm{j}ne^{\Omega_{k,m}}} u_0)|^2} \right. \\ &\quad \left. - \frac{\mathrm{E}[w_{k,n,m}^2] |B_k(e^{\mathrm{j}ne^{\Omega_{k,m}}} u_0)|^2}{2\nu^2} \right) + \ln p(\Theta).\end{aligned} \tag{5.31}$$

where λ denotes the set consisting of $\lambda_{k,n,l,m}$. We can derive update equations for the model parameters, using the above auxiliary function. First, the update of $q(w)$ is the calculation of the posterior distribution of w, and thus

$$q(w_{k,m,n}) \leftarrow \text{Nakagami}\left(w_{k,m,n}; \frac{\sum_{l,m} Y_{l,m} \lambda_{k,n,l,m}}{2} + 1, \frac{\frac{\sum_l Y_{l,m} \lambda_{k,n,l,m}}{2} + 1}{\frac{\sqrt{2\pi} U_{k,m} \sigma}{\xi_{k,m,n} \Delta_x} + \frac{|B_k(e^{\mathrm{j}ne^{\Omega_{k,m}}} u_0)|^2}{2\nu^2}} \right) \tag{5.32}$$

where Nakagami$(\zeta; a, b)$ denotes the Nakagami distribution:

$$\text{Nakagami}(\zeta; a, b) = \frac{2a^a}{\Gamma(a)b^a}\zeta^{2a-1}e^{-a\zeta^2/b}. \tag{5.33}$$

By setting at zero the partial derivative of $\mathcal{J}^+(\Theta, q(w), \lambda)$ with respect to $U_{k,m}$, we obtain

$$U_{k,m} \leftarrow \frac{\sum_{l,m} Y_{l,m} \sum_n \lambda_{k,n,l,m} + \alpha^{(U)} - 1}{\sum_{m,n} \frac{\sqrt{2\pi}\sigma}{\xi_{k,m,n}\Delta_x} \mathrm{E}[w_{k,m,n}^2] + \beta^{(U)}} \tag{5.34}$$

If we ignore the effect of $\ln p(w|\beta, \Omega)$ to the update of Ω, the update equation of Ω is written as

$$\boldsymbol{\Omega}_k \leftarrow \left(\frac{\alpha_{\mathrm{l}}}{\tau^2}D + \frac{\alpha_{\mathrm{g}}}{\upsilon_k^2}\boldsymbol{I}_M + \sum_{n,l}\mathrm{diag}(\boldsymbol{p}_{k,n,l})\right)^{-1}\left(\mu_k\frac{\alpha_{\mathrm{g}}}{\upsilon_k^2}\boldsymbol{1}_M + \sum_{n,l}(x_l - \ln n)\boldsymbol{p}_{k,n,l}\right), \tag{5.35}$$

$$\boldsymbol{p}_{k,n,l} := \frac{1}{\sigma^2}\left[Y_{l,1}\lambda_{k,n,l,1}, Y_{l,2}\lambda_{k,n,l,2}, \cdots, Y_{l,M}\lambda_{k,n,l,M}\right]^\top, \tag{5.36}$$

where $\mathrm{diag}(\boldsymbol{p})$ converts a vector $\boldsymbol{p}$ into a diagonal matrix with the elements of $\boldsymbol{p}$ on the main diagonal.

As for the update equations for the AR coefficients $\boldsymbol{\beta}$, we can invoke the method described in [54] with a slight modification since the terms in the auxiliary function that depend on $\boldsymbol{\beta}$ has the similar form as the objective function defined in [54]. For the simplicity of notation, we rewrite $|B_k(\omega)|^2$ as

$$|B_k(\omega)|^2 = \boldsymbol{\beta}_k^\top C(\omega)\boldsymbol{\beta}_k \tag{5.37}$$

$$\boldsymbol{\beta}_k = [\beta_{k,0}, \cdots, \beta_{k,P}]^\top \tag{5.38}$$

where $C(\omega)$ is a $(P+1)\times(P+1)$ Toeplitz matrix whose (p,q)-th entry is given by

$$(C(\omega))_{p,q} = \cos(\omega(p-q)). \tag{5.39}$$

With this notation, the partial derivative of $\mathcal{J}^+$ with respect to $\boldsymbol{\beta}_k$ is written as

$$\frac{\partial\mathcal{J}^+(\Theta, q(w), \lambda)}{\partial\boldsymbol{\beta}_k} = (R_k(\boldsymbol{\beta}_k) - R_k)\boldsymbol{\beta}_k. \tag{5.40}$$

$$R_k = \sum_{m,n}\frac{\mathrm{E}[w_{k,m,n}^2]}{\nu^2}C(e^{\mathrm{j}ne^{\Omega_{k,m}}u_0}) \tag{5.41}$$

$$\hat{R}_k(\boldsymbol{\beta}_k) = \sum_{m,n}\frac{2\nu^2}{\boldsymbol{\beta}_k^\top C(e^{\mathrm{j}ne^{\Omega_{k,m}}u_0})\boldsymbol{\beta}_k}C(e^{\mathrm{j}ne^{\Omega_{k,m}}u_0}) \tag{5.42}$$

Both the first and second term in Eq. (5.40) are positive definite matrices, and thus a multiplicative gradient ascent algorithm can be used similarly in [54, 88, 89] (see [54, 90] for the proof of the local convergence of the algorithm). $\mathcal{J}^+$ can be increased by the following update:

$$\boldsymbol{\beta}_k \leftarrow R_k^{-1} \hat{R}_k(\boldsymbol{\beta}_k) \boldsymbol{\beta}_k. \tag{5.43}$$

To avoid the indeterminacy in scales of the filters, $\boldsymbol{\beta}_k$ is normalized at each update such that $\beta_{k,p} = \beta_{k,p}/\beta_{k,0}$ for all k and p. For stability of the filters, all absolute values of the roots of $\sum_{p=0}^{P} \beta_{k,p} z^{P-p}$ are forced to be below 0.97 for all k. Although these steps violate the convergence of the algorithm, we found empirically that the steps increase the numerical stability of the algorithm.

5.5 Experiments

To examine the effect of the incorporation of the source-filter model, We conducted an experiment on monaural source separation and measured SDRs, SIRs and SARs, which were calculated with the BSSEval toolbox [75]. Similarly in Sec. 4.6, the magnitude CWT spectrograms were sampled every ten time points such that $t_m - t_{m-1} \simeq 10$ ms and then used for the estimation. For the separation, we designed a time-frequency mask as $\tilde{X}_{k,l,m}/X_{l,m}$ followed by linearly interpolating it. Here we used $\mathrm{E}[w_{k,n,m}^2]$ as the estimate of $w_{k,n,m}^2$.

The test data was the same as in Sec. 4.6. To compute CWT spectrograms of the signals, the fast approximate CWT algorithm was used with center frequencies ranging from 27.5 to 7994 Hz per 10 cent and a time shift of 1.832 ms. As an analyzing wavelet, we used the log-normal wavelet with $\sigma = 0.0116$, which corresponds to one fifth of a semitone interval. The parameters of the proposed algorithm was set as follows: $N = 20$, $\alpha_{\mathrm{g}} = \alpha_{\mathrm{l}} = 1$, $\nu = 1$, $\tau_k = \ln(2)/72 \times 9.16$, $v_k = \ln(2)/36$ and $\mu_k = \ln(27.5) + (k-1)\ln(2)/12$ for all k (A1 to A♯7, i.e. $K = 88$). The algorithm was stopped after 100 iterations. For the convenience of the implementation, we approximate $\mathrm{E}[w_{k,m,n}]$ and $\mathrm{E}[\ln w_{k,m,n}]$ with zero-order approximations as $\mathrm{E}[w_{k,m,n}] \simeq (\mathrm{E}[w_{k,m,n}^2])^{1/2}$ and $2\mathrm{E}[\ln w_{k,m,n}] \simeq \ln \mathrm{E}[w_{k,m,n}^2]$, respectively.

Table 5.1 compares SDR improvements, SIR improvements and SARs obtained with the proposed algorithm and HTFD. The results of the proposed algorithm were for $(\alpha^{(U)}, \beta^{(U)}) = (\epsilon, \epsilon)$. which scored the highest average SDR improvement in the settings $\alpha^{(U)} = \epsilon, 1.0 \times 10^{-3}, 1.0 \times 10^{-2}, 1.0 \times 10^{-1}$ and 1.0 and $\beta^{(U)} = \epsilon$. Here ϵ is the distance from 1.0 to the next

Table 5.1: Average SDR improvements, SIR improvements and SARs [dB] with standard errors obtained with the proposed algorithm (Proposed) with varying P and HTFD for overall data. The result of HTFD is the same as in Table 4.1.

Algorithm	P	SDR improvement	SIR improvement	SAR
	16	16.13 ± 0.72	26.55 ± 0.98	0.31 ± 0.43
Proposed	32	17.58 ± 0.56	$\mathbf{26.94 \pm 0.83}$	$\mathbf{1.16 \pm 0.41}$
	48	$\mathbf{17.60 \pm 0.58}$	$\mathbf{26.94 \pm 0.82}$	1.07 ± 0.40
	60	17.42 ± 0.56	26.84 ± 0.83	0.92 ± 0.40
HTFD	-	16.42 ± 0.62	26.07 ± 0.88	0.27 ± 0.39

double-precision number, i.e. $\epsilon = 2^{-52}$. The results of HTFD were the same in Tab. 4.1. The improvement in SAR indicates that the results obtained with the proposed algorithm have less artifacts, which is caused by the algorithm, than HTFD, and thus the proposed algorithm provides separation results whose spectral shapes are similar to realistic musical instruments. From these results, we confirmed that the incorporation of the source-filter model improves monaural source separation.

5.6 Summary

This chapter have incorporated the source-filter model into the spectrogram model of HTFD to improve the accuracy of monaural source separation. We have focused that parameters of the spectrogram model of HTFD in the CWT domain can be associated with those of an analytic signal model in the time domain, and have obtained the explicit relationship of parameters between the CWT domain and the discrete-time domain in which the source-filter model is defined as the AR model. We have experimentally confirmed that the incorporation of the source-filter model improves the accuracy of monaural source separation.

Chapter 6

Fast Signal Reconstruction from Magnitude Spectrogram of Continuous Wavelet Transform

6.1 Chapter Overview

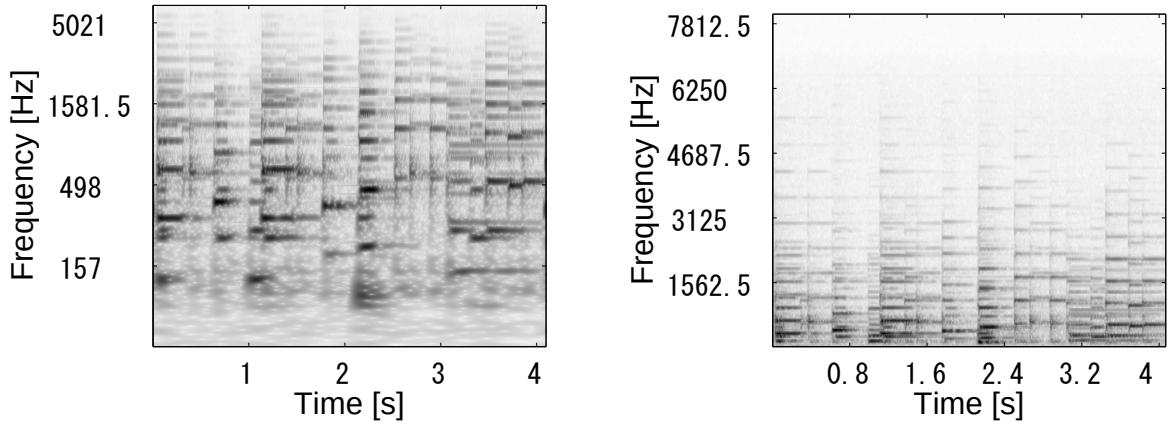
The complex spectrograms obtained with typical time-frequency transforms are redundant representations of a time domain signal. This means there is a certain condition that the spectrograms must satisfy to ensure they correspond to a time domain signal. We say that the complex spectrograms satisfying this condition are *consistent*. This chapter deals with the problem of estimating an unknown signal from a given magnitude spectrogram obtained with the CWT, based on a consistency criterion. A signal that is likely to yield the given magnitude spectrogram can be found by an iterative algorithm consisting of initializing the phase spectrogram estimate, performing inverse CWT followed by CWT, and replacing the magnitude part of the updated CWT spectrogram with the given magnitude spectrogram while leaving the phase part unchanged. Since CWT and inverse CWT can be computationally expensive, we may wish to accelerate the algorithm by employing some fast algorithms for computing CWT and inverse CWT. Here, when invoking a method to approximate CWT and inverse CWT, it is not always clear whether the convergence of the entire iterative algorithm is still guaranteed. The aim of this chapter is to show using the auxiliary function principle that the use of a particular type of an approximate algorithm does not affect the monotonicity of the convergence of the entire iterative algorithm. Experimental evaluations

show that the devised fast algorithms are around 75 times faster than the conventional algorithm while the reconstructed signals obtained with the proposed algorithms have almost the same audio quality as original sounds.

6.2 Introduction

The continuous wavelet transform (CWT), also known as the constant-Q transform, provides a time-frequency representation of a time domain signal with a logarithmically uniform frequency resolution. This agrees well with the human auditory system particularly at the high-frequency end, which may be one reason why the fundamental frequencies of the semitones on the musical scale are logarithmically spaced. This characteristic is in contrast to the short-time Fourier transform (STFT), which gives a spectrogram with a linearly uniform frequency resolution (Fig. 6.1). Thus, to develop auditory-motivated audio signal processing methods, one promising approach would be to model, analyze and modify the spectrogram given by the CWT (CWT spectrogram). Indeed, recent studies have reported that using the CWT instead of the STFT significantly improves the performances of source separation [7], multiple fundamental frequency estimation [1,8,9] and singing voice separation [10]. Specifically for those applications in which the aim is to generate audio signals, we must be able to construct a time domain signal from an estimated or modified magnitude CWT spectrogram, in which phase information is missing. To this end, this chapter addresses the problem of constructing a time-domain signal by estimation an appropriate phase from a magnitude CWT spectrogram.

A phase estimation algorithm for a magnitude CWT spectrogram has already been proposed by Irino *et al.* [60], which consists in iteratively performing the CWT and the inverse CWT. At each iteration, the magnitude part of the updated CWT spectrogram is replaced by the given magnitude CWT spectrogram while leaving the phase part unchanged. Since the CWT has a large computational cost, Irino's algorithm requires a long processing time for computation, which has been a serious obstacle for its practical uses. Thus, we consider it necessary to develop a faster algorithm. The convergence of the algorithm as well as the computational cost is an important issue. Efficient methods for computing the CWT and the inverse CWT have been recently proposed [61–64]. It may appear that simply carrying out one of these methods for the CWT and inverse CWT steps in Irino's algorithm would reduce the computational cost. However, it is not clear whether the convergence of such an



(a) CWT spectrogram. (b) STFT spectrogram.

Figure 6.1: Examples of spectrograms of a music audio signal given by (a) CWT and (b) STFT. A CWT (STFT) spectrogram has an equal resolution on a log-frequency scale (linear frequency scale, respectively).

algorithm is guaranteed.

Le Roux *et al.* have thus far proposed a fast algorithm for estimating the phase from a magnitude STFT spectrogram [58, 59] by using the fact that the waveforms in the overlapping part of consecutive frames must be consistent. This implies the fact that an STFT spectrogram is a redundant representation when the hop-size is shorter than the frame length and thus it satisfies a certain condition that it corresponds to a time domain signal. We have referred to this condition as *the consistency condition*. The problem of estimating the phase from a magnitude STFT spectrogram can be formulated as an optimization problem of minimizing the consistency criterion that describes how far an arbitrary complex array deviates from this condition. This formulation has provided a new insight into the well-known algorithm proposed in [57], allowing us to derive a fast approximate algorithm and give a very intuitive proof of its convergence. Since a CWT spectrogram is also a redundant representation, we can conjecture that we can develop a fast approximate method for estimating the phase from a magnitude CWT spectrogram with guaranteed convergence in the same way by using the concept of the spectrogram consistency.

This chapter presents two fast algorithms for estimating the phase from a magnitude CWT spectrogram. First, we first introduce a consistency condition for a CWT spectrogram and give its intuitive interpretation in analogy to the case of an STFT spectrogram (Sec. 6.3). Second, we formulate the phase estimation problem as an optimization problem based on the consistency condition, and derived an iterative algorithm based on an optimization principle

called the auxiliary function approach (Sec. 6.4). It becomes clear that the algorithm is equivalent to Irino's algorithm and gives a very clear proof of its convergence, though it should be noted that the proof of the convergence has already been mentioned in [91]. Third, on the basis of our proof, we show that the convergence of the iterative procedure is guaranteed if the CWT and inverse CWT steps are replaced with any linear and redundant transform and its inverse transform. Two efficient algorithms with guaranteed convergence are then presented based on a fast approximate method for computing the CWT [61, 62] (Sec. 6.5). Finally, we evaluate the efficiency and the signal reconstruction property of the proposed algorithms compared to Irino's algorithm through experiments on real audio signals (Sec. 6.6). The evaluations also show a trade off of the proposed algorithms between the approximation accuracy and the audio quality of reconstructed signals.

6.3 Spectrogram Consistency

6.3.1 Continuous Wavelet Transform

The CWT represents a time domain signal as a summation of wavelet basis waveforms, also known as analyzing wavelets, whose periods (the reciprocals of the center frequencies) correspond to a scale parameter. We here consider discretizing the scale parameter such that the center frequencies of the wavelet basis waveforms are equally spaced on a log-frequency scale. Let $l = 0, 1, \cdots, L-1$ and $m = 0, 1, \cdots, M-1$ be the indices of scale and time shift parameters, respectively, where L is the number of the discretized scale parameters and M is the length of an input signal. Given a discrete time domain signal $\boldsymbol{f} = [f_0, f_1, \cdots, f_{M-1}]^\top \in \mathcal{F} := \{\boldsymbol{f}; \boldsymbol{f} \in \mathbb{C}^M, \sum_m f_m = 0\}$, the component of a CWT spectrogram associated with scale $a_l > 0$, arranged as $\boldsymbol{s}_l = [s_{l,0}, s_{l,1}, \cdots, s_{l,M-1}]^\top$, is defined as

$$\boldsymbol{s}_l = W_l \boldsymbol{f}, \tag{6.1}$$

$$W_l := \begin{bmatrix} \psi^*_{l,0} & \psi^*_{l,M-1} & \cdots & \psi^*_{l,1} \\ \psi^*_{l,1} & \psi^*_{l,0} & \cdots & \psi^*_{l,2} \\ \vdots & \vdots & \ddots & \vdots \\ \psi^*_{l,M-1} & \psi^*_{l,M-2} & \cdots & \psi^*_{l,0} \end{bmatrix}. \tag{6.2}$$

Here $\psi^*_{l,m}$ is the complex conjugate of the wavelet basis waveform $\psi_{l,m} := \psi(t\Delta/a_l)/a_l$, where Δ denotes the sampling period of the input signal, $\psi(t\Delta)$ is a mother wavelet satisfying the admissibility condition. Each row of W_l contains the wavelet basis waveform of scale a_l with a different time shift parameter. Then, the CWT spectrogram $\boldsymbol{s} = [\boldsymbol{s}_0^\top, \boldsymbol{s}_1^\top, \cdots, \boldsymbol{s}_{L-1}^\top]^\top$ is given as

$$\boldsymbol{s} = W\boldsymbol{f}, \tag{6.3}$$

where W denotes the CWT matrix, defined as

$$W = [W_0^\top, W_1^\top, \cdots, W_{L-1}^\top]^\top. \tag{6.4}$$

Whether the inverse CWT of $W\boldsymbol{f}$ equals to $\boldsymbol{f}$ for all $\boldsymbol{f} \in \mathcal{F}$ depends on W. For simplicity, we hereafter assume that the equality holds. It is important to note that the following discussion is valid if the equality does not hold.

The inverse CWT can be defined by the pseudo inverse of W, defined as W^+, and the inverse of $\boldsymbol{s}$ is given as $W^+\boldsymbol{s}$. This implicitly means that the inverse CWT of $\boldsymbol{s}$ is the solution to the following minimization problem:

$$\underset{\tilde{\boldsymbol{f}} \in \mathcal{F}}{\operatorname{argmin}} \|\boldsymbol{s} - W\tilde{\boldsymbol{f}}\|_2^2, \tag{6.5}$$

where $\|\boldsymbol{s}\|_2 = \sqrt{\sum_{l,m} s_{l,m}^2}$ denotes the l^2 norm of $\boldsymbol{s}$.

6.3.2 Consistency Condition and Relation to Phase Estimation

As can be seen from Eq. (6.3), $\boldsymbol{s}$ belongs to the subspace $\mathcal{W}$ spanned by the column vectors of W. While the CWT spectrogram of a signal (i.e., a complex vector that belongs to $\mathcal{W}$) will be mapped to itself by applying the inverse CWT followed by the CWT, a complex vector that does not belong to $\mathcal{W}$ will not come back to the same point but will be projected onto the nearest point in $\mathcal{W}$ (Fig. 6.2). We can thus define a condition for a complex vector to be "consistent" (in the sense that it corresponds to a CWT spectrogram of a signal) as follows:

$$\boldsymbol{0}_{LM} = \boldsymbol{s} - WW^+\boldsymbol{s}, \tag{6.6}$$

where $\boldsymbol{0}_{LM}$ denotes an LM-dimensional zero vector. It is important to note that when W is replaced with a matrix in which each row is a basis waveform of the STFT, Eq. (6.6) becomes the consistency condition for an STFT spectrogram proposed in [59].

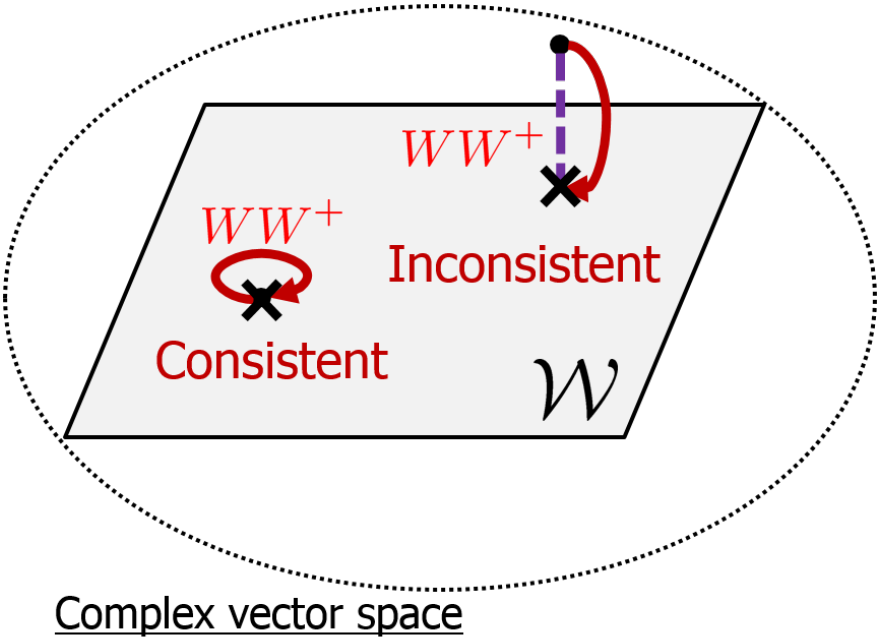


Figure 6.2: Consistent and inconsistent examples based on the concept of spectrogram consistency. The red curves represent WW^+ and the gray plane depicts the subspace spanned $\mathcal{W}$ by the column vectors of W. See text.

When given a magnitude CWT spectrogram, we can construct a signal by assigning phase to it to obtain a complex CWT spectrogram $\boldsymbol{s}$, and applying the inverse CWT, i.e., $W^+\boldsymbol{s}$. Here, if we assign "inconsistent" phase to the given magnitude CWT spectrogram, the complex CWT spectrogram $\boldsymbol{s}$ will not belong to $\mathcal{W}$ and so the CWT spectrogram of the constructed signal, $WW^+\boldsymbol{s}$, will be different from $\boldsymbol{s}$. As we want to equal the magnitude CWT spectrogram of the constructed signal to the given magnitude CWT spectrogram, we must find "consistent" phase such that $\boldsymbol{s}$ satisfies the consistency condition.

6.3.3 Intuitive Understanding of Consistency Condition

To obtain an intuitive understanding of the consistency condition, we consider the filter-bank interpretation of the CWT. The CWT can be thought of as a filter bank with subband filters whose impulse responses are given by the scaled analyzing wavelets. The filter bank does not pass all frequency components of a signal and blocks at least the DC component to satisfy the admissibility condition.

Now, by applying the M-point discrete Fourier transform (DFT) to each block of Eq. (6.6),

Eq. (6.6) can be written equivalently as

$$0 = \hat{\boldsymbol{s}} - \hat{W}\hat{W}^{+}\hat{\boldsymbol{s}}, \tag{6.7}$$

$$\hat{\boldsymbol{s}} := [\hat{\boldsymbol{s}}_0^{\top}, \hat{\boldsymbol{s}}_1^{\top}, \cdots, \hat{\boldsymbol{s}}_{L-1}]^{\top}, \tag{6.8}$$

$$\hat{W} := \begin{bmatrix} \hat{W}_0^{\top} & \hat{W}_1^{\top} & \cdots & \hat{W}_{L-1}^{\top} \end{bmatrix}^{\top}, \tag{6.9}$$

$$\hat{W}_l := F_M W_l F_M^{\mathsf{H}}, \tag{6.10}$$

where $\hat{\boldsymbol{s}}_l$ denotes the DFT of $\boldsymbol{s}_l$, $\hat{W}^{+}$ is the pseudo inverse of $\hat{W}$ and F_M^{H} is the Hermitian transpose of the M-point DFT matrix F_M. Since W_l is a circulant matrix, W_l is diagonalized by F_M and F_M^{H}. The diagonal elements of $\hat{W}_l$ represent the frequency response of the subband filter associated with scale a_l:

$$\hat{W}_l = \begin{bmatrix} \hat{\psi}_{l,0}^{*} & & & \\ & \hat{\psi}_{l,1}^{*} & & \\ & & \ddots & \\ & & & \hat{\psi}_{l,M-1}^{*} \end{bmatrix} \tag{6.11}$$

where $\{\hat{\psi}_{l,k}\}_{k=0}^{M-1}$ is the DFT of $\{\psi_{l,m}\}_{m=0}^{M-1}$ and k is the angular frequency index. Eq. (6.7) is explicitly written as

$$0 = \hat{s}_{l,k} - \frac{1}{\sum_{l'} |\hat{\psi}_{l',k}|^2} \sum_{l'} \hat{\psi}_{l,k}^{*} \hat{\psi}_{l',k} \hat{s}_{l',k}, \tag{6.12}$$

for $k = 1, \cdots, M-1$, where l' is the index of the scale parameter.

If the subbands of the filter bank overlap each other (more precisely, if there exists a pair of channels such that the product of their frequency responses is non-zero at every non-zero frequency), i.e. $\forall k = 1, \cdots, M-1, \exists l \neq l', \hat{\psi}_{l,k}^{*} \hat{\psi}_{l',k} \neq 0$, Eq. (6.6) becomes a nontrivial condition for a complex vector $\boldsymbol{s} \in \mathbb{C}^{LM}$ to correspond to a consistent CWT spectrogram. Otherwise, all the elements of $\mathbb{C}^{LM}$ trivially satisfy Eq. (6.6), implying that the consistency condition cannot be used as a criterion for phase estimation. Therefore, care must be taken in choosing the quantization intervals of the scale parameter and the type of the analyzing wavelet. The Morlet [92], the log-normal wavelet [1] and the wavelets used in the auditory wavelet transform [60] satisfy the above requirement when the quantization intervals of the scale parameter are appropriately chosen. We hereafter assume that the filter bank satisfies $\forall k = 1, \cdots, M-1, \exists l \neq l', \hat{\psi}_{l,k}^{*} \hat{\psi}_{l',k} \neq 0$.

The requirement for the subbands of the CWT to overlap each other is analogous to the requirement for the short time frames of the STFT to overlap. The consistency condition of STFT spectrograms can be understood as implying that the waveforms within the overlapping segment of consecutive frames must be consistent [59]. The consistency condition of CWT spectrograms, on the other hand, can be interpreted as implying that the outputs of adjacent channels within the overlapping subbands must be consistent.

6.4 Phase Estimation Based on CWT Spectrogram Consistency

6.4.1 Formulation of Phase Estimation Problem

Assume that we are given a magnitude CWT spectrogram, arranged as a LM-dimensional non-negative vector $\boldsymbol{a}$. We would like to estimate the phase of the given magnitude CWT spectrogram such that it meets the consistency condition. To allow for any LM-dimensional non-negative vector as the input, we here formulate the problem as that of finding a phase estimate $\boldsymbol{\phi} \in [-\pi, \pi)^{LM}$ that minimizes the consistency criterion

$$\mathcal{I}(\boldsymbol{\phi}) := \|\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) - WW^{+}\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})\|_2^2, \tag{6.13}$$

where $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ denotes the estimated CWT spectrogram, defined by

$$\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) := \boldsymbol{a} \odot \begin{bmatrix} e^{\mathrm{j}\phi_{0,0}} \\ e^{\mathrm{j}\phi_{0,1}} \\ \vdots \\ e^{\mathrm{j}\phi_{L-1,M-1}} \end{bmatrix}. \tag{6.14}$$

Here the operator $\odot$ denotes the element-wise product and $\phi_{l,m}$ is the element of $\boldsymbol{\phi}$ associated with scale a_l and time shift t. $\mathcal{I}(\boldsymbol{\phi})$ describes how far $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ deviates from the consistency condition (Fig. 6.3). Namely, the more consistent $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ becomes, the smaller $\mathcal{I}(\boldsymbol{\phi})$ becomes. $\mathcal{I}(\boldsymbol{\phi}) = 0$ indicates that $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ lies in the intersection of the set of consistent CWT spectrograms and the set of complex vectors that equal $\boldsymbol{a}$ up to a phase factor.

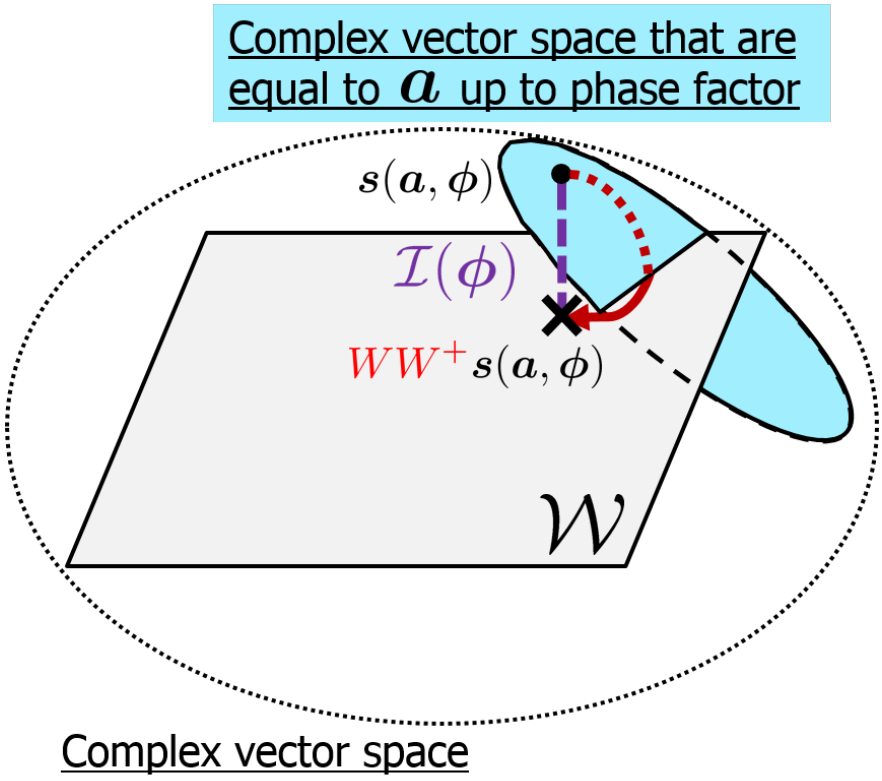


Figure 6.3: The consistency criterion $\mathcal{I}(\boldsymbol{\phi})$ (purple broken line) can be seen as a distance from $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ to the subspace $\mathcal{W}$ spanned by the column vectors of W. The red curves and the gray plane is the same in Fig. 6.2.

6.4.2 Iterative Algorithm with Auxiliary Function Approach

If we treat each $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ itself as the parameter, denoted by $\boldsymbol{s}$, the above minimization problem can be converted into the following quadratic programming with quadratic constraints:

$$\min_{\boldsymbol{s} \in \mathbb{C}^{LM}} \boldsymbol{s}^{\mathsf{H}}(I_{LM} - WW^{+})\boldsymbol{s} \tag{6.15}$$

$$\text{subject to } |s_{l,m}|^2 = a_{l,m}^2 \text{ for } \forall l, m. \tag{6.16}$$

The problem is difficult to solve with the method of Lagrange multiplier since more than a fourth order equation of the Lagrange multiplier need to be solved.

Conventional convex optimization techniques can be theoretically applicable to the problem. However, the number of rows and columns of WW^{+} are usually very large and some conventional techniques require large computational cost. To suppress it, we can invoke the auxiliary function approach [93] to derive an iterative algorithm that searches for the estimate of $\boldsymbol{\phi}$. To apply the auxiliary function approach to the current minimization problem,

the first step is to construct an auxiliary function $\mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ satisfying $\mathcal{I}(\boldsymbol{\phi}) = \min_{\tilde{\boldsymbol{s}}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$. We refer to $\tilde{\boldsymbol{s}}$ as an auxiliary variable. It can then be shown that $\mathcal{I}(\boldsymbol{\phi})$ is non-increasing under the updates $\boldsymbol{\phi} \leftarrow \operatorname{argmin}_{\boldsymbol{\phi}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ and $\tilde{\boldsymbol{s}} \leftarrow \operatorname{argmin}_{\tilde{\boldsymbol{s}}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$. Thus, $\mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ should be designed as a function that can be minimized analytically with respect to $\boldsymbol{\phi}$ and $\tilde{\boldsymbol{s}}$. Such a function can be constructed as follows.

Recall that the operator WW^+ is an orthogonal projection onto $\mathcal{W}$ and so $WW^+\boldsymbol{s}$ indicates the closest point in $\mathcal{W}$ from $\boldsymbol{s}$. Thus, we obtain

$$\mathcal{I}(\boldsymbol{\phi}) = \min_{\tilde{\boldsymbol{f}} \in \mathcal{F}} \|\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) - W\tilde{\boldsymbol{f}}\|_2^2 \tag{6.17}$$

$$= \min_{\tilde{\boldsymbol{s}} \in \mathcal{W}} \|\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) - \tilde{\boldsymbol{s}}\|_2^2. \tag{6.18}$$

We can confirm that

$$\mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}}) := \|\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) - \tilde{\boldsymbol{s}}\|_2^2, \quad \tilde{\boldsymbol{s}} \in \mathcal{W}, \tag{6.19}$$

satisfies $\mathcal{I}(\boldsymbol{\phi}) = \min_{\tilde{\boldsymbol{s}} \in \mathcal{W}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$. Eq. (6.19) can be used as an auxiliary function for $\mathcal{I}(\boldsymbol{\phi})$, and we can monotonically decrease $\mathcal{I}(\boldsymbol{\phi})$ by iteratively performing $\tilde{\boldsymbol{s}} \leftarrow \operatorname{argmin}_{\tilde{\boldsymbol{s}}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ and $\boldsymbol{\phi} \leftarrow \operatorname{argmin}_{\boldsymbol{\phi}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$. Here the iterative updates can be written explicitly as

$$\tilde{\boldsymbol{s}} \leftarrow WW^+\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}), \tag{6.20}$$

$$\boldsymbol{\phi} \leftarrow \angle\tilde{\boldsymbol{s}}, \tag{6.21}$$

respectively, where $\angle$ denotes an operator that gives the arguments of the components of a complex vector as a real vector in $[-\pi, \pi)^{LM}$.

Eq. (6.20) means applying the inverse CWT followed by the CWT to $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$. When $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ is already a complex vector corresponding to a consistent spectrogram, this update simply becomes $\tilde{\boldsymbol{s}} \leftarrow \boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$. Eq. (6.21) means replacing the phase estimate $\boldsymbol{\phi}$ with the phase of $\tilde{\boldsymbol{s}}$. The iterative algorithm is thus equivalent to Irino's algorithm.

Any phase estimated with the algorithm is accurate to up to an overall phase constant. This is because the CWT is a linear transform and the difference in an overall phase constant does not change the magnitude CWT spectrogram [91] (see [94] for more fundamental results).

The derivation of the algorithm depends only on the linearity of the CWT and the redundancy of the wavelet basis waveforms. Therefore, if the CWT and the inverse CWT are

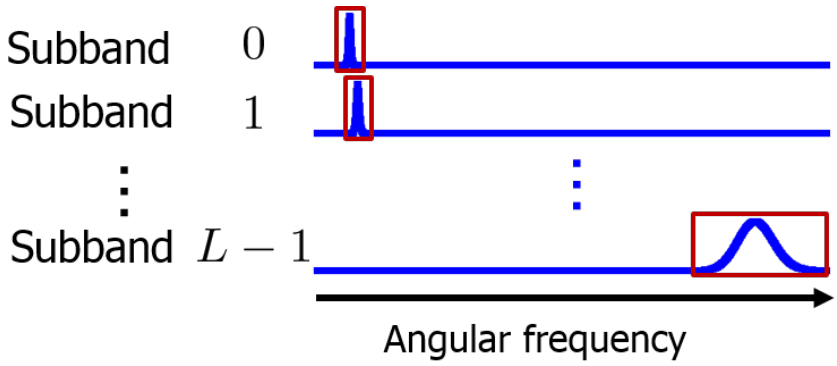


Figure 6.4: Examples of the frequency responses of different subband filters. The analyzing wavelet is the log-normal wavelet [1].

replaced with any linear and redundant time-frequency transform and its inverse transform, the iterative procedure is guaranteed to converge. In fact, when W is replaced with a matrix in which each row is a basis waveform of the STFT, the proposed algorithm becomes equivalent to the phase estimation algorithm for a magnitude STFT spectrogram proposed in [59].

6.5 Fast Phase Estimation Algorithm

6.5.1 Fast Approximate Continuous Wavelet Transform

The CWT and the inverse CWT are computationally expensive compared to the STFT and the inverse STFT. Here we briefly describe the fast approximate method (FACWT) for computing the CWT proposed in [61]. It uses the fact that the dominant part of the frequency response of each subband filter is concentrated around its center frequency (as shown in Fig. 6.4), as is common in many types of analyzing wavelets including the Morlet and log-normal wavelets [1]. The FACWT is equivalent to the method proposed in [64] if all frequency responses $\{\hat{\psi}_{l,k}\}_{l,k}$ have finite supports.

According to the filter bank interpretation of the CWT, the CWT of an input signal, $\boldsymbol{s}_l = [s_{l,0}, \ldots, s_{l,M-1}]^\top = W_l \boldsymbol{f}$, can be computed by multiplying the DFT of the entire signal, i.e, $\hat{\boldsymbol{f}} = [\hat{f}_0, \ldots, \hat{f}_{M-1}]^\top = F_M \boldsymbol{f}$, by the frequency response of the l-th subband, i.e., $\hat{W}_l$, and

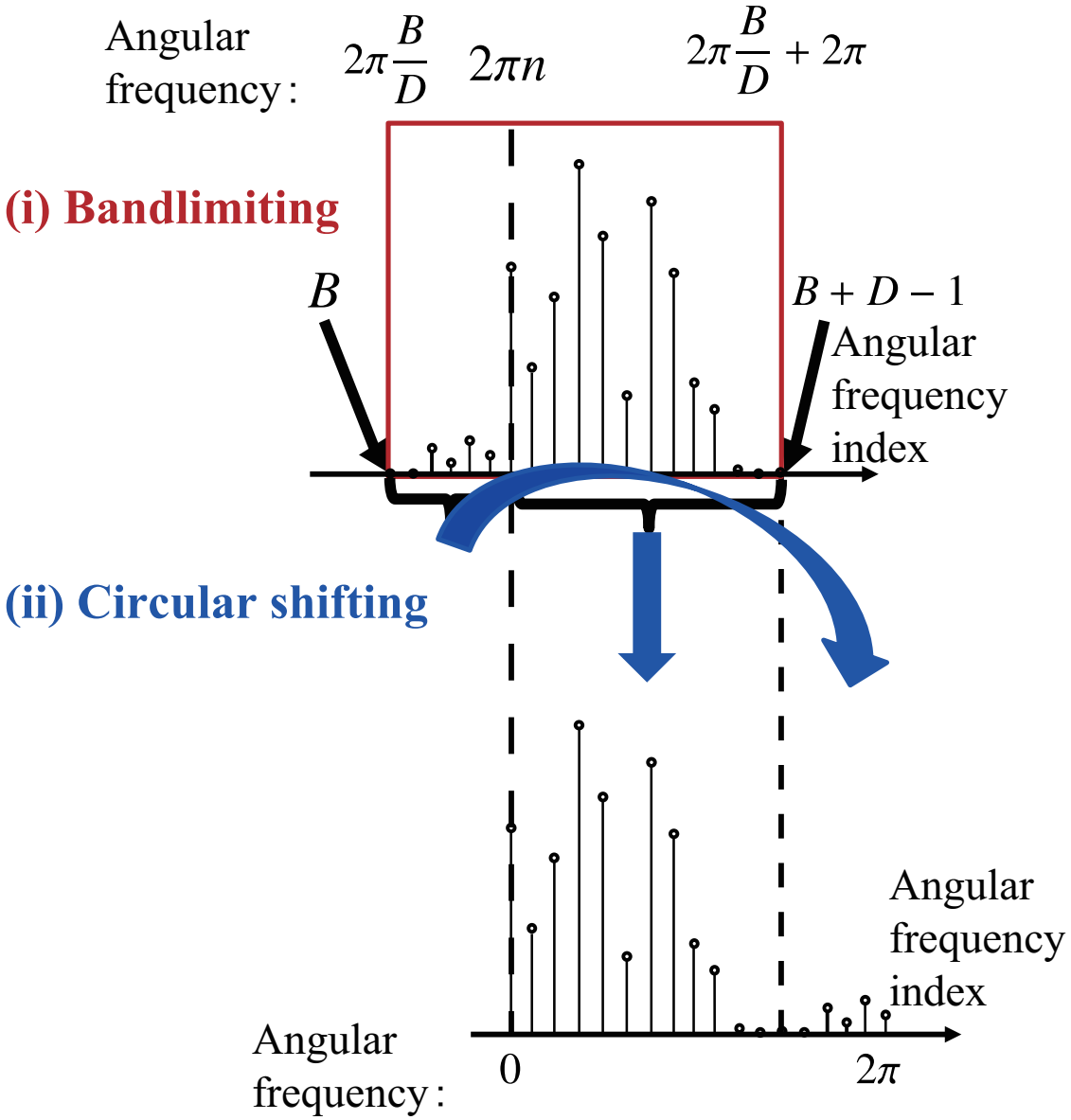


Figure 6.5: A circularly shifted version of $G_{l,B},\ldots,G_{l,B+D-1}$.

then computing the inverse DFT of $\hat{W}_l\hat{\boldsymbol{f}}$. This can be confirmed from

$$\boldsymbol{s}_l = W_l\boldsymbol{f} \tag{6.22}$$

$$= F_M^{\mathsf{H}}F_M W_l F_M^{\mathsf{H}}F_M\boldsymbol{f} \tag{6.23}$$

$$= F_M^{\mathsf{H}}\hat{W}_l\hat{\boldsymbol{f}}. \tag{6.24}$$

Note that the second equality follows from the fact that the DFT matrix F_M is a unitary matrix, i.e., $F_M^{\mathsf{H}}F_M = \boldsymbol{I}_M$. Here, if we can assume that the elements of $\{\hat{\psi}_{l,k}\}_k$ are dominant within and near zero outside the range $k = B, B+1, \cdots, B+D-1$ $(0 \le B, 0 < D \le M)$, we can approximate $\boldsymbol{s}_l$ reasonably well by using the elements of $\{\hat{\psi}_{l,k}\hat{f}_k\}_k$ only within that range and neglecting the remaining elements. This implies the possibility of computing an

approximation of $\boldsymbol{s}_l$ with a lower computational cost.

For simplicity of notation, let us put $G_{l,k} = \hat{\psi}_{l,k}\hat{f}_k$. We are concerned with computing an approximation of the full-band inverse DFT of $G_{l,k}$:

$$s_{l,m} = \sum_{k=0}^{M-1} G_{l,k} e^{j\frac{2\pi km}{M}}. \tag{6.25}$$

As mentioned above, $G_{l,0}, \ldots, G_{l,M-1}$ can be approximately viewed as a band-limited spectrum. In general, the inverse DFT of a band-limited spectrum can be computed by taking the inverse DFT over the finite support. In the time domain, this process corresponds to downsampling the signal given by the "full-band" inverse DFT. The proposed method uses this idea to approximate the inverse DFT of the full-band spectrum $G_{l,0}, \ldots, G_{l,M-1}$. Now, if we choose D such that M/D becomes an integer, we can approximate the downsampled version of $s_{l,m}$ by

$$y_{l,d} = \sum_{k=B}^{B+D-1} G_{l,k} e^{j\frac{2\pi kd}{D}} \tag{6.26}$$

$$= \sum_{k=B}^{B+D-1} G_{l,k} e^{j\frac{2\pi k(M/D)d}{M}}. \tag{6.27}$$

By comparing (6.25) and (6.27), we can confirm that

$$s_{l,(M/D)d} \simeq y_{l,d} \quad (d = 0, 1, \cdots, D-1), \tag{6.28}$$

if we assume $G_{l,k} \simeq 0$ outside the range $k = B, B+1, \cdots, B+D-1$. Since $y_{l,d}$ can be rewritten as

$$y_{l,d} = \sum_{k=0}^{D-1} G_{l,k+B} e^{j\left(\frac{2\pi k}{D} + 2\pi\frac{B}{D}\right)d} \tag{6.29}$$

$$= e^{j2\pi\frac{B}{D}d} \sum_{k=0}^{D-1} G_{l,k+B} e^{j\frac{2\pi kd}{D}}, \tag{6.30}$$

we notice that $y_{l,d}$ can be computed by multiplying the inverse DFT of $G_{l,B}, \ldots, G_{l,B+D-1}$ by $e^{j2\pi\frac{B}{D}d}$. Note that this is equivalent to computing the inverse DFT of a circularly shifted version of $G_{l,k}$ (see Fig. 6.5):

$$\tilde{G}_{l,k} = \begin{cases} G_{l,k+nD} & (k = 0, \ldots, B-(n-1)D-1) \\ G_{l,k+(n-1)D} & (k = B-(n-1)D, \ldots, D-1) \end{cases}, \tag{6.31}$$

Algorithm 1 Fast approximate continuous wavelet transform: $\{\boldsymbol{y}_l\}_l = \mathbf{FACWT}(\boldsymbol{f})$

Initialize $\hat{\psi}^*_{l,k}, B_l, D_l, n_l$ for $k = 1, 2, \cdots, M-1$ and $l = 0, 1, \cdots, L-1$.
$\hat{\boldsymbol{f}} \leftarrow \mathbf{FFT}_M(\boldsymbol{f})$
for $l = 0$ to $L-1$ **do**
  $B \leftarrow B_l, D \leftarrow D_l, n \leftarrow n_l,\ \hat{\boldsymbol{y}}_l \leftarrow \mathbf{0}_D$
  for $d = 0$ to $B-(n-1)D-1$ **do**
    $\hat{y}_{l,d} \leftarrow \hat{f}_{d+nD} \times \hat{\psi}^*_{l,d+nD}$
  end for
  for $d = B-(n-1)D$ to $D-1$ **do**
    $\hat{y}_{l,d} \leftarrow \hat{f}_{d+(n-1)D} \times \hat{\psi}^*_{l,d+(n-1)D}$
  end for
  $\boldsymbol{y}_l \leftarrow \mathbf{iFFT}_D(\hat{\boldsymbol{y}}_l)$
end for
return $\{\boldsymbol{y}_l\}_l$

where n is an integer such that

$$n - 1 < \frac{B}{D} \leq n. \tag{6.32}$$

We consider invoking the fast Fourier transform (FFT) for computing the inverse DFT and so we assume the size D to be a power of 2. Since $D < M$, the computational cost for computing $\boldsymbol{y}_l = [y_{l,0}, \ldots, y_{l,D-1}]^\top$ is obviously lower than that for computing $\boldsymbol{s}_l = [s_{l,0}, \ldots, s_{l,M-1}]^\top$. B and D are allowed to differ between subband filters, and we hereafter add the subscript l to B, D and n, i.e. B_l, D_l and n_l. The pseudo code of the FACWT is summarized in Algorithm 1, where (**i**)$\mathbf{FFT}_M$ denotes the M-point FFT (inverse FFT, respectively).

The processes of bandlimiting and circular shifting for the l-th subband can be represented by a matrix K_l:

$$K_l := \begin{bmatrix} 0_{(n_l D_l - B_l)\times\{B_l-(n_l-1)D_l\}} & I_{n_l D_l - B_l} \\ I_{B_l-(n_l-1)D_l} & 0_{\{B_l-(n_l-1)D_l\}\times(n_l D_l - B_l)} \end{bmatrix} \times \begin{bmatrix} 0_{D_l \times B_l} & I_{D_l} & 0_{D_l\times(M-D_l-B_l)} \end{bmatrix} \tag{6.33}$$

where I_D and $0_{D\times B}$ are the $D\times D$ identity matrix and the $D\times B$ zero matrix. The first matrix of the right-hand side of Eq. (6.33) represents the circular shift and the second matrix the

Algorithm 2 Inverse fast approximate continuous wavelet transform: $\boldsymbol{f} = \mathbf{iFACWT}(\{\boldsymbol{y}_l\}_l)$

1: **Initialize** $\hat{\psi}_{l,k}, B_l, D_l, n_l, C_k$ for $k = 1, 2, \cdots, M-1$ and $l = 0, 1, \cdots, L-1$
2: $\hat{\boldsymbol{f}} \leftarrow \mathbf{0}_M$
3: **for** $l = 0$ to $L-1$ **do**
4: $\quad D \leftarrow D_l,\ B \leftarrow B_l,\ n \leftarrow n_l,\ \hat{\boldsymbol{y}}_l \leftarrow \mathbf{FFT}_D(\boldsymbol{y}_l)$
5: $\quad$ **for** $k = B$ to $nD-1$ **do**
6: $\quad\quad \hat{f}_k \leftarrow \hat{f}_k + \hat{y}_{l,k-(n-1)D} \times (\hat{\psi}_{l,k}/C_k)$
7: $\quad$ **end for**
8: $\quad$ **for** $k = nD$ to $B+D-1$ **do**
9: $\quad\quad \hat{f}_k \leftarrow \hat{f}_k + \hat{y}_{l,k-nD} \times (\hat{\psi}_{l,k}/C_k)$
10: $\quad$ **end for**
11: **end for**
12: $\boldsymbol{f} \leftarrow \mathbf{iFFT}_M(\hat{\boldsymbol{f}})$
13: **return** $\boldsymbol{f}$

bandlimiting. The downsampled version of $\boldsymbol{s}_l$ obtained with the FACWT can be described as

$$\boldsymbol{y}_l = F_{D_l}^{\mathsf{H}} K_l \hat{W}_l F_M \boldsymbol{f}. \tag{6.34}$$

Similarly to the inverse CWT, the fast approximate version of the inverse CWT can be defined by the pseudo-inverse matrix of a $(\sum_l D_l) \times M$ matrix defined by vertically concatenating $\{F_{D_l}^{\mathsf{H}} K_l \hat{W}_l F_M\}_{l=0}^{L-1}$. The pseudo code of the inverse FACWT is summarized in Algorithm 2, where $\{\boldsymbol{y}\}_l$ denotes a CWT spectrogram obtained with the FACWT.

6.5.2 Fast Phase Estimation Algorithm

Now we consider the phase estimation algorithm in which the CWT and inverse CWT steps are replaced with the FACWT and the inverse FACWT. Both are linear and redundant transforms, and so the convergence of the algorithm is guaranteed as mentioned in Sec. 6.4.2. We call the algorithm the iterative FACWT algorithm. Its pseudo code is summarized in Algorithm 3, where we redefine a given magnitude CWT spectrogram associated with the l-th subband as $\boldsymbol{a}_l \in [0, \infty)^{D_l}$ and phase estimate as $\boldsymbol{\phi}_l$ for each subband filter, and $\{\boldsymbol{y}_l(\boldsymbol{a}_l, \boldsymbol{\phi}_l)\}_l$ denotes an estimated CWT spectrogram with magnitude $\{\boldsymbol{a}_l\}_l$ and phase $\{\boldsymbol{\phi}_l\}_l$.

Furthermore, the M-point inverse FFT in the inverse FACWT, (line 12 in Algorithm 2),

Algorithm 3 Iterative fast approximate continuous wavelet transform algorithm: $\{\boldsymbol{\phi}_l\}_l =$ **IterFACWT**$(\{\boldsymbol{a}_l\}_l)$

1: **Initialize** $\boldsymbol{\phi}_l$ for $l = 0, 1, \cdots, L-1$
2: **repeat**
3: $\quad \boldsymbol{f} \leftarrow$ **iFACWT**$(\{\boldsymbol{y}_l(\boldsymbol{a}_l, \boldsymbol{\phi}_l)\}_l)$
4: $\quad \{\tilde{\boldsymbol{y}}_l\}_l \leftarrow$ **FACWT**$(\boldsymbol{f})$
5: $\quad$ **for** $l = 0$ to $L-1$ **do**
6: $\quad\quad \boldsymbol{\phi}_l \leftarrow \angle\tilde{\boldsymbol{y}}_l$
7: $\quad$ **end for**
8: **until** a convergence criterion is satisfied
9: **return** $\{\boldsymbol{\phi}_l\}_l$

can be cancelled by the M-point FFT in the FACWT (line 2 in Algorithm 1) in the iterative FACWT algorithm. We call the algorithm the refined iterative FACWT algorithm.

6.5.3 Time and Space Complexity

The computational costs for the CWT and the FACWT mainly depend on the number of the points for the inverse DFT. Since the full band inverse DFT is of $O(M \log_2 M)$, the total complexity of the CWT is $O((L+1)M \log_2 M)$. By contrast, the band-limited DFT is of $O(D_l \log_2 D_l)$ and so the total complexity of the FACWT is $O(M \log_2 M + \sum_l D_l \log_2 D_l)$. Consequently, Irino's algorithm is of $O(2(L+1)M \log_2 M)$ per iteration while the iterative FACWT algorithm (the refined iteration FACWT algorithm) is of $O(2M \log_2 M + 2\sum_{l=0}^{L-1} D_l \log_2 D_l)$ ($O(2\sum_{l=0}^{L-1} D_l \log_2 D_l)$, respectively).

The space complexity of the proposed algorithms are small compared to Irino's algorithm [60]. When the signal length M is long enough, the space complexity depends primarily on the size of the CWT spectrogram. While the size of the CWT spectrogram of Irino's algorithm is LM, that of each proposed algorithm is only $\sum_l D_l$.

6.6 Experimental Evaluations

6.6.1 Processing Time

We measured processing times to evaluate the reduction of the computational complexity with the proposed algorithms. The processing time depends on signal length and not on signal content. We artificially synthesized random signals with the length of 2^x ($x = 10, 11, \cdots, 21$) samples at 16 kHz sampling rate, and used their magnitude CWT spectrograms as inputs. The log-normal wavelet [1] was used as an analyzing wavelet. Its Fourier transform is given by

$$\hat{\psi}(\omega) := \begin{cases} \exp\left(-\dfrac{(\ln \omega)^2}{4\sigma^2}\right) & (\omega > 0) \\ 0 & (\omega \leq 0) \end{cases} \tag{6.35}$$

where ω denotes an angular frequency. σ represents a standard deviation in the log-frequency domain and we put $\sigma = 0.02$ in the following. The scale parameters a_l were set such that the center frequencies ranged 27.5 from 7040 Hz with a 1/10 semitone interval. In the bandlimiting process of the proposed algorithms, we computed the elements within the range $[-3\sigma, 3\sigma]$ around the center frequency of each subband filter in the log-frequency domain. The algorithms were implemented in C++ on a PC with 3.50 GHz CPU (Intel(R) Core(TM) i7-3770K Processor) and 32 GB memory running Debian.

Processing times averaged over 50 iterations are shown in Fig. 6.6. It can be confirmed that the proposed algorithms outperformed Irino's algorithm in terms of processing times. For example, to process a signal of around 16 seconds length, Irino's algorithm took 18 seconds per iteration on average while the iterative FACWT algorithm took 0.24 second per iteration (75 times faster than Irino's algorithm). The refined iterative FACWT was faster than the iterative FACWT by around five percentage points on average. In addition, the proposed algorithms processed all signals, although Irino's algorithm could not work with the signals of over around 20 seconds length due to lack of memory. These results show that the proposed algorithms are efficient in both time and memory, which are consistent with the theoretical result described in Sec. 6.5.3.

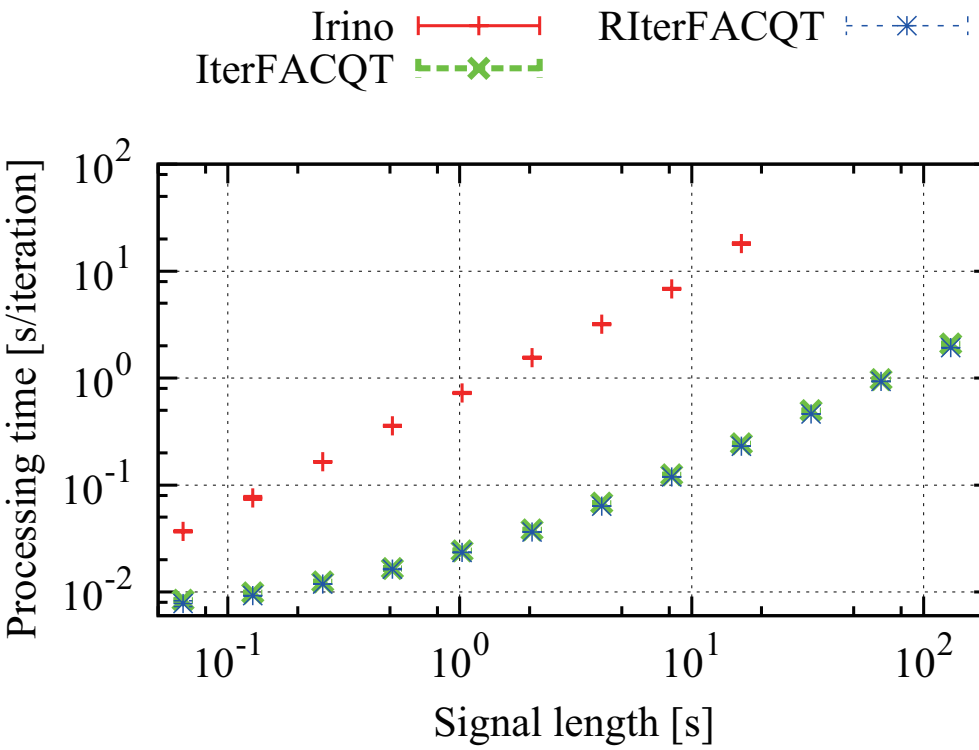


Figure 6.6: Average processing times per iteration and standard errors with respect to signal length. "Irino" denotes Irino's algorithm [60] and "IterFACWT" and "RIterFACWT" represent the iterative FACWT algorithm (Algorithm 3) and its refined version, respectively.

6.6.2 Audio Quality and Approximation Property

Experimental Conditions

The lower the approximation accuracy is, the faster the proposed algorithms becomes, but the poorer audio quality the resulting reconstructed signal has. To examine the trade off, we measured processing times and audio quality of signals reconstructed with the refined iterative FACWT algorithm. The proposed algorithms are same in approximation accuracy and the results obtained with the iterative FACWT algorithm were omitted. As input data, we used magnitude CWT spectrograms of music and speech audio signals. The music data consisted of 102 music audio signals in the RWC music genre database [2], and the speech data consisted of 485 speech signals (242 male and 243 female speeches) in the ATR Japanese speech database [95]. The audio signals were downsampled to 16 kHz and their durations were arranged to five seconds by cutting out a part of each audio. The analyzing wavelet was same as in Sec. 6.6.1, and the central frequencies ranged from 50 to 7040 Hz with an interval of 1/5 semitone. In the bandlimiting process of the refined iterative FACWT

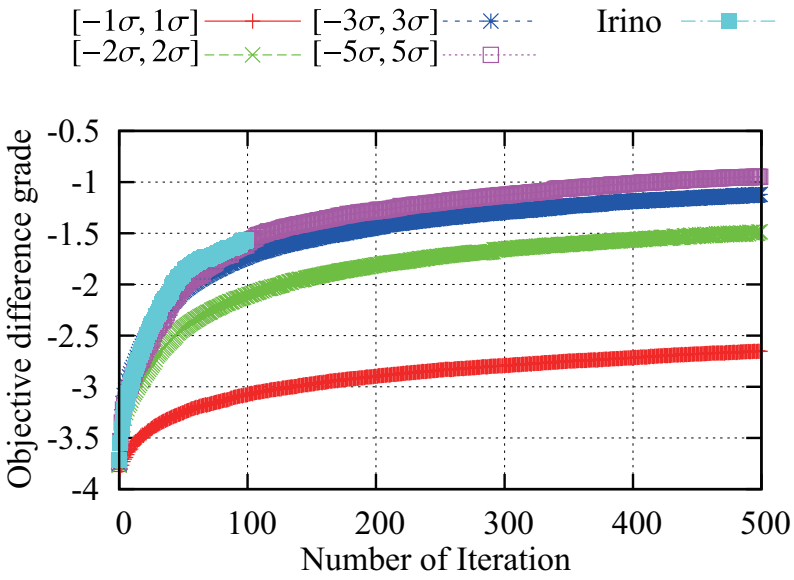
algorithm, we computed the elements within the range $[-P\sigma, P\sigma]$ ($P = 1, 2, 3, 5$) around the central frequencies in the log-frequency domain. The smaller P is, the approximation accuracy becomes lower. The proposed algorithm with enough large P is equivalent to Irino's algorithm, and we also compared the proposed algorithms with Irino's algorithm. The proposed algorithm and Irino's algorithm were started with randomly initialized phase and stopped after 500 iterations and 100 iterations, respectively. The algorithms ran on a computer with 3.30 GHz CPU (Intel (R) Core(IM) i3-2120 Processor) and 8 GB memory running Debian.

Results

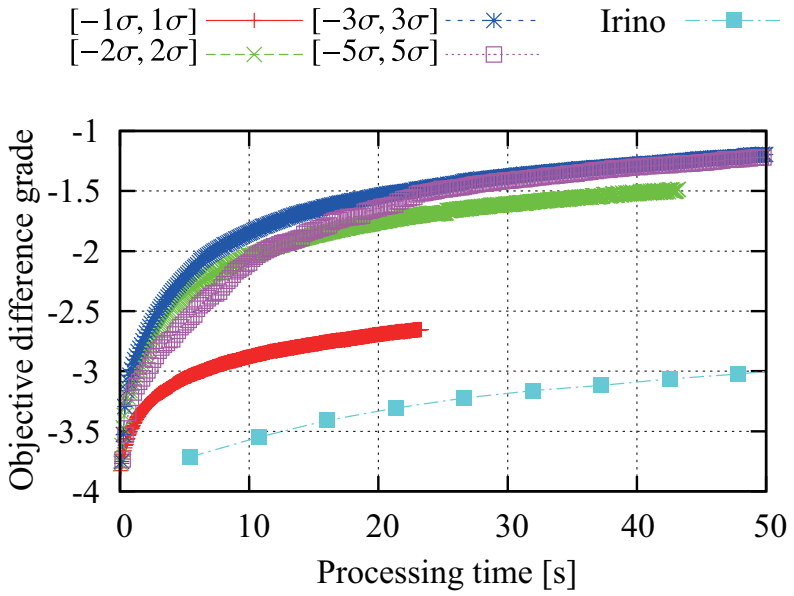
The audio quality of a reconstructed music signal was measured by the method of perceptual evaluation of audio quality (PEAQ) [96] in AFsp [97]. The PEAQ method provides an objective difference grade (ODG) between an original audio signal and a reconstructed signal, whose range is -4 to 0. The larger the ODG is, the higher the audio quality of a reconstructed signal becomes.

Fig. 6.7 displays average ODGs with respect to the number of iterations (left) and the processing times only required by the phase reconstruction part (right) on the music data. The processing times were averaged over all audio signals at each iteration. The proposed algorithms with $P = 3$ and 5 and Irino's algorithm provided larger ODGs than -2.0 on average after 100 iterations. These results show that the reconstructed signals obtained with the algorithms had almost the same audio quality as the original sounds. We notice again that the proposed algorithms with these Ps were much faster than Irino's algorithm. It can be seen that the proposed algorithm with $P = 3$ outperformed that with $P = 5$ in average ODG in the first 40 seconds of the processing time. This is not surprising since using small P increases the number of iterations that the algorithm can perform within the same processing time.

Results on the speech data showed similar trends. The audio quality of a reconstructed speech signal was measured by the method of the perceptual evaluation of speech quality (PESQ) [98]. The PESQ value ranges from -0.5 to 4.5. The higher PESQ value is, the higher the speech quality of a reconstructed signal becomes. Fig. 6.8 illustrates average PESQ values similarly in Fig. 6.7. Similarly to the results on the music data, the proposed algorithms with $P = 3$ and 5 reconstructed audio signals having almost the same speech quality as the original sounds and were much faster than Irino's algorithm. We finally concluded that using



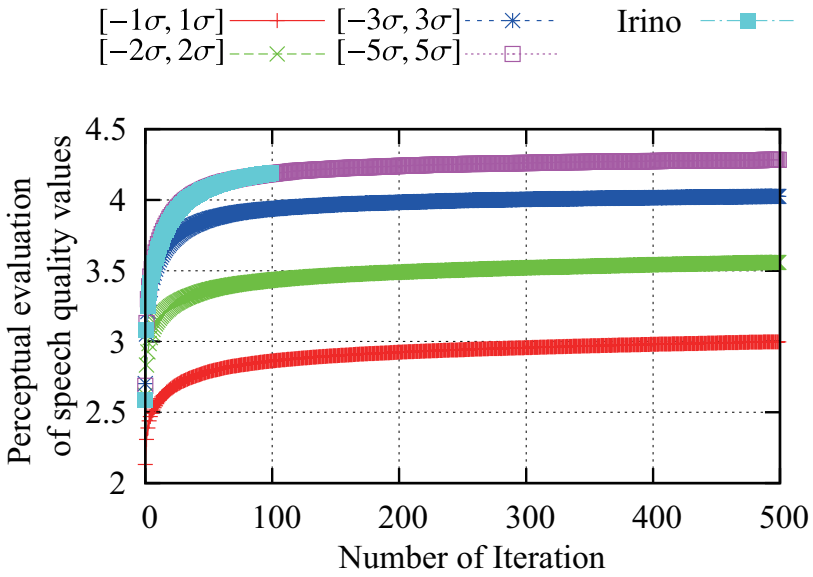
(a) Results for the number of iterations.



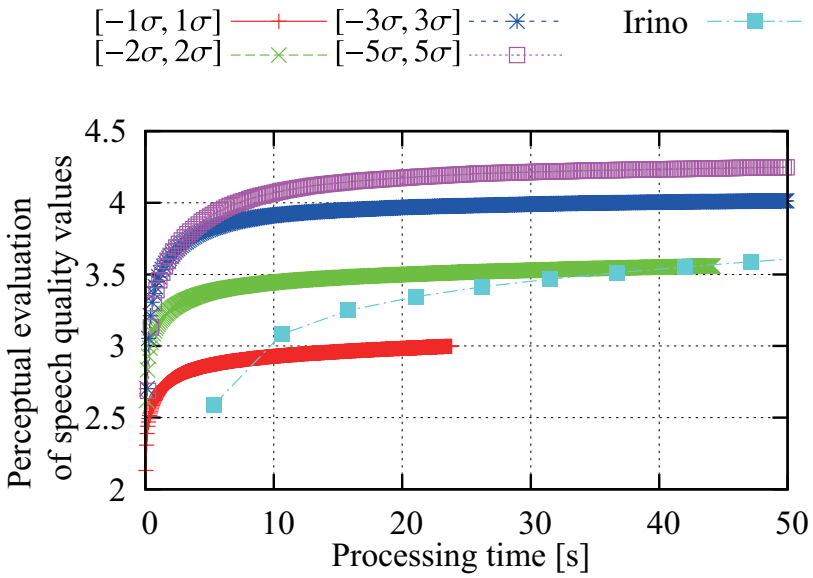
(b) Results for the processing time.

Figure 6.7: Evolution of average ODGs by PEAQ for the number of iterations and the processing time on the music data. "$[-P\sigma, P\sigma]$ $(P = 1, 2, 3, 5)$" denotes the refined iterative fast approximate CWT with varying approximations, and "Irino" represents Irino's algorithm [60].

around $P = 3$ is practical in terms of processing time and audio quality of reconstructed signals.



(a) Results for the number of iterations.



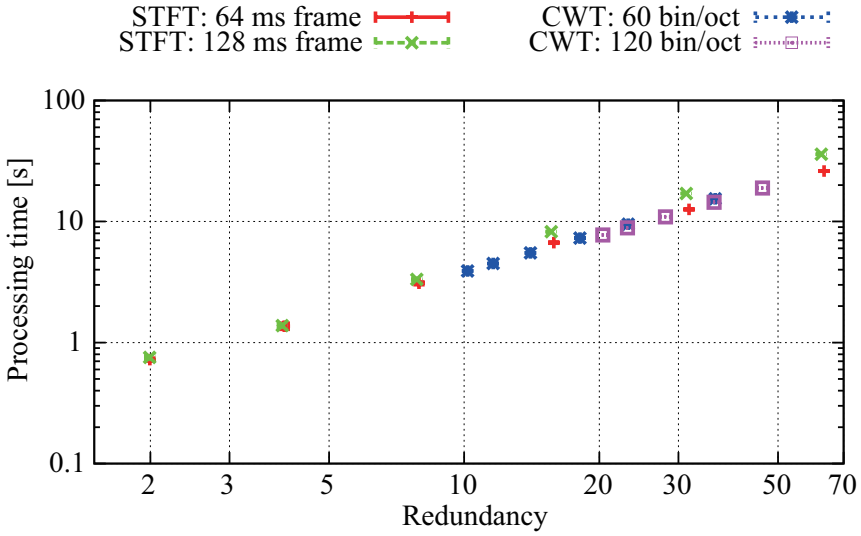
(b) Results for the processing time.

Figure 6.8: Evolution of average PESQ values for the number of iterations and the processing time on the speech data. The algorithms are same as in Fig. 6.7.

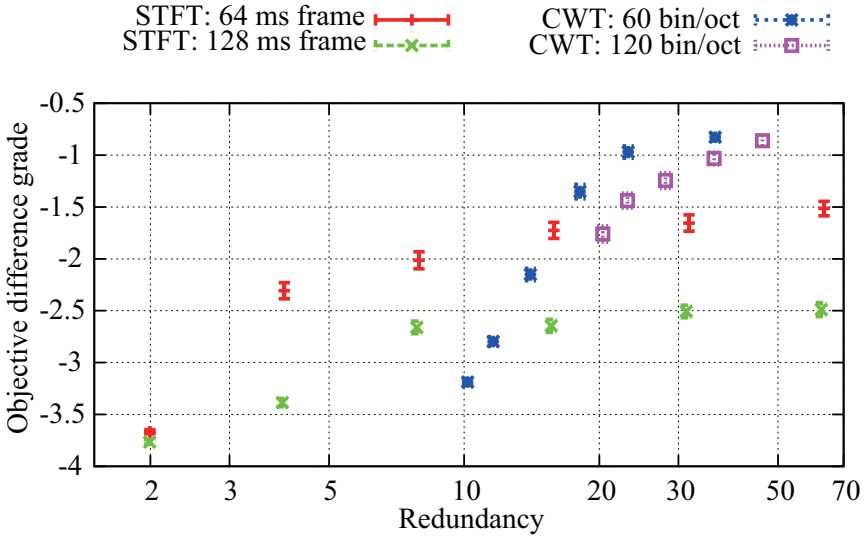
6.6.3 Comparison to Signal Reconstruction from Magnitude STFT spectrograms

Now, let us examine the effect of the difference of time-frequency transforms on signal reconstruction from magnitude spectrograms. To compare the redundancy of spectrograms having different frequency resolution, we defined a redundancy measure by the ratio of the

number of the number of the spectrograms to the number of signal elements, for example, $\sum_l D_l/T$ for a FACWT spectrogram. STFT spectrograms were computed with Gaussian windows and a wide variety of hopsizes. The parameter σ of the log-normal wavelet corresponds to the frame length and thus CWT spectrograms were computed with a wide range of σ and $P = 3$. We used the Griffin's algorithm [57] as the phase estimation algorithm for STFT and the refined iterative FACWT algorithm for CWT. The audio signals were the music data used in Sec. 6.6.2. The other conditions were the same as in Sec. 6.6.1.



(a) Average processing times with standard errors.



(b) Average ODGs with standard errors by PEAQ.

Figure 6.9: Comparison of signal reconstruction from magnitude STFT and CWT spectrograms in processing time and ODG by PEAQ for the redundancy measure. "STFT: 64 ms frame" and "CWT: 120 bin/oct" represent the STFT with a frame length of 64 and the CWT with 120 bins per octave, respectively.

The results after 500 iterations are shown in Fig. 6.9. The results of "STFT: 64 ms frame" ("STFT:128 ms frame") were obtained with $32, 16, 8, 4, 2$ and 1 ms ($64, 32, 16, 8, 4$ and 2 ms, respectively) in a left-to-right fashion of the redundancy measure. The results of "CWT: 60 bin/oct" and "120 bin/oct" were obtained with $\ln(2)/84, \ln(2)/72, \ln(2)/60, \ln(2)/48, \ln(2)/36$ and $\ln(2)/24$ in a left-to-right fashion of the redundancy measure. There were not significant differences in processing time between CWT and STFT, which shows the efficiency of the present algorithms. While the ODGs were mainly dependent on the redundancy, the ODGs of CWT increased more quickly with increase of the redundancy compared to those of STFT. This indicates that the CWT is more valid in the audio quality of reconstructed signals than the STFT when CWT provides spectrograms having enough redundancy.

6.6.4 Demonstration of Phase Estimation

We here demonstrate pitch transposition of an audio signal in the magnitude CWT spectrogram domain, using the refined iterative FACWT algorithm. When the center frequencies of the subbands are located uniformly in the log-frequency domain and $D_0 = D_1 = \cdots = D_{L-1}$ in the proposed algorithm, we simply shift the components of the CWT spectrograms to the lower or higher analysis frequency components, and the blank components by the move are filled by zero. However, the shifts cause the mismatches of phases, and the use of the original and zero phases leads to failure of the pitch transposition, hence we need to use the phase estimation for synthesizing the pitch-transposed audio signals. Although the shifts also result in the mismatches of magnitude CWT spectrograms, the proposed algorithm can be used as abovementioned. We obtained the synthesized signals with it as we expected, and they are available at `http://tomohikonakamura.github.io/Tomohiko-Nakamura/demo/fastCWT.html`.

6.7 Real-Time Extension of Fast Phase Estimation Algorithm

The algorithms presented in Sec. 6.5.2 rely on the FFT of an entire signal, and the space complexity of the algorithms increase with the signal length. This characteristics cause that the algorithms require large memory to compute music audio signals of practical length and in a situation where memories of computers and digital devices are limited, such as iPod,

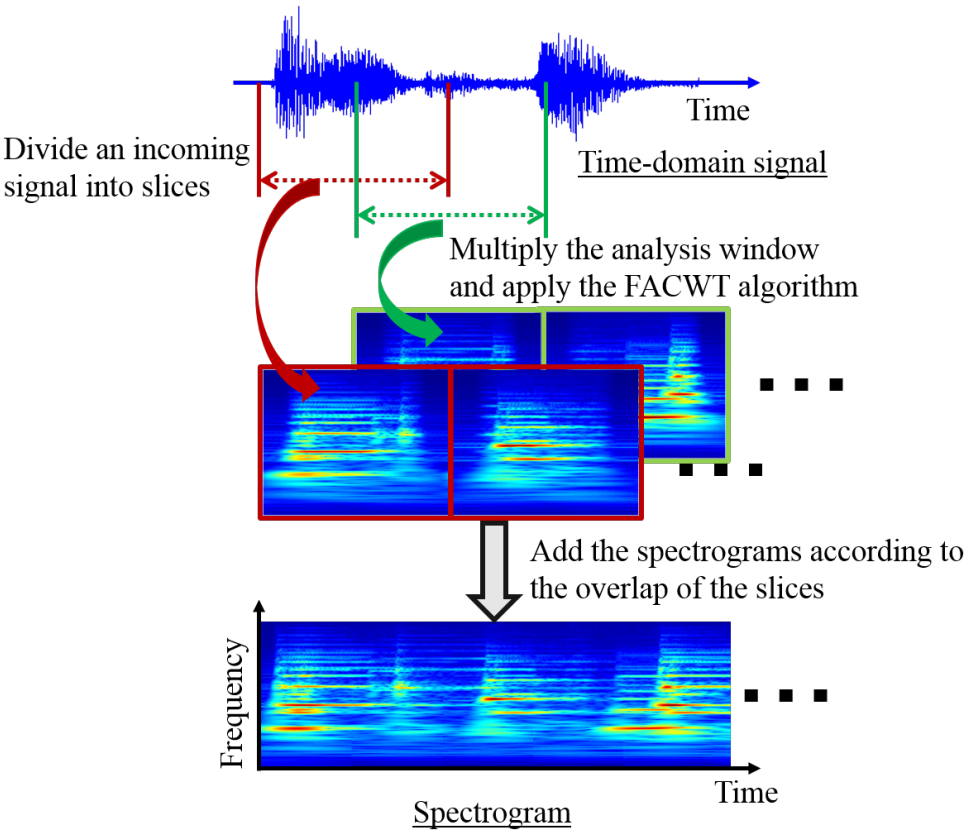


Figure 6.10: Schematic illustration of the online extension FACWT.

music audio signals that can be processed are limited. Furthermore, the algorithms, in principle, are not applicable to music applications that work in real time. To increase the applicability of the present algorithms, we now extend the iterative FACWT algorithm to work in real time. For the convenience of the notation we assume $D := D_0 = \cdots = D_{L-1}$, but the following discussion is valid for general D_l.

6.7.1 Online FACWT Algorithm

A real-time algorithm of calculating CWT have already been proposed in [63], which divides an incoming signal into segments overlapping with each other called slices and processes the signal of each slice. In the same way of the real-time algorithm, we can extend the FACWT algorithm to work in real time, which we call the online FACWT algorithm.

The online FACWT algorithm consists of three steps (Fig. 6.10):

(i): An incoming signal is divided into slices of $2N$ length with a N hopsize.

(ii): The FACWT algorithm is applied to the signal of each slice multiplied by an analysis

window $\boldsymbol{h} = [h_0, \cdots, h_{2N-1}]^\top$ to the windowed signal of each slice.

(iii): The FACWT spectrograms of the slices are added with each other according to the overlap of the slices.

To reconstruct the signal of each slice overlapping with the previous slice, we require only the spectrograms of the current slice and the previous slice obtained in Step (ii). The spectrograms of the two slices are transformed into signals by the inverse FACWT algorithm, the signals are multiplied by a synthesis window, arranged as $\boldsymbol{v} = [v_0, \cdots, v_{2N-1}]^\top$, and the windowed signals are added with each other according to the overlap of the slices. Here the synthesis window satisfies $h_n v_n + h_{n+N} v_{n+N} = 1$ for $n = 0, \cdots, N-1$. Since the slice-wise processing does not require the entire input signal to obtain the spectrograms of individual slices and the number of the spectrogram elements of a slice is independent of T, where T is the length of the input signal, the space complexity of the online FACWT algorithm is much lower than that of the FACWT algorithm when $N \ll T$.

For the convenience of the implementation, we use the following Tukey window defined in [63] as the analysis window:

$$h_n = \begin{cases} 0 & (0 \le n < \frac{N-M}{2}), \\ 0.5 - 0.5\cos\left(\pi \dfrac{n - \frac{N-M}{2}}{M-1}\right), & (\frac{N-M}{2} \le n < \frac{N+M}{2}), \\ 1 & (\frac{N+M}{2} \le n < \frac{3N-M}{2}), \\ 0.5 + 0.5\cos\left(\pi \dfrac{n - \frac{3N-M}{2}}{M-1}\right), & (\frac{3N-M}{2} \le n < \frac{3N+M}{2}), \\ 0 & (\frac{3N+M}{2} \le n < 2N), \end{cases} \tag{6.36}$$

where M $0 < M < N$ is an parameter of controlling the overlap of the consecutive slices. The larger M becomes, the more overlapping elements the consecutive slices have. The window satisfies $h_n + h_{n+N} = 1$ for $n = 0, \cdots, N-1$ and thus we can set $v_n = 1$ for all n.

6.7.2 Real-Time Iterative FACWT Algorithm

The online FACWT algorithm and its inverse transform are linear and redundant and one may think that the algorithm where the FACWT and inverse FACWT are replaced with their online versions in the iterative FACWT algorithm work well. We call the algorithm *the real-time baseline algorithm*. However, the algorithm fails to reconstruct signals from magnitude

spectrograms as we will show later in Sec. 6.7.3. This is because any phase estimated with the iterative FACWT algorithm is accurate to up to an overall phase constant and the estimated phases of consecutive slices are often inconsistent with each other. To suppress the inconsistency between the consecutive slices, we present a real-time iterative algorithm that takes into account signal components obtained in a previous slice, which we call the real-time iterative FACWT algorithm.

Let us define the signal components obtained in the previous slice by $\boldsymbol{g} \in \mathbb{C}^N$, where signal elements at the time outside the current slice are set as zero. We here redefine $\boldsymbol{a}, \boldsymbol{\phi}$ and $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ as a given magnitude spectrogram, a phase estimate and a spectrogram estimate of the current slice. If the inverse FACWT of $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ is consistent with $\boldsymbol{g}$, $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ satisfies

$$W^+\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) = \mathrm{diag}(\boldsymbol{h})\mathrm{diag}(\boldsymbol{v})(\boldsymbol{g} + W^+\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})) \tag{6.37}$$

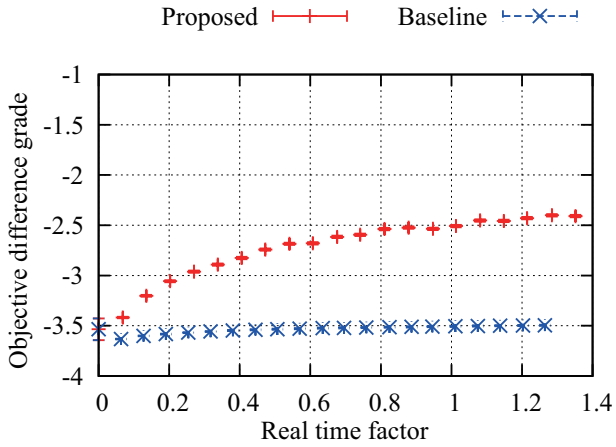
where where $\mathrm{diag}(\boldsymbol{p})$ converts a vector $\boldsymbol{p}$ into a diagonal matrix with the elements of $\boldsymbol{p}$ on the main diagonal. If we can see the right-hand side of Eq. (6.37) as an inverse transform from a complex FACWT spectrogram instead of the inverse FACWT algorithm, the algorithm that the inverse FACWT step is replaced with Eq. (6.38) in the iterative FACWT algorithm at each slice may suppress the inconsistency between the consecutive slices:

$$\tilde{s} \leftarrow W\mathrm{diag}(\boldsymbol{h})\mathrm{diag}(\boldsymbol{v})(\boldsymbol{g} + W^+\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})), \tag{6.38}$$

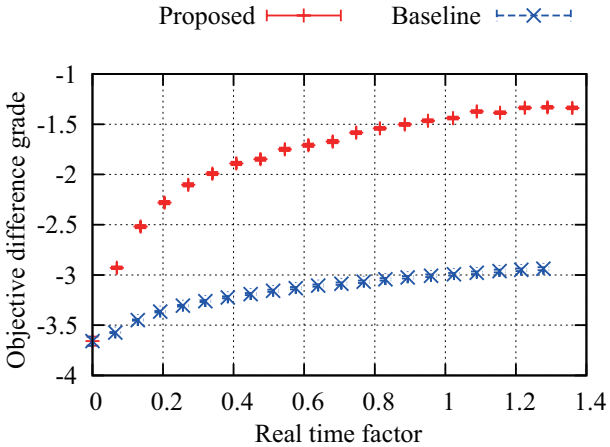
where $\tilde{s}$ is an auxiliary variable redefined for the real-time iterative FACWT algorithm. The present algorithm involves $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ and $\boldsymbol{g}$ for each slice, and thus the space complexity of the algorithm is reduced to $\mathcal{O}(N + LD)$ compared to the iterative FACWT algorithm.

6.7.3 Experiments

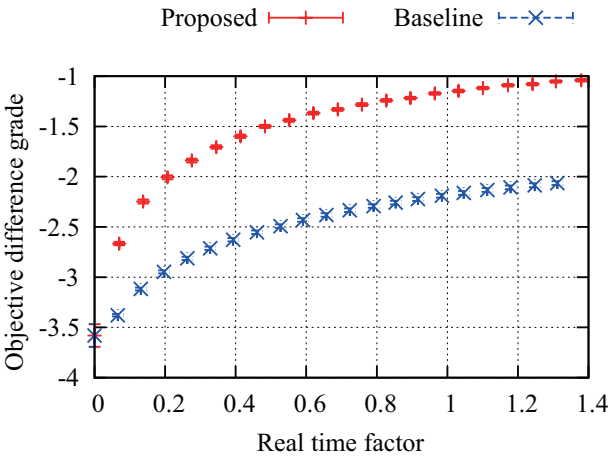
To examine the effect of the consistency between slices and the computation speed of the present algorithm, we conducted a signal reconstruction experiment. As a comparison, we use *the real-time baseline algorithm*. The test data were magnitude spectrograms of the first 30 s of 10 musical pieces in the RWC music genre database [2] and the sampling frequency was 48 kHz. The slice length was set as $170, 340$ and 680 ms ($N = 2^{12}, 2^{13}, 2^{14}$), and the analysis time window was the Tukey window defined by Eq. (6.36) with $M = N/4$. The analyzing wavelet was the log-normal wavelet with $\sigma = 0.02$. The center frequencies of the FACWT ranged from 27.5 to 23679.5 Hz with a 25 cent interval, and the elements within



(a) 170 ms slice length.



(b) 340 ms slice length.



(c) 680 ms slice length.

Figure 6.11: Average ODGs and standard errors obtained with the present algorithm and *the real-time baseline algorithm* for real time factor. Points correspond to the results finished with $0, 10, \cdots, 200$ iterations in a left-to-right fashion of real time factor.

the range $[-2\sigma, 2\sigma]$ around individual center frequencies in the log-frequency domain were computed in the bandlimiting process of the FACWT. Both algorithms were started with randomly initialized phase. The computation environment was the same in Sec. 6.6.1.

Fig. 6.11 displays the results obtained with the present algorithm and *the real-time baseline algorithm* for real-time factor, which is defined as the ratio of a processing time to a hopsize. Only the processing times required by the phase estimation part are reported here. The present algorithm provided reconstructed signals with higher ODGs than *the real-time baseline algorithm*. This shows that the consistency between slices is valid for signal reconstruction. The author listened to the reconstructed signals by *the real-time baseline algorithm*, and we found that the energy of the signals decreases quickly in the overlapping

parts of slices. This may be because the estimated phases were not coherent and the signals of the slices were cancelled with each other in the overlapping parts. While the convergence of the algorithms have yet been proved, we confirmed that the objective function decreased at each iteration.

6.8 Summary

We have proposed two fast algorithms for estimating the phase from a magnitude CWT spectrogram to construct a time domain signal. We have introduced the consistency condition for a CWT spectrogram and gave its intuitive interpretation in a viewpoint of the overlaps between the frequency responses of the neighboring subband filters. The problem of the phase estimation have been formulated as that of minimizing a numerical criterion describing how far a complex vector deviates from the consistency condition. To solve the problem, we have applied the auxiliary function approach and have derived an iterative algorithm. The derivation not only has turned out that the derived algorithm is equivalent to Irino's algorithm, but also have made it clear that the convergence of the phase estimation algorithm where the CWT and inverse CWT steps are replaced with the fast approximate versions is still guaranteed. On the basis of the proof of convergence, we have then derived two fast and convergence-guaranteed algorithms whose computational complexity and memory cost are far reduced from those of Irino's algorithm. Experimental evaluations have demonstrated that the present algorithms are around 75 times faster than Irino's algorithm and the reconstructed signals obtained with the present algorithms have almost the same audio quality as original sounds. Moreover, we extended the present algorithms to work in real time and showed the efficiency of the real-time version of the algorithms in experiments.

Chapter 7

L_p-Norm Non-Negative Matrix Factorization for Singing Voice Enhancement

7.1 Chapter Overview

Measures of sparsity are useful in many aspects of audio signal processing including speech enhancement, audio coding and singing voice enhancement. The well-known method for these applications is NMF, which decomposes a non-negative data matrix into two non-negative matrices. Although previous studies on NMF have focused on the sparsity of the two matrices, the sparsity of reconstruction errors between a data matrix and the two matrices is also important since designing the sparsity is equivalent to assuming the nature of the errors. We propose a new NMF technique, which we called L_p-norm NMF, that minimizes the L_p norm of the reconstruction errors, and derive a computationally efficient algorithm for L_p-norm NMF according to an auxiliary function principle. This algorithm can be generalized to complex-valued matrix factorizations. Since the spectrograms of singing voices can be seen as sparse matrices, we can apply L_p-norm NMF for enhancement of a singing voice in a monaural music signal. We confirmed experimentally that reasonably good enhancement results were obtained with appropriate choices of p.

7.2 Introduction

NMF [22] is a powerful technique that approximates a data matrix Y by using the product of two non-negative matrices W and H, and has been actively studied in many scientific and engineering fields in recent years (see [99]). In particular, in the field of music signal processing, successful results were obtained by regarding a magnitude spectrogram as a non-negative matrix [21].

NMF is formulated as the problem of minimizing a measure between a data matrix and a model. How to define the measure is very important since it corresponds to an assumption on statistical nature of observed data. For example, if we use the Frobenius norm, it implicitly assumes that observed data follow a normal distribution. However, when the data are contaminated with outliers, it becomes difficult to find the underlying low-rank structure of the data.

To cope with outliers, measures of sparsity (e.g. L_1 norm) have been widely used in many audio signal processing techniques such as NMF and robust principle component analysis [100, 101]. Many previous studies of NMF have used sparseness measures [28, 102], regularizers [103] and priors [34] on W and H, which induce sparse solutions. However, it is difficult to use the measures between a data matrix and a model directly since the measures are often non-linear and not differentiable and many conventional smooth optimization techniques are difficult to be directly applied.

In this chapter, we propose a new NMF (L_p-norm NMF) that minimizes the L_p norm of reconstruction errors between a data matrix and a model. We formulate L_p-norm NMF with $0 < p \leq 2$ and derive a convergence-guaranteed algorithm that consists of multiplicative update equations for W and H based on an optimization principle called the auxiliary function principle [71–73]. We further generalize this algorithm for complex-valued matrix factorizations. Since the spectrograms of singing voices can be seen as sparse matrices, we apply L_p-norm NMF for the enhancement of a singing voice in a monaural music signal and experimentally examined the effect of varying p.

We henceforth denote sets of real values and non-negative real values as $\mathbb{R}$ and $\mathbb{R}_{\geq 0}$, respectively.

7.3 L_p-Norm Non-Negative Matrix Factorization

7.3.1 Problem Setting

Let us define frequency, time, and basis indexes, respectively, as $\omega \in [0, \Omega-1]$, $t \in [0, T-1]$ and $k \in [0, K-1]$ such that $K < \Omega$ and $K < T$. Given a non-negative data matrix $Y := (Y_{\omega,t})$ L_p-norm NMF is formulated as the problem of minimizing the L_p norm

$$\mathcal{L}(H, U) := \sum_{\omega,t} \left| Y_{w,t} - \sum_{k} H_{\omega,k} U_{k,t} \right|^p \tag{7.1}$$

subject to

$$\forall k,\ \sum_{\omega} H_{\omega,k} = 1. \tag{7.2}$$

Here $H = (H_{\omega,k})$ and $U = (U_{k,t})$ are $\Omega \times K$ and $K \times T$ non-negative matrices. Eq. (7.2) is introduced to avoid an indeterminacy in the scaling. When Y is a magnitude spectrogram, the columns of H represent spectral templates and the rows of U represent the temporal activities of the spectral templates.

The constant p $(0 < p < 2)$ controls the sparsity of the difference between Y and HU. The smaller p becomes, the sparser the reconstruction errors tend to be. Note that when $p = 2$, the objective function equals the NMF with Frobenius norm. In a statistical viewpoint, this formulation assumes that the observed data follow a generalized normal distribution.

7.3.2 Efficient Algorithm Based on Auxiliary Function Approach

The objective function $\mathcal{L}(H, U)$ involves a summation over k in the L_p norm, and so many conventional smooth optimization techniques cannot be directly applied. Now, we propose a computationally stable algorithm based on an optimization principle called the auxiliary function approach [71–73]. If we can construct an upper bound of the objective function such that $\mathcal{L}(H, U) = \min_{\Theta} \mathcal{L}^+(H, U, \Theta)$, the objective function is guaranteed to be non-increasing under the updates, $H, U \to \operatorname*{argmin}_{H,U} \mathcal{L}^+(H, U, \Theta)$ and $\Theta \to \operatorname*{argmin}_{\Theta} \mathcal{L}^+(H, U, \Theta)$. We call the upper bound the auxiliary function.

To construct it, we focus on the fact that a quadratic function tangent to the power p function is an upper bound of the power p function. By writing the tangent point as $\xi_{\omega,t} \in \mathbb{R}_{\geq 0}$, the upper bound can be specifically written as

$$\left| Y_{w,t} - \sum_{k} H_{\omega,k} U_{k,t} \right|^p \leq p \xi_{\omega,t}^{p-2} \left| Y_{w,t} - \sum_{k} H_{\omega,k} U_{k,t} \right|^2 + (2-p) \xi_{\omega,t}^{p} \tag{7.3}$$

(see Lemma 2 of [26] for the proof of the inequality). The equality of the inequality (7.3) holds if and only if

$$\xi_{\omega,t} = \left| Y_{\omega,t} - \sum_k H_{\omega,t} U_{k,t} \right|. \tag{7.4}$$

Next, we focus on the fact that a quadratic function is a convex function, and so we can employ Jensen's inequality:

$$\Big(\sum_k H_{\omega,k} U_{k,t} \Big)^2 \leq \sum_k \frac{1}{\lambda_{\omega,t,k}} (H_{\omega,k} U_{k,t})^2 \tag{7.5}$$

where $\lambda_{\omega,t,k} \in \mathbb{R}_{\geq 0}$ are auxiliary variables that sum to unity, i.e. $\sum_k \lambda_{\omega,t,k} = 1$. The equality of the inequality (7.5) holds if and only if

$$\lambda_{\omega,t,k} = \frac{H_{\omega,k} U_{k,t}}{\sum_{k'} H_{\omega,k'} U_{k',t}}. \tag{7.6}$$

In summary, the upper bound of $\mathcal{L}(H, U)$ can be described as

$$\mathcal{L}^+(H, U, \Theta) = \sum_{\omega,t} p \xi_{\omega,t}^{p-2} \Big\{ Y_{w,t}^2 - Y_{\omega,t} \sum_k H_{\omega,k} U_{k,t} + \sum_k \frac{1}{\lambda_{\omega,t,k}} (H_{\omega,k} U_{k,t})^2 \Big\} + \sum_{\omega,t} (2-p) \xi_{\omega,t}^p \tag{7.7}$$

where $\Theta := (\{\lambda_{\omega,t,k}\}_{\omega,t,k}, \xi_{\omega,t}\}_{\omega,t})$. By setting the partial derivatives of $\mathcal{L}^+(H, U, \Theta)$ with respect to H and U at zeros and substituting Eqs. (7.4) and (7.6) in $\lambda_{\omega,t,k}$ and $\xi_{\omega,t}$, we obtain

$$H \leftarrow H \odot [\{(Y \oslash C) U^\top\} \oslash \{(HU \oslash C) U^\top\}] \tag{7.8}$$

$$U \leftarrow U \odot [\{H^\top (Y \oslash C)\} \oslash \{H^\top (HU \oslash C)\}] \tag{7.9}$$

where $\odot$ and $\oslash$ denote element-wise product and division, and C is an $\Omega \times T$ matrix whose (ω, t)-th element is

$$C_{\omega,t} = \left| Y_{\omega,t} - \sum_k H_{\omega,k} U_{k,t} \right|^{2-p}. \tag{7.10}$$

It is worth noting that each update equation consists of multiplication by a non-negative factor. Hence, the entries of H and U are guaranteed to be non-negative whenever their initial values are set at non-negative values.

7.4 Extension to Complex-Valued Matrix Factorizations

Although the target of the algorithm in the previous section is for non-negative matrices, it can be generalized for complex-valued matrices similarly. We call the problem L_p-norm

matrix factorization (L_p-norm MF).

L_p-norm MF is the problem of finding complex-valued matrices H and U for a given complex-valued data matrix Y such that

$$\min_{H\in\mathbb{C}^{\Omega\times K},U\in\mathbb{C}^{K\times T}} \mathcal{J}(H,U)=\sum_{\omega,t}|Y_{\omega,t}-\sum_k H_{\omega,k}U_{k,t}|^p,$$
$$\text{subject to } \forall k,\ \sum_\omega H_{\omega,k}=1$$

where $0<p<2$ and $\mathcal{J}(H,U)$ is the objective function.

Similarly in Sec. 7.3, we derive the upper bound of $\mathcal{J}(H,U)$ for applying the auxiliary function approach to this problem. The inequality used in Eq. (7.3) is applicable to $\mathcal{J}(H,U)$, and then its upper bound is the same as the right-hand side of Eq. (7.3). To derive the upper bound of the square function of Eq. (7.3), we can use a generalization of Jensen's inequality for complex values, which was employed in [26]:

$$\Big|Y_{\omega,t}-\sum_k H_{\omega,k}U_{k,t}\Big|^2\le\sum_k\frac{|\alpha_{\omega,t,k}-H_{\omega,k}U_{k,t}|^2}{\beta_{\omega,t,k}} \tag{7.11}$$

where $\alpha_{\omega,t,k}\in\mathbb{R}, \beta_{\omega,t,k}\in[0,1]$ are auxiliary variables subject to $\sum_k\alpha_{\omega,t,k}=Y_{\omega,t},\ \sum_k\beta_{\omega,t,k}=1$. The equality of the inequality (7.11) holds if and only if

$$\alpha_{\omega,t,k}=H_{\omega,k}U_{k,t}-\beta_{\omega,t,k}\Big(\sum_{k'}H_{\omega,k'}U_{k',t}-Y_{\omega,t}\Big). \tag{7.12}$$

The upper bound of $\mathcal{J}(H,U)$ can thus be given as

$$\mathcal{J}^+(H,U,\{\xi_{\omega,t}\}_{\omega,t},\{\alpha_{\omega,t,k}\}_{\omega,t,k},\{\beta_{\omega,t,k}\}_{\omega,t,k})=\sum_{\omega,t}p\xi_{\omega,t}^{p-2}\sum_k\frac{|\alpha_{\omega,t,k}-H_{\omega,k}U_{k,t}|^2}{\beta_{\omega,t,k}}+(2-p)\xi_{\omega,t}^p. \tag{7.13}$$

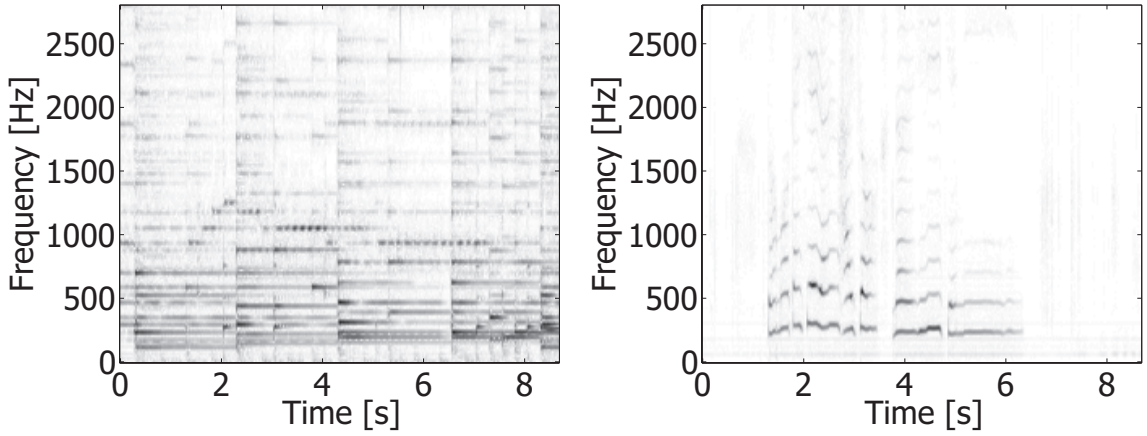
By differentiating $\mathcal{J}^+(H,U,\{\xi_{\omega,t}\}_{\omega,t},\{\alpha_{\omega,t,k}\}_{\omega,t,k},\{\beta_{\omega,t,k}\}_{\omega,t,k})$ partially with respect to H and U and setting them at zeros, we can obtain the following update equations:

$$H_{\omega,k}\leftarrow\frac{\sum_t C_{\omega,t}^{-1}U_{k,t}\Big(\beta_{\omega,t,k}^{-1}H_{\omega,k}h_{k,t}+Y_{\omega,t}-\sum_{k'}H_{\omega,k'}U_{k',t}\Big)}{\sum_t C_{\omega,t}^{-1}\beta_{\omega,t,k}^{-1}U_{k,t}^2} \tag{7.14}$$

$$U_{k,t}\leftarrow\frac{\sum_\omega C_{\omega,t}^{-1}H_{\omega,k}\Big(\beta_{\omega,t,k}^{-1}H_{\omega,k}U_{k,t}+Y_{\omega,t}-\sum_{k'}H_{\omega,k'}U_{k',t}\Big)}{\sum_\omega C_{\omega,t}^{-1}\beta_{\omega,t,k}^{-1}H_{\omega,k}^2} \tag{7.15}$$

where $C_{\omega,t}$ is defined as Eq. (7.10).

The parameters $\beta_{\omega,t,k}$ can be chosen arbitrarily subject to $\beta_{\omega,t,k}\in\mathbb{R}_{\ge0}$ and $\sum_k\beta_{\omega,k,t}=1$ for all ω and t, and so $\beta_{\omega,t,k}$ is allowed to be different at each iteration. The update rule of β can be derived similarly in the above.



(a) The spectrogram of an accompaniment sound

(b) The spectrogram of a singing voice

Figure 7.1: Examples of spectrograms of an accompaniment sound and a singing voice.

7.5 Application to Singing Voice Enhancement

7.5.1 Singing Voice Enhancement

In this section, we apply L_p-norm NMF to singing voice enhancement, of which aim is to enhance a singing voice enhancement in a monaural music signal. Singing voice enhancement is often used in music information retrieval (MIR) applications such as automatic lyrics recognition [104,105], automatic singer identification [106], and automatic karaoke generators [107].

To do this, we focus a difference in spectrogram between between accompaniment sounds and singing voices (Fig. 7.1). Music instruments can reproduce almost the same sounds each time they are played and music has a repeating musical structure. We can see the spectrograms of accompaniment signals as a low-rank matrix. By contrast, the spectra of singing voices are highly time-varying in pitch, timbre and loudness, and so the rank of the spectrograms of singing voices are relatively higher than those of accompaniment sounds. In addition, the spectrograms of singing voices are sparse as shown in Fig. 7.1 (b). Thus, if we try to approximate Y with HU, HU may correspond to the spectrograms of accompaniment sounds, and the spectrograms of singing voices can be viewed as outliers.

The enhancement scheme is depicted in Fig. 7.2. First, the spectrogram of an input sound is computed by the STFT. Second, the magnitude spectrogram of an input sound is regarded as a non-negative matrix, and L_p-norm NMF is applied to it. Third, some

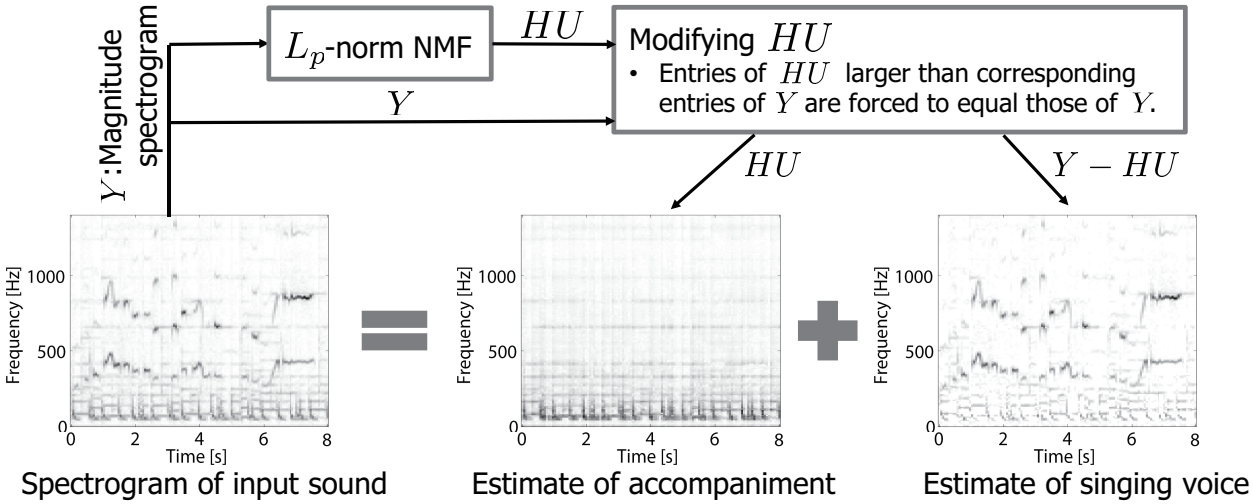


Figure 7.2: Proposed enhancement scheme using L_p-norm NMF.

time-frequency components of the obtained model spectrogram HU may be larger than corresponding components of Y, and so such components of HU are forced to equal those of Y. This means that the corresponding time-frequency components of Y are estimated not to contain the components of the singing voice. In summary, the estimated magnitude spectrogram of a singing voice $\hat{S}$ is derived as

$$\hat{S}_{\omega,t} = \begin{cases} Y_{\omega,t} - \sum_k H_{\omega,k}U_{k,t} & (Y_{\omega,t} \geq \sum_k H_{\omega,k}U_{k,t}) \\ 0 & (Y_{\omega,t} < \sum_k H_{\omega,k}U_{k,t}) \end{cases}. \tag{7.16}$$

The magnitude spectrograms of the accompaniment sound and the singing voice are converted into an audio signal by the inverse STFT, using the phase of the input spectrogram.

7.6 Experimental Evaluation

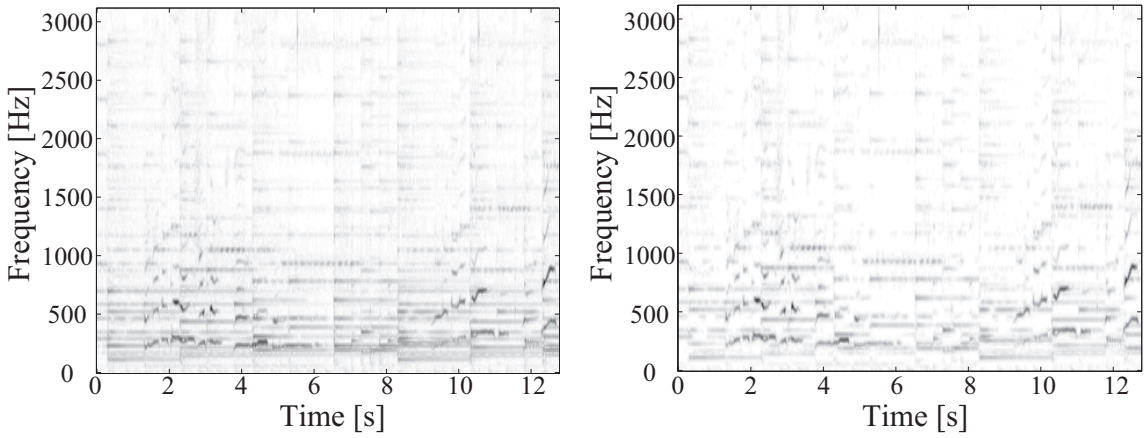
7.6.1 Experimental Conditions

To evaluate the performance of the proposed method, we conducted two experiments on singing voice enhancement: an evaluation of the effect of p, which controls sparsity, and frame length F, and a comparison of our results with the state-of-the-art [101, 108, 109].

The criteria for evaluating the singing voice enhancement were the normalized signal-to-distortion ratio (NSDR) and the global NSDR (GNSDR), given as

$$\mathrm{NSDR}(\{\hat{f}_t\}_t; \{f_t\}_t, \{x_t\}_t) = \mathrm{SDR}(\{\hat{f}_t\}_t, \{f_t\}_t) - \mathrm{SDR}(\{x_t\}_t, \{f_t\}_t), \tag{7.17}$$

$$\mathrm{GNSDR} = \frac{\sum_i w_i \mathrm{NSDR}(\{\hat{f}_{i,t}\}_t; \{f_{i,t}\}_t, \{x_{i,t}\}_t)}{\sum_i w_i}, \tag{7.18}$$



(a) Spectrogram of an input audio signal. (b) Spectrogram obtained with the proposed method.

Figure 7.3: Singing-voice-enhanced results obtained with the proposed method.

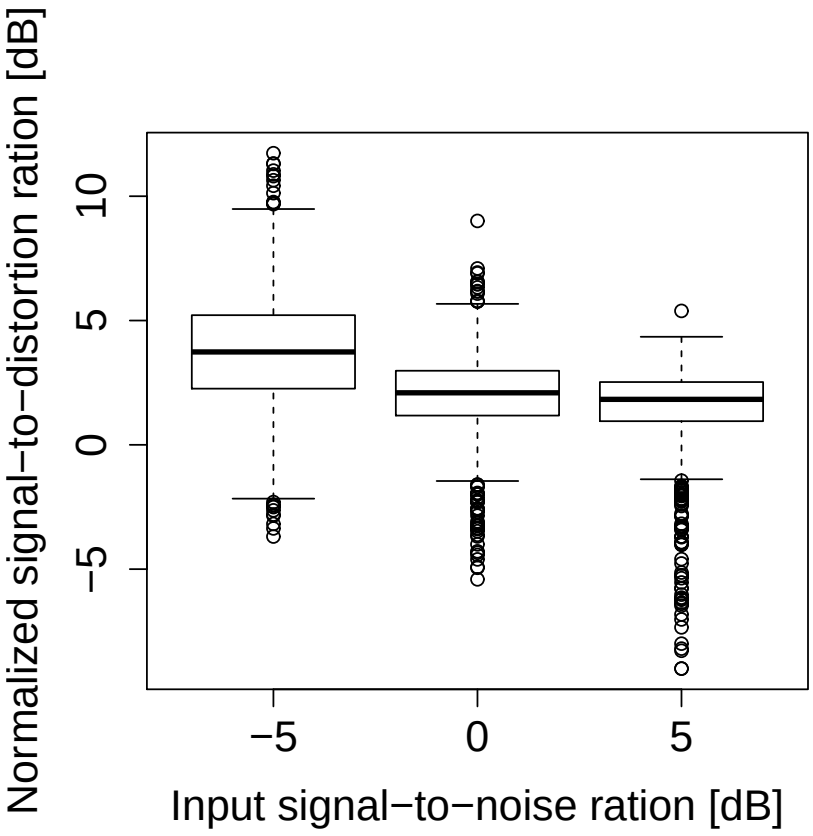


Figure 7.4: Box plot of the NSDRs by the proposed method for the MIR-1K dataset. The results for SNRs of $-5, 0$ and 5 dB are for $(p, F) = (1.7, 2048), (1.0, 2048)$ and $(0.8, 1024)$, respectively.

$$\mathrm{SDR}(\{\hat{f}_t\}_t, \{f_t\}_t) = 10 \log_{10} \frac{\sum_t \hat{f}_t f_t}{\Big(\sum_t \hat{f}_t\Big)\Big(\sum_t f_t\Big) - \sum_t \hat{f}_t f_t}, \tag{7.19}$$

Table 7.1: Comparison in GNSDR of the proposed method and methods presented in previous studies. F denotes the length of a frame in sample point. "Hsu", "Rafii" and "Huang" represent the corresponding methods [101, 108, 109].

Input SNR [dB]	Proposed method $F = 1024$	Proposed method $F = 2048$	Hsu	Rafii	Huang
-5	2.84 ($p = 1.6$)	3.70 ($p = 1.7$)	-0.51	0.52	1.51
0	1.93 ($p = 1.0$)	1.95 ($p = 1.0$)	0.91	1.11	2.37
5	1.43 ($p = 0.8$)	1.04 ($p = 0.8$)	0.17	1.10	2.57

where $\hat{f}_{i,t}$, $f_{i,t}$ and $x_{i,t}$ denote the estimated signal, the target signal and the input signal of the i-th piece. NSDR represents the improvement in SDR, and GNSDR denotes the weighted averages of the NSDR of all the music pieces by the length of the i-th piece, $\{w_i\}_i$. These criteria have also been employed in many previous studies [101, 108–112]. To calculate the SDR, we used the BSS Eval Toolbox [75, 113].

As an evaluation dataset, we used the MIR-1K dataset [114], following the evaluation framework in [101, 108, 109]. The dataset consists of 1000 Chinese song clips performed by amateur singers. The durations of the clips range from 4 to 13 s, and the audio signals are monaural with a sampling rate of 16 kHz. The accompaniment and vocal parts were recorded separately, and we could mix them with any signal-to-noise ratio (SNR), where the SNR corresponds to the voice to accompaniment ratio. The accompaniment and vocal parts for each clip were mixed at -5 dB (accompaniment is louder), 0 dB (same level) and 5 dB (vocal is louder) SNRs.

7.6.2 Effect of Sparsity and Frame Lengths

We first compared the proposed method in p and F. We used $p = 0.1, 0.2, \cdots, 2.0$ and $F = 512, 1024, 2048, 4096$ sample points. For STFT, the window function was the sine window, and the frame shifts were half the length of the frames. The number of bases was set at $K = 10$. The entries of W and H were initialized randomly. There were 200 iterations, which is supposed to be sufficient empirically.

Fig. 7.3 shows one of the enhanced results obtained with the proposed method. The figures show the spectrograms of the input signal and the enhanced result. We can see

that most of the accompanying sounds (vertically and horizontally smooth components) are suppressed, and the singing voice component of the spectrogram is clearer than that of the input spectrogram.

Fig. 7.4 shows the distributions of NSDRs for each SNR. The results were for $(p, F) = (1.7, 2048), (1.0, 2048), (0.8, 1024)$ for SNRs of $-5, 0$ and 5 dB. Since most of the NSDRs exceeded 0 dB, and we can confirm that the proposed method worked well for most of the input signals.

As illustrated in Fig. 7.5 for SNRs of $-5, 0, 5$ dB, the results show that the GNSDRs depended strongly on p for all frame lengths. The highest GNSDRs for all SNRs were 3.7 at $(p, F) = (1.7, 2048)$ for -5 dB SNR, 1.95 at $(p, F) = (1.0, 2048)$ for 0 dB SNR, and 1.43 at $(p, F) = (0.8, 1024)$ for 5 dB SNR. We can see that p at which GNSDR was the highest for each input SNR decreased as the input SNR became higher. With a high input SNR, the non-zero time-frequency components of the singing voice spectrogram are large, and increasing the sparsity is preferred. On the other hand, with a low input SNR, the non-zero time-frequency components are small, and decreasing the sparsity is preferred.

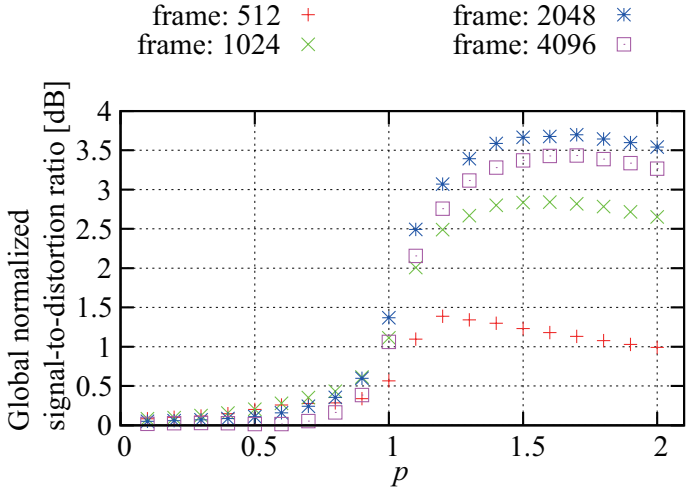
7.6.3 Comparison with Previous Studies

Finally, we compared the proposed method with the state-of-the-art [101, 108, 109]. The results are summarized in Tab. 7.1. The proposed method with $F = 1024$ outperformed two previous methods for all input SNRs. While the GNSDRs of the proposed method were lower than those of [101] at SNRs of 0 and 5 dB, the GNSDR with the proposed method was 0.4 to 2.4 dB larger than those of three previous methods at a SNR of -5 dB. This result indicates that the proposed method works well particularly in a low SNR environment.

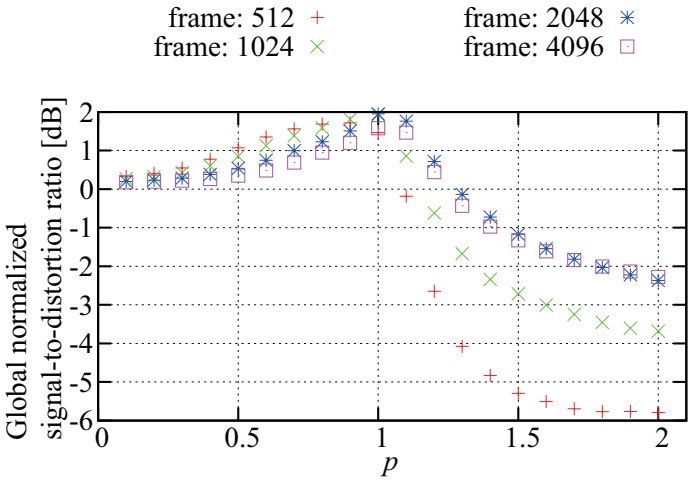
7.7 Summary

We have proposed a new NMF that minimizes the L_p norm of the reconstruction errors between a data matrix and the model. We have derived a computationally efficient algorithm according to the auxiliary function principle. This algorithm consists of multiplicative update equations and guarantee the non-negativity of H and U at each iteration. We have further generalized this algorithm for complex-valued matrix factorizations. We have applied L_p-norm NMF to singing voice enhancement and have showed experimentally that adequately selecting p improves the enhancement quality, and the proposed method outperformed three

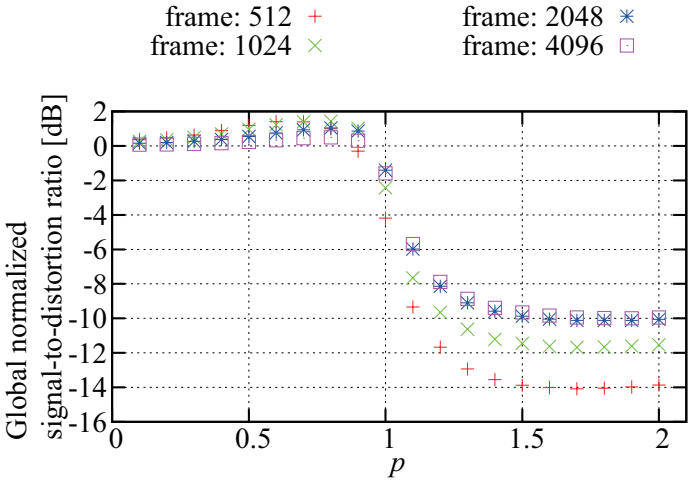
previous works under a low SNR situation. There are several ways to extend L_p-norm NMF to other applications. We think one promising application is speech enhancement since the spectrogram of background noise is sometimes approximated as low rank and the speech spectrogram is relatively sparse.



(a) GNSDR at −5 dB SNR.



(b) GNSDR at 0 dB SNR.



(c) GNSDR at 5 dB SNR.

Figure 7.5: GNSDRs of the proposed method with respect to p of the L_p norm and frame length F. Red, blue, green and purple points correspond to $F = 512, 1024, 2048, 4096$ sample points. The results are for (a) −5 dB, (b) 0 dB, and (c) 5 dB SNRs.

Chapter 8

Timbre Replacement of Drum Components in Music Audio Signals

8.1 Chapter Overview

This chapter presents a system that allows users to customize an audio signal of polyphonic music (*input*), without using musical scores, by replacing the frequency characteristics of harmonic sounds and the timbres of drum sounds with those of another audio signal of polyphonic music (*reference*). To develop the system, we first use a method that can separate the magnitude spectra of the input and reference signals into harmonic and percussive spectra. We characterize frequency characteristics of the harmonic spectra by two envelopes tracing spectral dips and peaks roughly, and the input harmonic spectra are modified such that their envelopes become similar to those of the reference harmonic spectra. The input and reference percussive spectrograms are further decomposed into those of individual drum instruments, and we replace the timbres of those drum instruments in the input piece with those in the reference piece. Through the subjective experiment, we show that our system can replace drum timbres and frequency characteristics adequately.

8.2 Introduction

Customizing existing musical pieces according to users' preferences is a challenging task in music signal processing. We would sometimes like to replace the timbres of instruments and audio textures of a musical piece with those of another musical piece. Professional audio

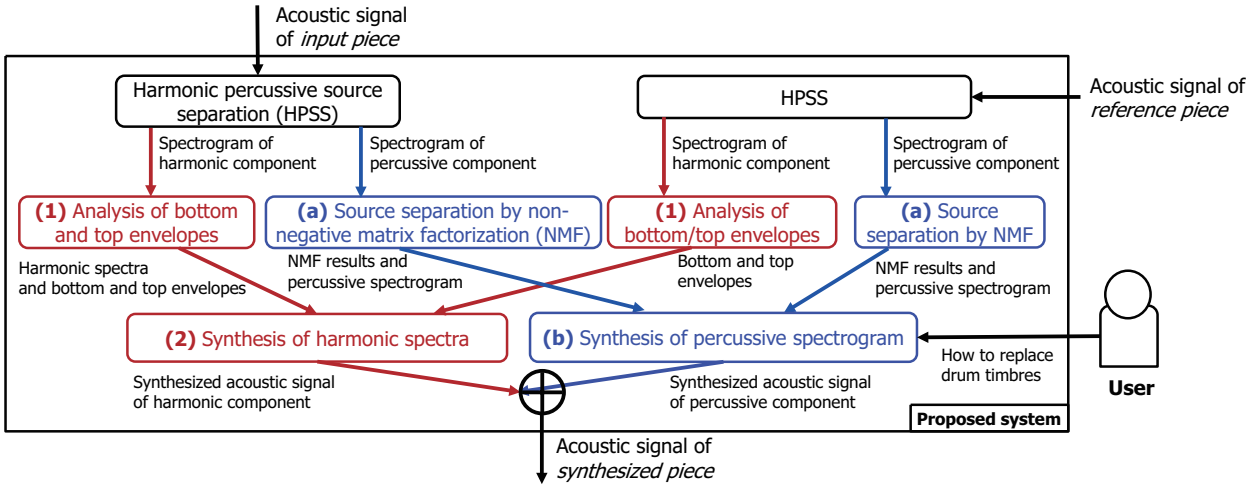


Figure 8.1: System outline for replacing drum timbres and frequency characteristics of the harmonic component. Red and blue modules relate to harmonic and percussive components of input and reference pieces.

engineers are able to perform such operations in the music production process by using effect units such as equalizers [115–119] that change the frequency characteristics of audio signals. However, sophisticated audio engineering skills are required for handling such equalizers effectively. It is therefore important to develop a new system that we can use intuitively without special skills.

Several highly functional systems have recently been proposed for intuitively customizing the audio signals of existing musical pieces. Itoyama *et al.* [120], for example, proposed an instrument equalizer that can change the volume of individual musical instruments independently. Yasuraoka *et al.* [121] developed a system that can replace the timbres and phrases of some instrument with users' own performances. Note that these methods are based on score-informed source separation techniques that require score information about the musical pieces (MIDI files). Yoshii *et al.* [122], on the other hand, developed a drum instrument equalizer called *Drumix* that can change the volume of bass and snare drums and replace their timbres and patterns with others prepared in advance. To achieve this, audio signals of bass and snare drums are separated from polyphonic audio signals without using musical scores. In this system, however, only the drum component can be changed or replaced. In addition, users would often need to prepare isolated drum sounds (called *reference*) with which they want to replace original drum sounds. Here we are concerned with developing an easier-to-handle system that only requires the users to specify a different musical piece as a reference.

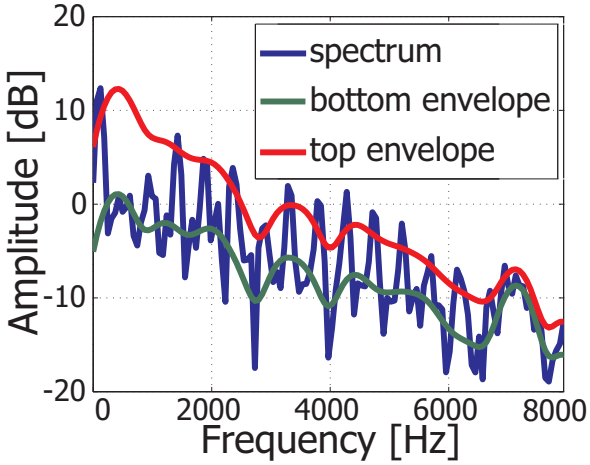


Figure 8.2: Bottom (green) and top (red) envelopes of a spectrum (blue). The envelopes trace dips and peaks of a spectrum roughly.

In this chapter, we propose a system that allows users to customize a musical piece (called *input*), without using musical scores, by replacing the timbres of drum instruments and the frequency characteristics of pitched instruments including vocals with those of another music piece (reference). We consider the problems of customizing the drum sounds and the pitched instruments separately, because they have different effects on audio textures. As illustrated in Fig. 8.1, the audio signals of the input and reference pieces are separated into harmonic and percussive components, respectively, by using a harmonic percussive source separation (HPSS) method [123] based on spectral anisotropy. The system then (1) analyzes the frequency characteristics of the spectra of the harmonic component (hereafter *harmonic spectra*) of the input piece by using a spectral-envelope-based method presented by [93], and (2) adapts those characteristics to the frequency characteristics of the reference harmonic spectra. Moreover, (a) the spectrograms of the percussive components (hereafter *percussive spectrograms*) of the input and reference pieces are further decomposed into individual drum instruments such as bass and snare drums, and (b) the drum timbres of the input piece are replaced with those of the reference piece. In the following, we describe a replacement method of frequency characteristics for harmonic spectra and a replacement method of drum timbres for percussive spectrograms.

8.3 Frequency Characteristics Replacement

The goal is to modify the frequency characteristics of the harmonic spectra obtained with HPSS from an input piece by referring to those of a reference piece. The frequency

characteristics of a musical piece are closely related to the timbres of the musical instruments used in that piece. If score information is available, a music audio signal could be separated into individual instrument parts [120, 121]. However, blind source separation is still difficult when score information is not available. We therefore take a different approach to avoid the need for perfect separation.

We here modify the input magnitude spectrum using two envelopes, named *bottom and top envelopes*, which trace the dips and peaks of the spectrum roughly as illustrated in Fig. 8.2. The bottom envelope expresses a flat and wide-band component in the spectrum, and the top envelope represents a spiky component in the spectrum. We can assume that the flat component corresponds to the spectrum of vocal consonants and attack sounds of musical instruments, while the spike component corresponds to the harmonic structures of musical instruments. Thus, individually modifying these envelopes allows us to approximately change the frequency characteristics of the musical instruments. The modified magnitude spectra are converted into an audio signal using the phases of the input harmonic spectra.

8.3.1 Mathematical Model for Bottom and Top Envelopes

We describe each envelope using a Gaussian mixture model (GMM) as a function of the frequency ω:

$$\Psi(\omega;\mathbf{a}) := \sum_k a_k \psi_k(\omega),\ \psi_k(\omega) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}\Big(\omega - \frac{k f_{\rm nyq}}{K}\Big)\Big] \tag{8.1}$$

where $\mathbf{a} := \{a_k\}_{k=1}^K$, and $f_{\rm nyq}$ stands for a Nyquist frequency. $a_k \geq 0$ denotes the power of the k-th Gaussian $\psi_k(\omega)$ with the average $k f_{\rm nyq}/K$ and the variance σ^2.

We first estimate $\mathbf{a}$ for the bottom envelopes of the input and reference pieces respectively by fitting $\Psi(\omega;\mathbf{a})$ to their harmonic spectra, and also estimate $\mathbf{a}$ for the top envelopes (see Sec. 8.3.3). We then design a filter that converts the input envelopes so that their time averages and variances equal those of the reference envelopes. Finally, by using the converted version of the input envelopes, we convert the input magnitude spectra.

8.3.2 Spectral Synthesis via Bottom and Top Envelopes

We consider converting the input piece so that the bottom and top envelopes of the converted version become similar to those of the reference piece. Let us define the averages and variances in time of the envelopes of the input and reference harmonic spectra as $\mu_\omega^{(l)}$ and

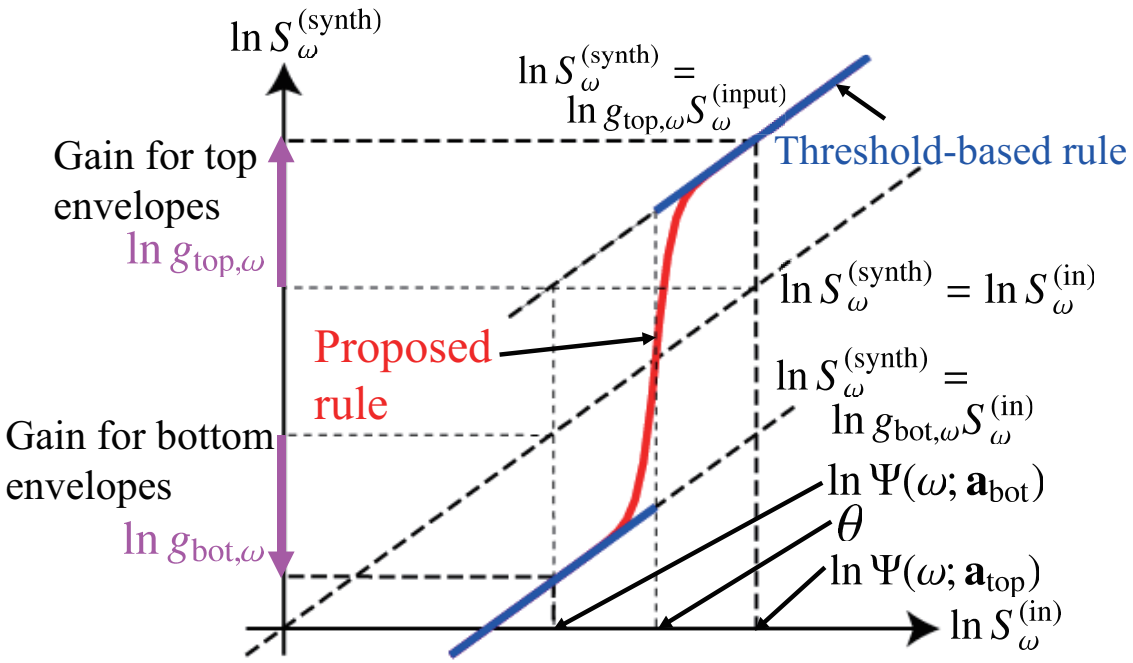


Figure 8.3: The proposed (red curve) and threshold-based (blue lines) rules of modifying a spectrum in the log-spectral domain. The horizontal and vertical axes are an magnitude spectral elements of input and synthesized pieces.

$V_\omega^{(l)}$ for $l = \mathrm{in}, \mathrm{ref}$, respectively. Assuming that the envelopes follow normal distributions, the distributions of the converted input envelopes approach those of the reference envelopes by minimizing a measure between the distributions. As one such measure, we can use the Kullback-Leibler divergence, and derive the gains as

$$g_\omega = \frac{\mu_\omega^{(\mathrm{in})}\mu_\omega^{(\mathrm{ref})} + \sqrt{(\mu_\omega^{(\mathrm{in})}\mu_\omega^{(\mathrm{ref})})^2 - 4\{V_\omega^{(\mathrm{in})} + (\mu_\omega^{(\mathrm{in})})^2\}V_\omega^{(\mathrm{ref})}}}{2\{V_\omega^{(\mathrm{in})} + (\mu_\omega^{(\mathrm{in})})^2\}}. \tag{8.2}$$

Next, we show the conversion rule for the harmonic magnitude spectrum ($S_\omega^{(\mathrm{in})}$) of the input piece by using the gains for the bottom and top envelopes in the log-spectral domain. When modifying the bottom envelope, we want to modify only the flat component (and keep the spiky component fixed). On the other hand, when modifying the top envelope, we want to modify only the spiky component (and keep the flat component fixed). To do this, we multiply the spectral components above or near the top envelope by $g_{\mathrm{top},\omega}$ (the gain factor for the top envelope), and multiply the spectral components below or near the bottom envelope by $g_{\mathrm{bot},\omega}$ (the gain factor for the bottom envelope). One such rule is a threshold-based rule which means that we divide the set of spectral components into two sets, one consisting of the components above or near the top envelope and the other consisting of the components below or near the bottom envelope. We multiply the former and latter sets

by $g_{\mathrm{top},\omega}$ and $g_{\mathrm{bot},\omega}$, respectively. Fig. 8.3 illustrates the rule where $S_\omega^{(\mathrm{synth})}$ is a synthesized magnitude spectrum and a threshold $\theta := \{\ln(\Psi(\omega;\mathbf{a}_{\mathrm{bot}})\Psi(\omega;\mathbf{a}_{\mathrm{top}}))\}/2$ is the midpoint of the bottom and top envelopes ($\Psi(\omega;\mathbf{a}_{\mathrm{bot}})$ and $\Psi(\omega;\mathbf{a}_{\mathrm{top}})$) of the input piece in the log-spectral domain. However, the rule changes spectral elements near θ with discontinuity. To avoid the discontinuity, we use the relaxed rule as shown in Fig. 8.3:

$$\ln S_\omega^{(\mathrm{synth})} = \ln g_{\mathrm{bot},\omega} S_\omega^{(\mathrm{in})} + \ln \frac{g_{\mathrm{top},\omega}}{g_{\mathrm{bot},\omega}} f\Big(\frac{\ln S_\omega^{(\mathrm{in})} - \theta}{\rho \ln(\Psi(\omega;\mathbf{a}_{\mathrm{top}})/\Psi(\omega;\mathbf{a}_{\mathrm{bot}}))}\Big) \tag{8.3}$$

$$f(x) := \frac{1}{1+\exp(-x)} = \begin{cases} 0 & (x \to -\infty) \\ 1 & (x \to \infty) \end{cases} \tag{8.4}$$

where $\rho > 0$. Note that (8.3) is equivalent to the threshold-based rule when $\rho \to 0$.

8.3.3 Estimation of Bottom and Top Envelopes

Estimation algorithms of spectral envelopes presented in this section have already been presented in [93]. However, it has not been published in English and so we will review it details in the rest of this section.

Estimation of Bottom Envelopes

When estimating the bottom envelope $\Psi(\omega;\mathbf{a})$, we can use the Itakura-Saito divergence (IS divergence) [124] as a cost function. The estimation requires a cost function that is lower for the spectral dips than for the spectral peaks. The IS divergence meets the requirement as illustrated in Fig. 8.4. Let S_ω be an magnitude spectrum. The cost function is described as

$$\mathcal{J}_{\mathrm{bot}}(\mathbf{a}) := \sum_\omega D_{IS}(\Psi(\omega;\mathbf{a})||S_\omega), \tag{8.5}$$

$$D_{IS}(\Psi(\omega;\mathbf{a})||S_\omega) := \frac{\Psi(\omega;\mathbf{a})}{S_\omega} - \ln\frac{\Psi(\omega;\mathbf{a})}{S_\omega} - 1 \tag{8.6}$$

where $D_{IS}(\cdot||\cdot)$ is the IS divergence. Minimizing $\mathcal{J}_{\mathrm{bot}}(\mathbf{a})$ directly is difficult, because of the non-linearity of the second term of (8.5).

We can use the auxiliary function method [71–73]. Given a cost function $\mathcal{J}$, we introduce an auxiliary variable λ and an auxiliary function $\mathcal{J}^+(x,\lambda)$ such that $\mathcal{J}(x) \le \mathcal{J}^+(x,\lambda)$. We can then monotonically decrease $\mathcal{J}(x)$ indirectly by minimizing $\mathcal{J}^+(x,\lambda)$ with respect to x and λ iteratively.

The auxiliary function of $\mathcal{J}_{\rm bot}(\mathbf{a})$ can be defined as

$$\mathcal{J}^{+}_{\rm bot}(\mathbf{a},\boldsymbol{\lambda}) := \sum_{\omega}\Big\{\sum_{k}\Big(\frac{a_k\psi_k(\omega)}{S_\omega} - \lambda_k(\omega)\ln\frac{a_k\psi_k(\omega)}{\lambda_k(\omega)S_\omega}\Big) - 1\Big\} \tag{8.7}$$

where $\boldsymbol{\lambda} = \{\lambda_k(\omega)\}^{K,W}_{k=1,\omega=1}$ is a series of auxiliary variables such that $\forall\omega, \sum_k \lambda_k(\omega) = 1,\ \lambda_k(\omega) \geq 0$. The auxiliary function is obtained by Jensen's inequality based on the concavity of the logarithmic function in the second term of (8.5). By solving $\partial\mathcal{J}^{+}_{\rm bot}(\mathbf{a},\boldsymbol{\lambda})/\partial a_k = 0$ and the equality condition of $\mathcal{J}_{\rm bot}(\mathbf{a}) = \mathcal{J}^{+}_{\rm bot}(\mathbf{a},\boldsymbol{\lambda})$, we can obtain

$$a_k \leftarrow \frac{\sum_\omega \lambda_k(\omega)}{\sum_\omega \psi_k(\omega)/S_\omega},\ \lambda_k(\omega) \leftarrow \frac{a_k\psi_k(\omega)}{\sum_{k'} a_{k'}\psi_{k'}(\omega)}. \tag{8.8}$$

Estimation of Top Envelopes

The estimation of the top envelope $\Psi(\omega;\mathbf{a})$ requires a cost function that is higher for the spectral dips than for the spectral peaks. This is the opposite requirement for that in Sec. 8.3.3. The IS divergence is asymmetric as shown in Fig. 8.4, thus exchanging $\Psi(\omega;\mathbf{a})$ with S_ω of (8.6) leads to the opposite property to (8.6), and $D_{IS}(S_\omega||\Psi(\omega;\mathbf{a}))$ meets the requirement. Suppose that the bottom envelope $\Psi(\omega;\mathbf{a}_{\rm bot})$ was estimated. The cost function is defined as

$$\mathcal{J}_{\rm top}(\mathbf{a}) := P(\mathbf{a};\mathbf{a}_{\rm bot}) + \sum_{\omega} D_{IS}(S_\omega||\Psi(\omega;\mathbf{a})) \tag{8.9}$$

where $P(\mathbf{a};\mathbf{a}_{\rm bot}) := \sum_k \eta_k a_{{\rm bot},k}/a_k$ is a penalty term for the closeness between the bottom and top envelopes, and $\eta_k \geq 0$ is the weight of $a_{{\rm bot},k}/a_k$. Direct minimization of $\mathcal{J}_{\rm top}(\mathbf{a})$ is also difficult because the IS divergence in the second term of (8.9) includes non-linear terms as described in (8.6).

Here we can define the auxiliary function of $\mathcal{J}_{\rm top}(\mathbf{a})$ as

$$\begin{aligned}\mathcal{J}^{+}_{\rm top}(\mathbf{a},\boldsymbol{\nu},\boldsymbol{h}) :=& P(\mathbf{a};\mathbf{a}_{\rm bot}) + \sum_{\omega}\Big\{\sum_{k}\frac{(\nu_k(\omega))^2 S_\omega}{a_k\psi_k(\omega)} + \ln h(\omega) \\ &+ \frac{1}{h(\omega)}\Big(\sum_k a_k\psi_k(\omega) - h(\omega)\Big) - \ln S_\omega - 1\Big\}\end{aligned} \tag{8.10}$$

where $\boldsymbol{\nu} = \{\nu_k(\omega)\}^{K,W}_{k=1,\omega=1}$ and $\mathbf{h} = \{h(\omega)\}^{W}_{\omega=1}$ are series of auxiliary variables such that $\forall\omega, \sum_k \nu_k(\omega) = 1,\ \nu_k(\omega) \geq 0, h(\omega) > 0$. This inequality is derived from the following two inequalities for the non-linear terms:

$$\frac{1}{\sum_k x_k} \leq \sum_k \frac{\nu_k^2}{x_k},\ \ln x \leq \ln h + \frac{1}{h}(x-h). \tag{8.11}$$

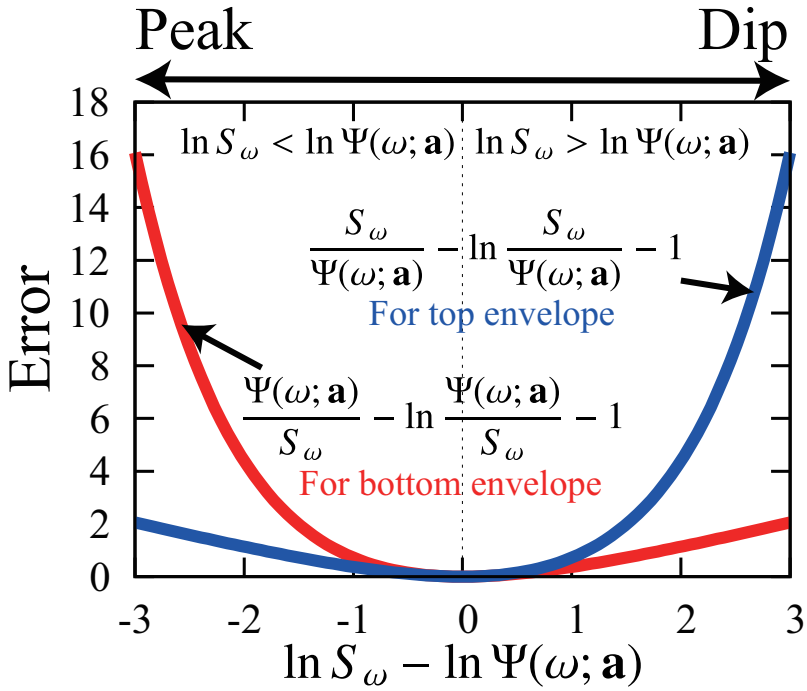


Figure 8.4: The Itakura-Saito divergences for bottom and top envelopes.

where $\forall k, \nu_k \geq 0$ and $h > 0$ are auxiliary variables such that $\sum_k \nu_k = 1$. The first inequality is obtained by Jensen's inequality for $1/\sum_k x_k$ and the second inequality is a first-order Taylor-series approximation of $\ln x$ around h. By solving $\partial \mathcal{J}^+_{\text{top}}(\mathbf{a}, \boldsymbol{\nu}, \boldsymbol{h})/\partial a_k = 0$ and the equality condition of $\mathcal{J}_{\text{top}}(\mathbf{a}) = \mathcal{J}^+_{\text{top}}(\mathbf{a}, \boldsymbol{\nu}, \boldsymbol{h})$, update rules can be derived as

$$a_k \leftarrow \left\{ \frac{\eta_k a_{\text{bot},k} + \sum_\omega (\nu_k(\omega))^2 S_\omega / \psi_k(\omega)}{\sum_\omega \psi_k(\omega)/h(\omega)} \right\}^{1/2}, \tag{8.12}$$

$$\nu_k(\omega) \leftarrow \frac{a_k \psi_k(\omega)}{\sum_{k'} a_{k'} \psi_{k'}(\omega)}, \quad h(\omega) \leftarrow \sum_k a_k \psi_k(\omega). \tag{8.13}$$

(8.12) does not guarantee $a_k \geq a_{\text{bot},k}$, and we set $a_k = a_{\text{bot},k}$ when $a_k < a_{\text{bot},k}$.

8.4 Drum Timbre Replacement

To replace drum timbres, we first decompose the percussive magnitude spectrograms into approximately those of individual drum instruments. The decomposition can be achieved by non-negative matrix factorization (NMF) [125] and Wiener filtering. We call a component of the decomposed spectrograms *a basis spectrogram*. NMF approximates the magnitude spectrograms by a product of two non-negative matrices, one of which is a basis matrix. Each column of the basis matrix corresponds to the magnitude spectrum of an individual drum sound, and the corresponding row of the activation matrix represents its temporal

activity. The users are then allowed to specify which drum sounds (bases) in the input piece they want to replace with which drum sounds in the reference piece. According to this choice, the chosen drum timbres of the input piece are replaced with those of the reference piece for each basis.

8.4.1 Equalizing Method

One simple method for replacing drum timbres, called *the equalizing (EQ) method*, is to apply gains to a basis spectrogram of the input piece such that the drum timbre of the input basis becomes similar to that of the reference basis. The input and reference bases represents the timbral characteristics of their drum sounds, and we use the gain that equalize the input and reference bases for each frequency bin. Let us define the complex basis spectrogram of the input piece and its basis as $Y^{(\mathrm{in})}_{\omega,t}$ and $H^{(\mathrm{in})}_{\omega}$. Using the corresponding reference basis $H^{(\mathrm{ref})}_{\omega}$, we can obtain the synthesized complex spectrogram $Y^{(\mathrm{synth})}_{\omega,t}$ for the basis as $Y^{(\mathrm{synth})}_{\omega,t} = Y^{(\mathrm{in})}_{\omega,t} H^{(\mathrm{ref})}_{\omega} / H^{(\mathrm{in})}_{\omega}$ for $\omega \in [1, W]$ and $t \in [1, T]$.

This method only requires applying gains to the input basis spectrograms uniformly in time. However, when there is a large difference between the timbres of the specified drum sounds, the method often amplifies low-energy frequency elements excessively, and so the resulting converted version would sound very noisy and the method fails to replace the drum timbres adequately.

8.4.2 Copy and Paste Method

To avoid the problem of the EQ method, we directly use basis spectrograms of the reference piece. The reference basis spectra include the drum timbre which we want, and by appropriately copying and pasting the reference basis spectra, we can obtain the percussive spectrogram with the reference drum timbres and the input temporal activities. We call the method *the copy and paste (CP) method*.

This method requires how to copy and paste the reference basis spectra with keeping the input temporal activities and how to reduce noise occurred by this method. Features should be less sensitive to the drum timbres but reflect temporal activities. As the features, the NMF activations are available. Furthermore, there are three requirements related to the noise reduction. Noise occurs when previously remote high-energy spectra are placed adjacent to each other. To suppress the noise, (i) time-continuous segments should be used

and (ii) the segment boundaries should be established when the activation is low. Since unsupervised source separation is still a challenging problem, the basis spectra may include a non-percussive component due to imperfect source separation, and (iii) the use of basis spectra that include non-percussive components should be avoided.

The problem can be formulated as an alignment problem. The requirements of (i), (ii), and (iii) are described as cost functions, and the cumulative cost $\mathcal{I}_t(\tau)$ can be written recursively as

$$\mathcal{I}_t(\tau) := \begin{cases} O_{t,\tau} & (t=1) \\ O_{t,\tau} + \max_{\tau'}\{C_{\tau',\tau} + \mathcal{I}_{t-1}(\tau')\} & (t>1) \end{cases}, \tag{8.14}$$

$$O_{t,\tau} := \alpha D(\tilde{U}_t^{(\mathrm{in})} || \tilde{U}_\tau^{(\mathrm{ref})}) + \beta P_\tau \tag{8.15}$$

where τ is a time index of the reference piece, $\alpha > 0$ and $\beta > 0$ are the weights of $D(\tilde{U}_t^{(\mathrm{in})} || \tilde{U}_\tau^{(\mathrm{ref})})$ and P_τ, and $\tilde{U}_t^{(l)} := U_t^{(l)} / \max_t\{U_t^{(l)}\}$ for $l = \mathrm{in}, \mathrm{ref}$. The first term of (8.15) indicates the generalized I-divergence between the two normalized activations. P_τ represents the degree to which the reference basis spectrum at the τ-th frame includes non-percussive components: the term becomes larger as the number of non-percussive components in the spectrum (requirement (iii)). $C_{\tau',\tau}$ is the transition cost from the τ'-th frame to the τ-th frame of the reference piece:

$$C_{\tau',\tau} = \begin{cases} 1 & (\tau = \tau'+1) \\ c + \gamma(\tilde{U}_{\tau'}^{(\mathrm{ref})} + \tilde{U}_\tau^{(\mathrm{ref})}) & (\tau \neq \tau'+1) \end{cases}. \tag{8.16}$$

The constant c expresses a cost for all other transitions except for a straight one. We set $c > 1$ and this ensures that a straight transition occurs more frequently than the others (requirement (i)). The second term of (8.16) for $\tau \neq \tau' + 1$ indicates that transitions to remote frames tend to occur when the activations are low (requirement (ii)), and $\gamma > 0$ is the weight of $\tilde{U}_{\tau'}^{(\mathrm{ref})} + \tilde{U}_\tau^{(\mathrm{ref})}$. We can obtain the alignment as an optimal path that minimizes the cumulative cost by the Viterbi algorithm [126].

The input basis spectra may include the non-percussive components because of imperfect source separation. In this case, the input basis spectra which may include the non-percussive components are replaced with the reference basis spectra by the CP method, and the input basis spectra loses the input non-percussive components. To recover the components, we make an extra processing. The components tend to have low energy, and they would probably be included in the input percussive spectra with low energy. We replace synthesized

percussive spectra $\{Y_{\omega,t}^{(\mathrm{synth})}\}_\omega$ with the corresponding input percussive spectra $\{Y_{\omega,t}^{(\mathrm{in})}\}_\omega$ when $\sum_\omega Y_{\omega,t}^{(\mathrm{in})}$ is lower than a threshold ϵ.

8.5 Experimental Evaluation

8.5.1 Experimental Conditions

We conducted an experiment to evaluate the performance of the system subjectively. We prepared three audio signals of musical pieces (10 s for each piece) from the RWC popular music and music genre databases [2] as input and reference pieces, and they were down-sampled from 44.1 to 22.05 kHz. Then, we synthesized six pairs of these musical audio signals. Some synthesized sounds are available at `http://tomohikonakamura.github.io/Tomohiko-Nakamura/demo/TimbreReplacer.html`. The signals of the input and reference pieces were converted into spectrograms with the short time Fourier transform (STFT) with a 512-sample Hanning window and a 256-sample frame shift, and the synthesized spectrograms were converted into audio signals by the inverse STFT with the same window and frame shift. The parameters of the frequency characteristics replacement were set at $\sigma = 240$ Hz and $(K, \rho, \eta_k) = (30, 0.2, 100/k)$ for $k \in [1, K]$. Then, the parameter a_k of the envelope model was initialized by $\sum_\omega S_\omega / K$ for $k \in [1, K]$, all frames and all pieces. For the NMF of the percussive spectrograms, we set the number of bases at 4, and used the generalized I-divergence. The CP method was compared with the EQ method, and one of the authors chose which drum sounds in the input piece were replaced with which drum sounds in the reference piece. The parameters for the drum timbre replacement were set at $(M, \alpha, \beta, \gamma, c, \epsilon) = (4, 0.5, 3, 10, 3, 100)$. A negative log posterior, which was computed by the L2-regularized L1-loss support vector classifier (SVC) [127], was used as P_τ, and the SVC was trained to distinguish between percussive and non-percussive instruments, using the RWC instrument database [2].

We asked 9 subjects how adequately they felt that (1) the drum timbres of the input piece were replaced with those of the reference piece and (2) the timbres of the input harmonic components were replaced with those of the reference piece. The subjects were allowed to listen to the input, reference, and synthesized pieces as well as their harmonic and percussive components as many times as they liked. They then evaluated (1) and (2) for each synthesized piece on a scale of 1 to 5. 1 point means that the timbres were not replaced and 5

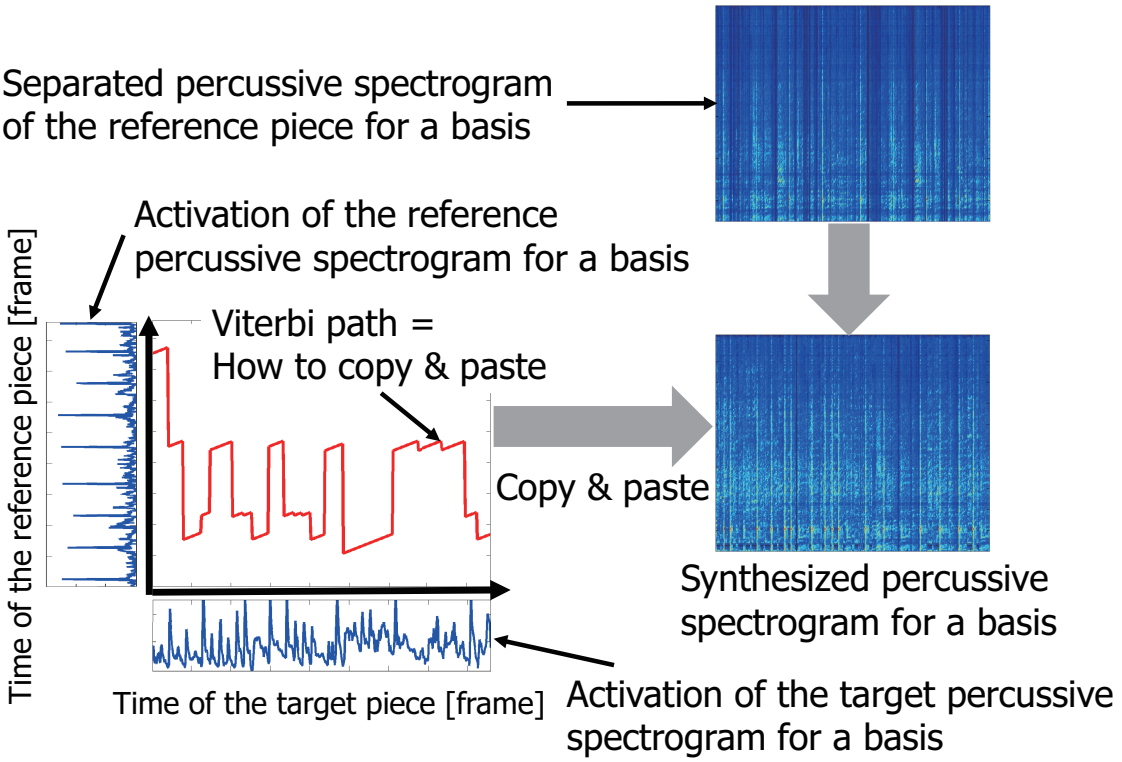


Figure 8.5: Outline of the copy and paste method.

points indicates that the timbres were replaced perfectly.

8.5.2 Result and Discussion

The average scores of (1) with standard errors were 2.37 ± 0.15 and 2.83 ± 0.15 for the EQ and the CP methods. The CP method result was provided prior to that provided by the EQ method, in particular when the drum timbres were very different as we mentioned in Sec. 8.4. The average score of (2) with standard errors was 2.5 ± 0.1. The results show that the subjects perceived the replaced drum timbres and frequency characteristics, and that the system works well.

We asked the subjects to comment about the synthesized pieces. One subject said that he wanted to control the degree to which drum timbres and frequency characteristics were converted. This opinion indicates that it is important to enables users to adjust the conversions. Additionally, another subject mentioned that replacing vocal timbres separately would change the moods of the musical pieces more drastically. We plan to replace vocal timbres by using an extension of HPSS [128] for vocal extraction.

8.6 Summary

We have described a system that can replace the drum timbres and frequency characteristics of harmonic components in polyphonic audio signals without using musical scores. We have proposed an algorithm that can modify a harmonic magnitude spectrum via its bottom and top envelopes. We have also discussed two methods for replacing drum timbres. The EQ method applies gains to basis spectrograms by the proportions of the NMF bases of the input percussive spectrograms and those of the reference percussive spectrograms. The CP method copies and pastes the basis spectra of a reference piece, according to NMF activations of the input and reference pieces. Through the subjective experiment, we confirmed that the system can replace drum timbres and frequency characteristics adequately.

Chapter 9

Conclusion

We discussed a spectrogram-aware approach for monaural audio source separation. To realize the approach, we considered the three principles:

[P1] Use spectrograms having a log-frequency resolution obtained with the CWT.

[P2] Utilize the source-filter model.

[P3] Take into account the spectral leakage.

To develop methods that satisfy all the principles simultaneously, we addressed the following issues in Chapters 3, 4 and 5, respectively.

[I1] How can we incorporate the source-filter model in the CWT domain ?

[I2] How can we describe the spectral leakage in the CWT domain ?

[I3] How can we simultaneously incorporate the source-filter model and the spectral leakage in the CWT domain ?

In Chapter 3, we presented a monaural audio source separation method that satisfies the principles [P1] and [P2] simultaneously. The method describes the spectrogram of a mixture signal as the sum of the products between the shifted copies of excitation spectrum templates, which represent spectra of different F_0s, and filter spectrogram templates, which represent the dynamics of the timbre. Iterative algorithms of estimating parameters for the I divergence and IS divergence criterions were derived based on the auxiliary function approach. Experiments revealed that the incorporation of the source-filter model is effective in terms of the source separation accuracy.

In Chapter 4, we presented a new approach for monaural source separation, called HTFD, that satisfies the principles [P1] and [P3]. HTFD combines the features of the models employed in the NMF and HTC approaches. Since the present spectrogram model was derived from an analytic signal model, the parameters of the spectrogram model in the CWT domain can be associated with those of the analytic signal model in the time domain. The parameter relationship enables us to describe the spectral leakage of the signal model in the log-frequency domain. We conducted an experiment to compare HTFD with the harmonic NMF, which does not take into account the spectral leakage, and found that the incorporation of the spectral leakage is effective in the CWT and STFT domains. Furthermore, we implemented HTFD for STFT spectrograms and compared it with HTFD for CWT spectrograms, and obtained that the CWT-domain HTFD outperformed the STFT-domain HTFD in source separation accuracy. Moreover, we confirmed that the CWT-domain HTFD outperformed the harmonic NMF for STFT spectrograms in audio quality of separated signals through an subjective experiment. These results show that CWT is more suited for monaural source separation of harmonic audio signals compared to STFT.

Using the explicit relationship of parameters between the CWT domain and the time domain, we incorporate the source-filter model into the spectrogram model of HTFD in Chapter 5. The present method satisfies all the principles simultaneously. The source-filter model is defined in the discrete time domain, and thus we can associate the parameters of the source-filter model with those of the spectrogram model defined in the CWT domain via the analytic signal model. An iterative algorithm was derived based on the auxiliary function approach. Experiments showed that the incorporation of the source-filter model improves the source-separation accuracy.

In Chapter 6, we addressed the problem of estimating an unknown signal from a given magnitude CWT spectrogram. Due to the redundancy of CWT spectrograms, any CWT spectrograms satisfy a certain condition to ensure they correspond to time domain signals, which we call the consistency condition. The problem of the phase estimation was formulated as that of minimizing a numerical criterion describing how far a complex vector deviates from the consistency condition. We presented fast phase estimation algorithms with guaranteed convergence using the auxiliary function principle. Experimental evaluations have demonstrated that the present algorithms are around 75 times faster than an algorithm proposed in previous literature, and the reconstructed signals obtained with the present algorithms have almost the same audio quality as original sounds. Moreover, we extended the present

algorithms to work in real time and showed the efficiency of the real-time version of the algorithms in experiments.

In Chapter 7, we present a method of enhancing singing voices in music audio signals using NMF with the L_p norm criterion by focusing on that spectrograms of singing voices can be seen as sparse matrices while spectrograms of accompaniment sounds can be seen as low-rank matrices. An experimental evaluation showed that reasonably good enhancement results were obtained with appropriate choices of p.

In Chapter 8, we develop a system that allows users to edit a music audio signal without using musical scores by replacing the timbres of drum sounds and the frequency characteristics of harmonic sounds with those of another music signal. The present system was confirmed to work well through a subjective experiment.

Bibliography

[1] H. Kameoka, "Statistical Approach to Multipitch Analysis," Ph.D. dissertation, The University of Tokyo, Mar. 2007.

[2] M. Goto, "Development of the RWC Music Database," in *Proceedings of International Congress on Acoustics*, vol. 1, 2004, pp. 553–556.

[3] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Brill, 2013.

[4] C. C. Wier, W. Jesteadt, and D. M. Green, "Frequency discrimination as a function of frequency and sensation level," *Journal of the Acoustical Society of America*, vol. 61, no. 1, pp. 178–184, 1977.

[5] I. Daubechies, *Ten lectures on wavelets*. SIAM, 1992, vol. 61.

[6] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, p. 425, 1991.

[7] J. J. Burred and T. Sikora, "Comparison of frequency-warped representations for source separation of stereo mixtures," in *Proceedings of Audio Engineering Society Convention*, 2006.

[8] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proceedings of International Conference Independent Component Analysis and Blind Signal Separation*, 2006, pp. 700–707.

[9] J. P. de León, F. Beltrán, and J. R. Beltrán, "A complex wavelet based fundamental frequency estimator in single-channel polyphonic signals," in *Proceedings of International Conference on Digital Audio Effects*, 2013, pp. 47–54.

[10] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent F0 estimation and source separation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 574–578.

[11] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[12] K. Nishi, M. Abe, and S. Ando, "Multiple pitch tracking and harmonic segregation algorithm for auditory scene analysis," *Transactions of the Society of Instrument and Control Engineers*, vol. 34, no. 6, pp. 483–490, 1998, in Japanese.

[13] M. Unoki and M. Akagi, "A method of extracting the harmonic tone from noisy signal based on auditory scene analysis," *The IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (Japanese Edition)*, vol. 82, no. 10, pp. 1497–1507, 1999, in Japanese.

[14] M. Abe and S. Ando, "Auditory scene analysis based on time-frequency integration of shared FM and AM(I): Lagrange differential features and frequency-axis integration," *The IEICE Transactions on Information and Systems (Japanese Edition)*, vol. 83, no. 2, pp. 458–467, 2000, in Japanese.

[15] ——, "Auditory scene analysis based on time-frequency integration of shared FM and AM(II) : Optimum time-domain integration and stream sound reconstruction," *The IEICE Transactions on Information and Systems (Japanese Edition)*, vol. 83, no. 2, pp. 468–477, 2000.

[16] M. Wu, D. W., and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[17] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[18] A. de Cheveigné, *Multiple F0 estimation*, C. Wang and G. Brown, Eds. IEEE press, Wiley, 2006.

[19] A. Klapuri and M. Davy, *Signal processing methods for music transcription*. Springer Science & Business Media, 2007.

[20] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[21] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2003, pp. 177–180.

[22] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[23] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[24] I. S. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with bregman divergences," in *Proceedings of Advances in Neural Infomation Processing Systems*, 2005, pp. 283–290.

[25] M. Nakano, J. Le Roux, H. Kameoka, Y. Kitano, N. Ono, and S. Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," in *Proceedings of Latent Variable Analysis and Signal Separation*, 2010, pp. 149–156.

[26] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3437–3440.

[27] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proceedings of International Conference Independent Component Analysis and Blind Signal Separation*, 2004, pp. 494–499.

[28] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[29] A. Ozerov, C. Feévotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 121–124.

[30] M. Nakano, J. Le Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama, "Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden markov model," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 325–328.

[31] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[32] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proceedings of International Conference on Music Infomation Retrieval*, 2007, pp. 381–386.

[33] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Mar. 2008, pp. 109–112.

[34] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Mar. 2008, pp. 1825–1828.

[35] A. Ozerov, C. Févotte, R. Blouet, and J. L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2011, pp. 257–260.

[36] T. Nakamura, K. Shikata, N. Takamune, and H. Kameoka, "Harmonic-temporal factor decomposition incorporating music prior information for informed monaural source separation," in *Proceedings of International Symposium Music Information Retrieval*, 2014, pp. 623–628.

[37] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted non-negative matrix factorisation for sound source separation," in *IEEE/SP 13th Workshop on Statistical Signal Processing*. IEEE, 2005, pp. 1132–1137.

[38] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Mar. 2008, pp. 2069–2072.

[39] B. Fuentes, R. Badeau, and G. Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 21, no. 9, Sep. 2013.

[40] R. Jaiswal, D. FitzGerald, E. Coyle, and S. Rickard, "Towards shifted NMF for improved monaural separation," in *Proc. 24th IET Irish Signals and Systems Conference*, Jul. 2013.

[41] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2007, pp. I–53–I–56.

[42] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 45–48.

[43] J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Canadas-Quesada, "Musical instrument sound multi-excitation model for non-negative spectrogram," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1144–1158, Oct. 2011.

[44] H. Kirchhoff, S. Dixon, and A. Klapuri, "Missing template estimation for user-assisted music transcription," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 26–30.

[45] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Latent Variable Analysis and Signal Separation*, 2010, pp. 140–148.

[46] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method," *Journal of the Acoustical Society of America*, vol. 92, no. 3, pp. 1394–1402, 1992.

[47] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama, "Specmurt analysis of polyphonic music signals," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 16, no. 3, pp. 639–650, Mar. 2008.

[48] [Online]. Available: http://www.music-ir.org/mirex/wiki/MIREX_HOME

[49] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Proc. Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*, 2006.

[50] H. Kameoka and K. Kashino, "Composite autoregressive system for sparse source-filter representation of speech," in *IEEE International Symposium on Circuits and Systems*, 2009, pp. 2477–2480.

[51] F. Itakura and S. Saito, "Analysis-synthesis telepohny based upon the maximum likelihood method," in *Proceedings of International Congress on Acoustics*, 1968, pp. C17–20.

[52] F. Itakura, S. Saito, T. Koike, H. Sawabe, and M. Nishikawa, "An audio response unit based on partial autocorrelation," *IEEE Transactions on Communications*, vol. 20, no. 4, pp. 792–797, Aug 1972.

[53] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[54] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, Feb. 1991.

[55] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[56] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," in *Proceedings of the Stockholm Music Acoustics Conference*, 2013, pp. 287–292.

[57] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[58] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction." in *Proceedings of ISCA*

Tutorial and Research Workshops on Statistical and Perceptual Audition, 2008, pp. 23–28.

[59] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proceedings of International Conference on Digital Audio Effects*, Sep. 2010, pp. 397–403.

[60] T. Irino and H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3549–3554, 1993.

[61] H. Kameoka, T. Tahara, T. Nishimoto, and S. Shigeki, "Signal processing method and device," Nov. 2008, japan Patent JP2008-281898.

[62] T. Nakamura and H. Kameoka, "Fast signal reconstruction from magnitude spectrogram of continuous wavelet transform based on spectrogram consistency," in *Proceedings of International Conference on Digital Audio Effects*, 2014, pp. 129–135.

[63] N. Holighaus, M. Dorfler, G. Velasco, and T. Grill, "A framework for invertible, real-time constant-Q transforms," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 21, no. 4, pp. 775–785, Apr. 2013.

[64] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörlfer, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Proceedings of AES International Conference on Semantic Audio*, Jan. 2014.

[65] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of International Conference on Digital Audio Effects*, 2010.

[66] H. Tachibana, N. Ono, H. Kameoka, and S. Sagayama, "Harmonic/percussive sound separation based on anisotropic smoothness of spectrograms," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2059–2073, 2014.

[67] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals," in *Proceedings of International Symposium Music Information Retrieval*, 2014, pp. 611–616.

[68] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, pp. 47–61, 1940.

[69] R. D. Patterson, "Auditory filter shape," *Journal of the Acoustical Society of America*, vol. 55, no. 4, pp. 802–809, 2005.

[70] (2016, Feb.). [Online]. Available: http://www.music-ir.org/mirex/wiki/2014:Multiple_Fundamental_Frequency_Estimation_%26_Tracking_Results

[71] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. Society for Industorial and Applied Mathematics, 1970.

[72] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 60–77, 2000.

[73] H. Kameoka, "Statistical speech spectrum model incorporating all-pole vocal tract model and f_0 contour generating process model," in *IEICE Technical Report*, vol. 110, no. 297, Nov. 2010, pp. 29–34, in Japanese.

[74] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds., 2006, pp. 267–296.

[75] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[76] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[77] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.

[78] G. Hu and D. L. Wang, "An auditory scene analysis approach to monaural speech segregation," *Topics in Acoustic Echo and Noise Control*, pp. 485–515, 2006.

[79] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 69–72.

[80] U. Simsekli and A. T. Cemgil, "Score guided musical source separation using generalized coupled tensor factorization," in *Proceedings of European Signal Processing Conference*. IEEE, 2012, pp. 2639–2643.

[81] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[82] K. Yoshii and M. Goto, "Infinite latent harmonic allocation: A nonparametric bayesian approach to multipitch analysis," in *Proceedings of International Society Music Information Retrieval*, 2010, pp. 309–314.

[83] D. Sakaue, T. Otsuka, K. Itoyama, and H. G. Okuno, "Bayesian nonnegative harmonic-temporal factorization and its application to multipitch analysis," in *Proceedings of International Society Music Information Retrieval*, Oct. 2012, pp. 91–96.

[84] R. Hennequin, R. Badeau, and B. David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," in *Proceedings of International Conference on Digital Audio Effects*, 2010, pp. 246–253.

[85] H. Kameoka, J. Le Roux, Y. Ohishi, and K. Kashino, "Music Factorizer: A note-by-note editing interface for music waveforms," in *Special Interest Group Technical Reports of IPSJ*, vol. 2009-MUS-81, no. 9, Jul. 2009, in Japanese.

[86] (2016, Feb.). [Online]. Available: http://www.fluidsynth.org/

[87] K. Yoshii, K. Itoyama, and M. Goto, "Infinite superimposed discrete all-pole modeling for source-filter decomposition of wavelet spectrograms," in *Proceedings of International Symposium Music Information Retrieval*, 2015, pp. 86–92.

[88] R. Hennequin, R. Badeau, and B. Davide, "NMF with time-frequency activations model nonstationary audio events," *IEEE Transactions on Acoustics, Speech, and Language Processing*, pp. 744–753, 2011.

[89] R. Badeau and A. Ozerov, "Multiplicative updates for modeling mixtures of non-stationary signals in the time-frequency domain," in *Proceedings of European Signal Processing Conference*, 2013, pp. 1–5.

[90] R. Badeau, N. Bertin, and E. Vincent, "Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, vol. 21, no. 12, pp. 1869–1881, 2010.

[91] D. M. Lopes and P. R. White, "Signal reconstruction from the magnitude or phase of a generalised wavelet transform," in *Proceedings of European Signal Processing Conference*, 2000, pp. 2029–2032.

[92] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, 1998.

[93] H. Kameoka, M. Goto, and S. Sagayama, "Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelops," in *Proceedings of SIG Technical Reports on Music and Computer of IPSJ*, vol. 2006-MUS-66, no. 13, Aug. 2006, pp. 77–84, in Japanese.

[94] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 345–356, 2006.

[95] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[96] "ITU-T recommendation BS.1387-1, Perceptual evaluation of audio quality (PEAQ): Method for objective measurements of perceived audio quality," Sep. 2001.

[97] (2016, Feb.). [Online]. Available: http://www-mmsp.ece.mcgill.ca/Documents/Software/Packages/AFsp/AFsp.html

[98] "ITU-T recommendation P.862, Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[99] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Transactions on Knowledge Data Engineering*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.

[100] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proceedings of Advances in Neural Infomation Processing Systems*, 2009, pp. 2080–2088.

[101] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 57–60.

[102] P. O. Hoyer, "Non-negative sparse coding," in *IEEE Workshop Neural Networks for Signal Processing*, 2002, pp. 557–565.

[103] B. Shen, L. Si, R. Ji, and B.-D. Liu, "Robust nonnegative matrix factorization via l1 norm regularization by multiplicative updating rules," in *IEEE International Conference on Image Processing*, Oct. 2014, pp. 5282–5286.

[104] M. Suzuki, T. Hosoya, A. Ito, and S. Makino, "Music information retrieval from a singing voice using lyrics and melody information," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 151–151, 2007.

[105] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 4, pp. 1–7, 2010.

[106] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proceedings of International Symposium Music Information Retrieval*, 2005, pp. 329–336.

[107] M. Ryynänen, M. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," in *Proceedings of IEEE International Conference on Multimedia and eExpo*, 2008, pp. 1417–1420.

[108] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[109] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 221–224.

[110] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 90–93.

[111] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 228–237, 2014.

[112] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitch-based methods for background and melody separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1884–1893, 2014.

[113] (2016, Feb.). [Online]. Available: http://bass-db.gforge.inria.fr/bss_eval/

[114] (2016, Feb.). [Online]. Available: https://sites.google.com/site/unvoicedsoundseparation/mir-1k

[115] M. N. S. Swamy and K. S. Thyagarajan, "Digital bandpass and bandstop filters with variable center frequency and bandwidth," *Proceedings of IEEE*, vol. 64, no. 11, pp. 1632–1634, 1976.

[116] S. Erfani and B. Peikari, "Variable cut-off digital ladder filters," *International Journal on Electron*, vol. 45, no. 5, pp. 535–549, 1978.

[117] E. C. Tan, "Variable lowpass wave-digital filters," *Electronics Letters*, vol. 18, pp. 324–326, 1982.

[118] P. A. Regalia and S. K. Mitra, "Tunable digital frequency response equalization filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 1, pp. 118–120, 1987.

[119] S. J. Orfanidis, "Digital parametric equalizer design with prescribed nyquist-frequency gain," *Journal of Audio Engineering Society*, vol. 45, no. 6, pp. 444–455, 1997.

[120] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2007, pp. I–57–I–60.

[121] N. Yasuraoka, T. Abe, K. Itoyama, T. Takahashi, T. Ogata, and H. G. Okuno, "Changing timbre and phrase in existing musical performances as you like: manipulations of single part using harmonic and inharmonic models," in *Proceedings of the ACM International Conference on Multimedia*, 2009, pp. 203–212.

[122] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Drumix: An audio player with real-time drum-part rearrangement functions for active music listening," *Transactions on Information Processing Society of Japan*, vol. 48, no. 3, pp. 1229–1239, 2007.

[123] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative evaluation of multiple harmonic/percussive sound separation techniques based on anisotropic smoothness of spectrogram," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 465–468.

[124] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proceedings of International Congress on Acoustics*, 1968, c-17–C-20.

[125] H. S. Seung and D. D. Lee, "Algorithms for non-negative matrix factorization," in *Proceedings of Advances in Neural Infomation Processing Systems*, vol. 13, 2001, pp. 556–562.

[126] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transaction on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[127] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[128] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal variability of melodic source," in

Proceedings of International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 425–428.

Appendix A

Additional Experimental Results of Low-Rankness of Spectrograms

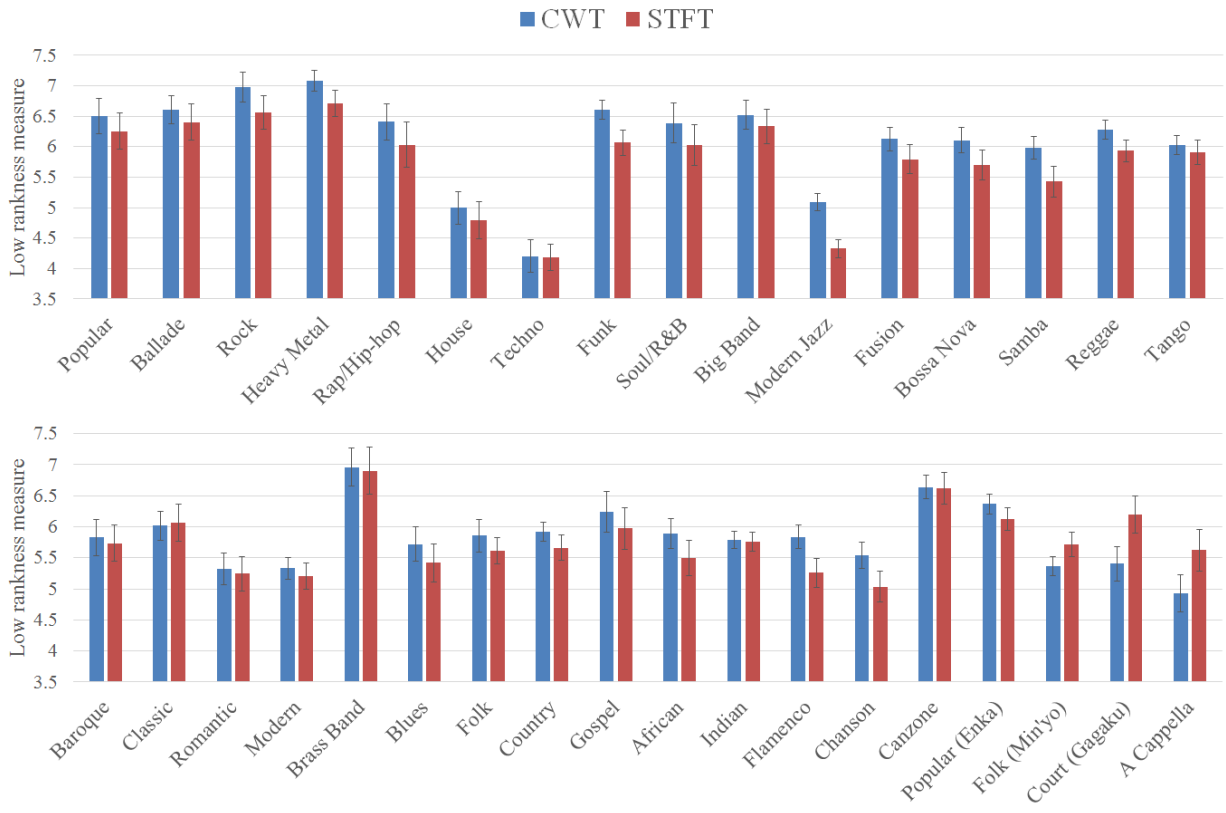


Figure A.1: Comparison of STFT and CWT spectrograms in low-rankness for subcategories of music genre.

List of Publications

Journal Paper

[J1] Tomohiko Nakamura and Hirokazu Kameoka, "Fast Signal Reconstruction from Magnitude Spectrogram of Constant-Q Transform," *IEEE/ACM Transaction on Audio, Speech and Language Processing*, in preparation to resubmission, 2015. (Chapter 6.)

Peer-Reviewed International Conferences

[I1] Tomohiko Nakamura and Hirokazu Kameoka, "Shifted and Convolutive Source-Filter Non-Negative Matrix Factorization for Monaural Audio Source Separation," *Proceedings of International Conference on Audio, Speech and Signal Processing*, to appear, Mar. 2016. (Chapter 3.)

[I2] Tomohiko Nakamura and Hirokazu Kameoka, "L_p-Norm Non-Negative Matrix Factorization and Its Application to Singing Voice Enhancement," *Proceedings of International Conference on Audio, Speech and Signal Processing*, pp. 2115–2119, Apr. 2015. (Chapter 7.)

[I3] Tomohiko Nakamura and Hirokazu Kameoka, "Harmonic-Temporal Factor Decomposition Incorporating Music Prior Information for Informed Monaural Source Separation," *Proceedings of International Society for Music Information Retrieval Conference*, pp. 623–628, Oct. 2014. (Chapter 4 and 5.)

[I4] Tomohiko Nakamura and Hirokazu Kameoka, "Fast Signal Reconstruction from Magnitude Spectrogram of Continuous Wavelet Transform based on Spectrogram Consistency," *Proceedings of International Conference on Digital Audio Effects*, pp. 129–135, Sep. 2014. (Chapter 6.)

[I5] Tomohiko Nakamura, Hirokazu Kameoka, Kazuyoshi Yoshii and Masataka Goto, "Timbre Replacement of Harmonic and Drum components for Music Audio Signals," *Proceedings of International Conference on Audio, Speech and Signal Processing*, pp. 7520–7524, May. 2014. (Chapter 8.)

Domestic (Japanese) Conferences

[D1] 中村友彦, 亀岡弘和, "高速近似連続ウェーブレット変換による振幅スペクトログラムに対する実時間位相推定法," 日本音響学会春季研究発表会講演集, to appear, Mar. 2016.

[D2] 中村友彦, 亀岡弘和, "全極スペクトログラムモデルと擬似周期信号モデルのウェーブレット変換表現を用いた多重音スペクトログラムの調波時間因子分解," 情報処理学会研究報告, vol. 2015-MUS-107, no. 50, May 2015.

[D3] 中村友彦, 亀岡弘和, " 全極スペクトルモデルを用いた調波時間因子分解による多重音解析," 情報処理学会研究報告, vol. 2015-MUS-106, no. 26, Mar. 2015.

[D4] 中村友彦, 吉井和佳, 後藤真孝, 亀岡弘和, "音楽音響信号中の調波音の周波数特性およびドラムの音色の置換システム," 情報処理学会研究報告, vol. 2014-MUS-104, no. 11, Aug. 2014.

[D5] 中村友彦, 亀岡弘和, "無矛盾性規準に基づく連続ウェーブレット変換スペクトログラムへの位相推定法と高速化," 情報処理学会研究報告, vol. 2014-MUS-103, no. 41, May 2014.

[D6] 四方紘太郎, 高宗典玄, 中村友彦, 亀岡弘和, "調波時間因子分解法に基づく事前情報付き多重音解析," 情報処理学会研究報告, vol. 2014-MUS-103, no. 18, May 2014.

[D7] 中村友彦, 亀岡弘和, "連続ウェーブレット変換の高速近似アルゴリズムに基づく振幅スケーログラムへの無矛盾位相付加法の検討," 日本音響学会春季研究発表会講演集, pp. 745–746, Mar. 2014.

[D8] 中村友彦, 吉井和佳, 後藤真孝, 亀岡弘和, "音楽音響信号に含まれる調波音の周波数特性とドラムの音色の転写システム," 日本音響学会春季研究発表会講演集, pp. 1043–1044, Mar. 2014.

[D9] 四方紘太郎, 高宗典玄, 中村友彦, 亀岡弘和, "調波時間因子分解に基づく音楽事前情報付き多重音解析," 日本音響学会春季研究発表会講演集, pp. 1049–1052, Mar. 2014.

Presentation in Symposium (Japanese)

[S1] 中村友彦, 吉井和佳, 後藤真孝, 亀岡弘和, “音楽音響信号中の調波音の周波数特性およびドラムの音色の置換システム,” *OngaCREST* シンポジウム *2014*-音楽情報処理研究が切り拓く未来を探る-, Aug. 23, 2014.

Awards and Funds

[A1] 情報処理学会 2015 年度山下記念研究賞

[A2] 音学シンポジウム 2015 年ポスター賞

[A3] 日本音響学会 2014 年春季研究発表会学生優秀発表賞

[F1] 平成 27 年度日本学術振興会特別研究員 DC2（2015 年 4 月〜2016 年 3 月）

[F2] 立石科学技術振興財団　後期国際交流助成（2014 年 10 月）

[F3] 原総合知的通信システム基金　第 46 回国際論文発表者助成（2014 年 9 月）