Monaural Music Separation via Supervised Non-negative Matrix Factor with Side-information

by

Ce Peng, **M.A.Sc**

A dissertation submitted to the
Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy of Electrical: Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario
July, 2017

Abstract

In this dissertation, a supervised source template nonnegative matrix factorization (NMF) algorithm is proposed to solve the monaural music source separation problem. Different from the previous state-of-the-art algorithms, the basic theoretical concept of the proposed algorithm considers the spectrogram from an audio mixture as linear combinations of note templates. Having prior knowledge of these note templates for each source, we can estimate and determine the activities of each template in recordings to build a mask of each source. Through the masks, the audio of target tracks can be reconstructed.

We reviewed previous research on source separation for monaural music audio separation and compared these work with our proposed algorithm not only in mathematical expressions but also in separation performances. First, the prior knowledge of note templates is informed by musical instrument audio dataset. The spectrograms from these instruments are obtained and factored into a source resonance character matrix and a source impulse excitation matrix by assuming that the spectrum of the different notes are formed by the resonance effects from an impulse excitation. Secondary, according to the prior informed note templates, their onset-offset-like features are estimated by using the multiplicative update rule and supervised by the proposed pitch-checking algorithm to remove misleading estimations. Finally, the supervised note onset-offset-like features alternatively become a constraint to help the proposed model evolve its prior informed note templates into the forms given by the recorded instruments.

We employed the TRIOS and the Bach-10 dataset for our multi-source separation performance tests. Among the source separation algorithms, our proposed supervised source template NMF and the state-of-the-art algorithms including the sound-prism and the Oracle-toolbox methods were selected to make comparisons. Furthermore, we added white Gaussian noise into the audio mixture to simulate the background full of the random noise to test the noise characteristics of each algorithm. The experimental results SDR (signal to distortion ratio), SIR (signal to interferences ratio), and SAR (signal to artifacts ratio) indicate that with the note templates from side-information, the proposed supervised source template NMF algorithm can have equivalent or higher performance in two-source separation and have a better performance under noise.

This is dedicated to my wife and the deepest love, Ruoxi Du, who loves me, believes in me, inspires me and has supported me every step of the way.

"The fastest horse cannot reach ten steps at one leap, yet an ordinary horse can gallop a great distance by continuous trotting." I am the ordinary horse and cannot be still trotting toward to the destination without her steadfast love.

Acknowledgments

First of all, I wish to thank Prof. Richard Dansereau, my supervisor, for bringing me into the fantastic world of Artificial Intelligence and Machine Learning, continually giving me generous funding support, confidence and knowledge, and guiding me on this road.

Special thanks to Tony Wacheski and Sean Kormilo who are the CEO, Co-Founder and Senior Software Engineer at Anystone Inc., who have opened a gate for me by programming and coding our research findings into commercial software.

Finally, my sincere appreciation to my wife, especially for her great love and care which keeps me smiling even in the darkest time of illness.

Table of Contents

List of Tables

List of Figures

Nomenclature

Table of Symbols

Symbol name	Description
$V_{f,t}$	STFT spectrogram of the observed signal $s(t)$
$W_{f,\phi,d}$	note templates matrix
$\varOmega_{\hat{f},d}$	impulse excitation matrix
$T_{d,f,\phi,\hat{f}}$	resonance characteristic matrix
$H_{d,\phi,t}$	activation matrix
$\hat{V}_{d,f,t}$	rebuilt STFT of the source
$\hat{V}_{f,t}$	rebuilt STFT of the mixed sources
$G(\cdot\|\cdot)$	auxiliary function
$\mathcal{D}_{\mathcal{E}}(\cdot\|\|\cdot)$	cost or distance function
$\mathcal{W}_{f,d}$	basis vector tensor
$\mathcal{T}_{f,k,f}$	unit constant translation tensor
$\mathcal{H}_{d,k,t}$	coefficient vector tensor
d	amount of sources
ϕ	amount of potential components (notes)
f	frequency domain in Hz
$\hat{f}$	potential components domain
η	step-size in gradient descent algorithm

Chapter 1

Introduction

1.1 Problem Statement

The source separation problem is more difficult than its mixing process. A simplified example is that it is easy to mix salt in water, but hard to separate them. A similar problem also exists in music sound separation. However, unlike the separation of salt and water, a human's ears and brain can focus on the sound of the interested instrument and mask it from others. This characteristic is very useful in applications of signal separation and understood by using machine learning to find an efficient method for teaching the computer to get a solution.

The phenomenon of audio mixture separation by brain was first brought out by Colin Cherry in 1953 [13] as a problem faced by air traffic controllers and is well known now as the cocktail party problem, where the human auditory system can pay attention to a particular stimulus while filtering out a range of "other noises." In the early years, the solution was given by the application of classic filters such as high pass [14], low pass [15], band pass [16] or filter banks [17], which keep the attention on a particular frequency band and get rid of the range of other noises. Though filters work, in general, their performance encounters severe challenges when there are overlaps of useful signal and "noise" in the frequency domain. Lately, adaptive filter theory [18, 19] has been developed to point at the frequency overlap problem and pick up useful data, if the useful data is predictable or the "noise" can be depicted in a statistical way. But the filter theory would not work well when the useful data is unpredicted, or the reference signal is non-stationary or non-stochastic series. Compared with filter theory, the advantage of a machine learning technique is that it can build a model using the data of non-stationary or non-stochastic series, therefore a machine learning technique can become an alternative to classic filters and adaptive filters in getting useful data from "other noise."

To blend the sound of different instruments is as easy as blending salt and water together. But to separate and rebuild the audio of the targeted track or instrument from a monaural musical audio mixture is much harder because there are an infinite number of mathematical solutions. This problem can be described as how to work out a set of numbers, of which the sum is 9. In mathematics, the possible solutions could be $1+8$, $1-2i+8+2i$, $0.5+0.05+8.45$ and so on. For our needs, we have to consider several factors from prior settings. The first is the domain of solutions, e.g. negative, positive, real or imaginary, and the second is the number of required numbers, e.g. two or three. These prior settings are called constraints and depending on them, we will get different possible answers. The other metaphor of explaining how constraints work in source separation is to consider a fruit recognition between bananas, apples, tomatoes or pears. If we separate them according to shape, the banana is different from the others; if we separate them from the way they are cooked, the tomato is the only fruit normally cooked as a vegetable; and if we separate them from the taste, the four fruits have four unique flavours. The constraints such as the shape, colour, the preparation method and the taste determine the outputs of the separation. From a machine learning perspective, the most challenging problem in source separation is choosing the prior constraints to fit the needs of the separation task and determining if the coefficients of the prior constraints can be calculated in mathematics, or if the prior templates can be evolved to approach the actual ones which created the blended data.

Fortunately, the non-negative matrix factorization (NMF) algorithm [20, 21], an unsupervised classifier like the form of under-determined simultaneous equations, is applied in many areas and shows great potential value in our research. In its basic form, a non-negative matrix, a spectrogram of an observed musical audio mixture, can be factored into two non-negative underlying factors, which are most likely defined as the frequency profile templates in Hertz and their actions in time by a number of notes from a musical clip. These underlying factors are summarized as the note templates and their coefficients. Therefore, the multi-source separation work is turned into a problem of finding the instrument labelled note templates used in actual recordings and calculating their onset-offset-like features. The sub-spectrogram of the targeted track can be reconstructed based on the time combinations of the same instrument labelled note templates as the mask for filtering.

Though the templates play a crucial role to avoid getting lost in infinite solutions in the under-determined problem, how to obtain proper templates is still in the research phase. If we decide to establish the separation algorithm based on the source labelled prior template, there will be a sequence of questions, such as where we should find them, whether they have the same form with the actual recording templates, and if the prior templates can be evolved to approach the form of actual recorded ones. It is just like a scenario of teaching a person to know the sounds of one violin he has heard before and letting him recognize the sounds of other violins from audio mixture. This needs the flexibility of the templates, which supposes to be a constraint and an updatable factor at the same time.

1.2 Overview of Proposed Plan

For solving the problems mentioned above, we propose a plan to build up a new type of constraint and side-information to supervise monaural music source separation. The proposed model is divided into two parts: the note templates labelled by instrument and their onset-offset-like features describing how strongly these templates are activated in the observed music. For the note templates, we propose to use isolated note audio from a standard dataset as one of the side-information to build up the prior note templates which can be decomposed further into a resonance matrix of different notes and a column vector representing the impulse excitation. On the one hand, a resonance matrix has to be in the form of an identity matrix and this design constraint has the restriction to keep the note resonances of an instrument from one basis note. On the other hand, the resonance matrix should have the flexibility in the degrees-of-freedom of being updated in the separation to adopt the unique resonance characteristics of each note rather than being strictly defined by a single score for the entire instrument. From the perspective of onset-offset-like features, we propose the note range of the actual recordings and their fundamental frequency components as another side-information, which allows the features to be modified dynamically.

According to the degrees-of-freedom, we classify the constraints used in NMF research as strong, medium and weak. In previous work of shifted non-negative tensor factorization (Shifted NTF) [10, 22], 2-D non-negative matrix factorization (2-D

NMF) [5,6], and score informed non-negative tensor factorization [1,5,22], they perform the strong constraint with low level degrees-of-freedom on unit constant magnification frequency shifted operands on a basis note vector to translate other notes of an instrument, but in fact the frequency spectrum often varies from note to note. In our proposed plan, the resonance matrix has a medium constraint with a medium level degrees-of-freedom, which means the constraint can be self evolved from its original form to fit the separation work. In the work of sparse non-negative tensor factorization [23], J. Kim and H. Park applied a weak constraint with a high-level degrees-of-freedom of measuring the norms of the underlying factors. Whatever the case, building the constraints or side-information is aimed at adjusting the degrees-of-freedom to obtain high accuracy separation results. The research of instrument

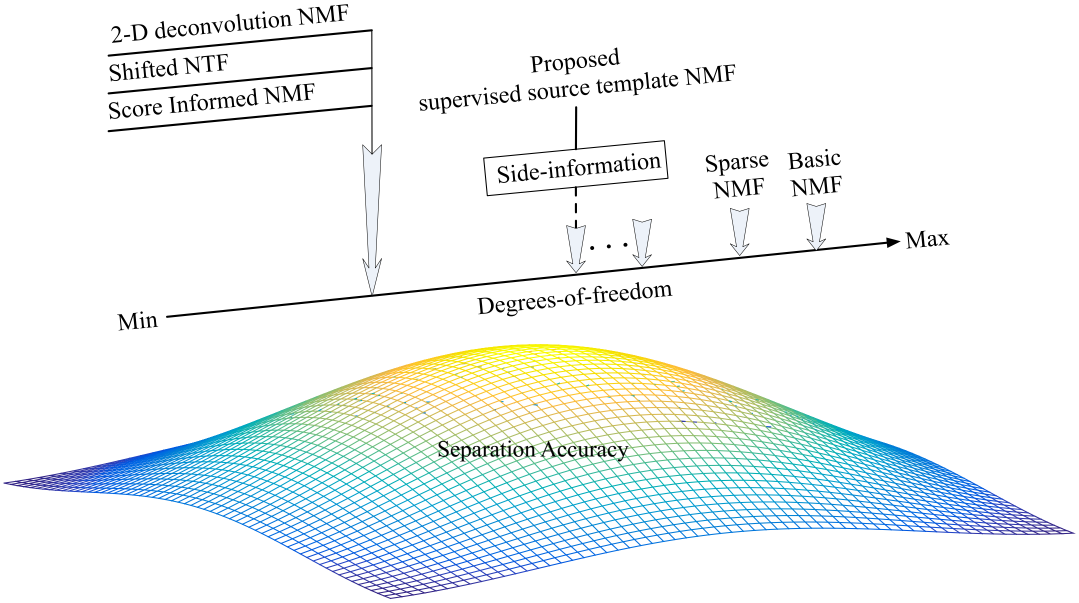


Figure 1.1: The technique position of proposed work.

separation based on NMF gives its contribution to find out the balance between the degree-of-freedom and separation accuracy. In Fig 1.1, we illustrate the positioning of this proposed work among state-of-the-art algorithms. Below the degrees-of-freedom axis, a quadratic surface is used to illustrate the separation accuracy. In the source separation problem, the underlying factors with too small or large degrees-of-freedom both cause poor accuracy in source separation. The degrees-of-freedom can be changed by forcing the constraint on factors, and there is an optimized point of degrees-of-freedom corresponding to the highest level of separation accuracy. The work of the shifted NTF, 2-D NMF and score informed NMF, and the work of sparse NMF and basic NMF are located on the sides of the small level and high level of degrees-of-freedom. Our task is to find suitable constraints by adjusting the degrees-of-freedom close to the optimized point with the help of side-information or prior knowledge.

1.3 Bulleted List of Contributions

- **Contribution 1** is the setup of the source labelled constraints. By our assumption the source labelled constraints include a note resonance characteristic part and a note impulse excitation part, which are initialized by a prior note templates through our proposed supervised learning algorithm in Section 3.3.1. This part of work explores the note resonance characteristic and the note impulse excitation information by instrument from a prior source labelled constraints.

- **Contribution 2** is the introduction of an updatable constraint of a note resonance characteristic part which is obtained from the supervised learning algorithm and responsible for the notes translation function in Section 3.4.3. In previous work [5, 6, 10, 22], the constraint of the notes translation from a basis note is a fixed identity matrix and non-updatable. Our proposed work breaks the restriction of the constraints from the previous work and enlarges the degrees-of-freedom of the constraints.

- **Contribution 3** is supervising the convergence process iteratively by our proposed pitch-checking algorithm in Section 3.4.2. This part of work is a necessary supplement for gradient descent [24, 25] based convergence algorithms e.g. multiplicative update rule [26–28]. Because the multiplicative update rule gives its optimized solutions following the constraints of gradient descent direction, which may not the optimized solutions for the source separation problem. The aim of this part of work is adding our side-information of pitch-checking as one of the constrains on convergence process in order to get a side-information oriented optimized solutions.

1.4 Proposal Organization

The organization of the proposal is listed by item structures.

- **Chapter 2** is a background introduction. In its first part, we present the general procedure of source separation based on the NMF algorithm. Then in its second part, we review the literature in the previous work of NMF, sparseness NMF [23], de-convolution NMF [1,5,22], score-informed NMF [29], shifted NTF [10,22] and so on.

- **Chapter 3** is mainly composed of four sub-topics. The first sub-topic talks about the motivation of our proposed work. The second sub-topic describes the basic expression of the proposed source template NMF. The third sub-topic mainly introduces the supervised learning work which aims at building the source labelled note templates and estimating their corresponding onset-offset-like features.

- **Chapter 4** designs the source separation performances test for the proposed source template NMF, sound-prism algorithm [2] and Oracle-toolbox algorithm [30] based on the TRIOS dataset [1] and the Bach-10 dataset [31]. It also tests the noise characteristics of all three algorithms on simulated audio mixtures which are created by adding the white noise as the background.

- **Chapter 5** makes conclusions and analysis of our proposed source template NMF algorithm. Furthermore, it discusses the advantages and disadvantages in the proposed algorithm and gives the potential development in the future.

Chapter 2

Research Background

This chapter introduces the work regarding monophonic sound source separation based on NMF methods which are related to our proposed work. The main topics include the general procedure of source separation and a review of previous work putting the emphasis on three branches: the development of constraints on the note basis vector matrix **W**, the innovation of constraints on the coefficient or weight vector matrix **H** and the growth of the method on convergence.

2.1 Audio Mixtures and Source Separation

Instrument separation is an important and challenging problem in computational analysis and music research [32,33]. Acoustic signal mixtures result, when several instruments, such as the saxophone, trumpet and violin, play together simultaneously shown in Fig. 2.1. Coccurring sounds from others instruments makes it difficult to estimate an individual instrument. The separation task would become easier depending on the spatial locations of recording microphone arrays. However, if the scenario does not have multi-channel recording, then the estimation task only involves source separation of monaural music signals [33]. This acoustic signal mixture is depicted in the time domain in the form of sound waveforms which may be played by the sum of musical notes in Fig. 2.2. In this figure, the music notes are arranged in a piano keyboard. We selected the notes from C, D, E, F, G, A, B and C_1 to indicate a completed range of an octave. Compared with the concept of audio mixtures in Fig. 2.1, the idea of source separation is addressed in a different point of view completely. It describes that the STFT of the audio mixtures is defined as V can be decomposed

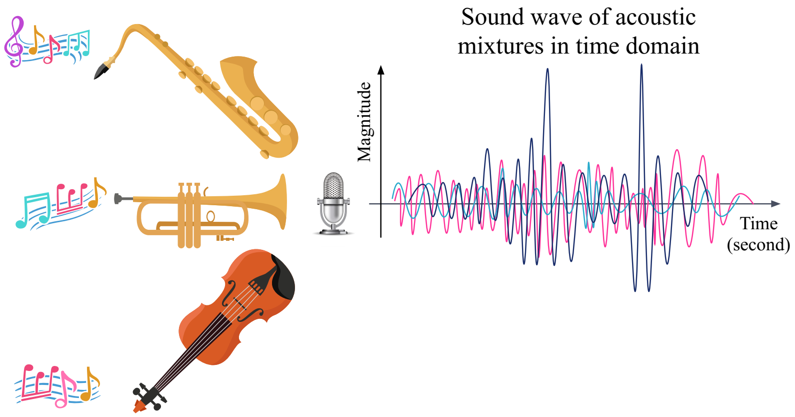


Figure 2.1: Monaural audio mixtures from several instruments

into the sum of the products of a note dictionary matrix W and a note activation matrix H by source. The number of sources and notes are side information provided by the prior knowledge. When the source defined dictionary matrix and activation matrix are obtained, their product will build up the STFT of a targeted channel. Through the inverse operand of STFT, the time domain signal of a targeted channel can be rebuilt as the separation solution. It has to be clear that all the separation procedure is called masking-based separation and the necessary phase information is obtained by the original spectrogram [1].

There are several ways such as mean opinion score (MOS) [34–36], perceptual evaluation of audio quality (PEAQ) [37–39] and signal-to-distortion ratio (SDR) [40, 41] to evaluate the quality of sound. MOS method is a measurement in telecommunication engineering domain. Though it is commonly used for measuring audio quality, it needs a large number of people to give a rating in scaled numbers between 1 and 5. Furthermore it also takes a long time to finish the rating by a large group of people. PEAQ measurements are commonly used in perception coding, which may introduce different levels of noise into a signal. But for separation operands, the main difference between the reconstructed signal and the target signal are the interference from other sources and the distortion from the separation algorithm. So in this thesis, we employed the SDR as a measure to outline the quality. The definitions of SDR and the feature set are introduced in Chapter 3.

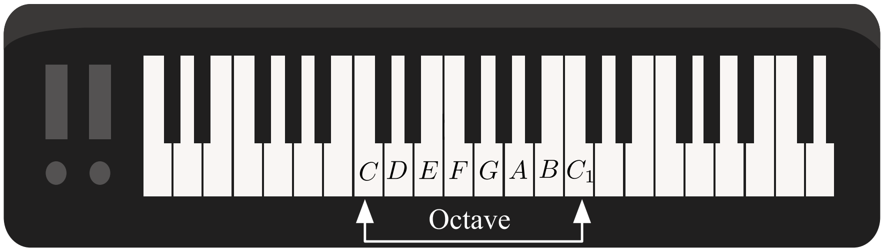


Figure 2.2: Music note and its octave relation

2.2 Procedure of Source Separation Based on NMF Algorithm

In the 1990s, the beginning work of NMF [20, 21, 42] was created by Paatero, Tapper, Lee, and Seung. NMF has widespread application in such domains as chemo metrics [43, 44], audio and image signal processing [45–47], data clustering, [48–50] and computer vision [48, 51]. But the concentration of this proposal only focuses on its application in side information aided monophonic sound source separation (MSSS). NMF is a way to factor any non-negative matrix, vector or data **V** into two matrices with no negative elements: matrix **W** which has linear combination basis vectors and matrix **H** which contains the weights of corresponding hidden components that make the contribution of each basis vector [52]. The NMF basic sketch is given in Fig 2.4, which makes the resulting matrices easily understood, especially in the factorization process. The basic expression for NMF is

$$\begin{aligned}\mathbf{V} &\approx \mathbf{W}\cdot\mathbf{H}\\ &= \sum_k W_{f,k}\cdot H_{k,t},\end{aligned} \tag{2.1}$$

where **V** is a non-negative matrix with size of $F \times T$ and its grayscale indicates the data intensity. **W** is a collection of non-negative basis vectors formed in an $F \times K$ matrix, in which F is the domain of non-negative basis components and K is the number of these potential bases. H is a collection of row vectors which represent the basis activation sequences in a $K \times T$ matrix. The numbers within **H** determine the amplitude of an activation process and the subscript t represents the length of the activation sequence.

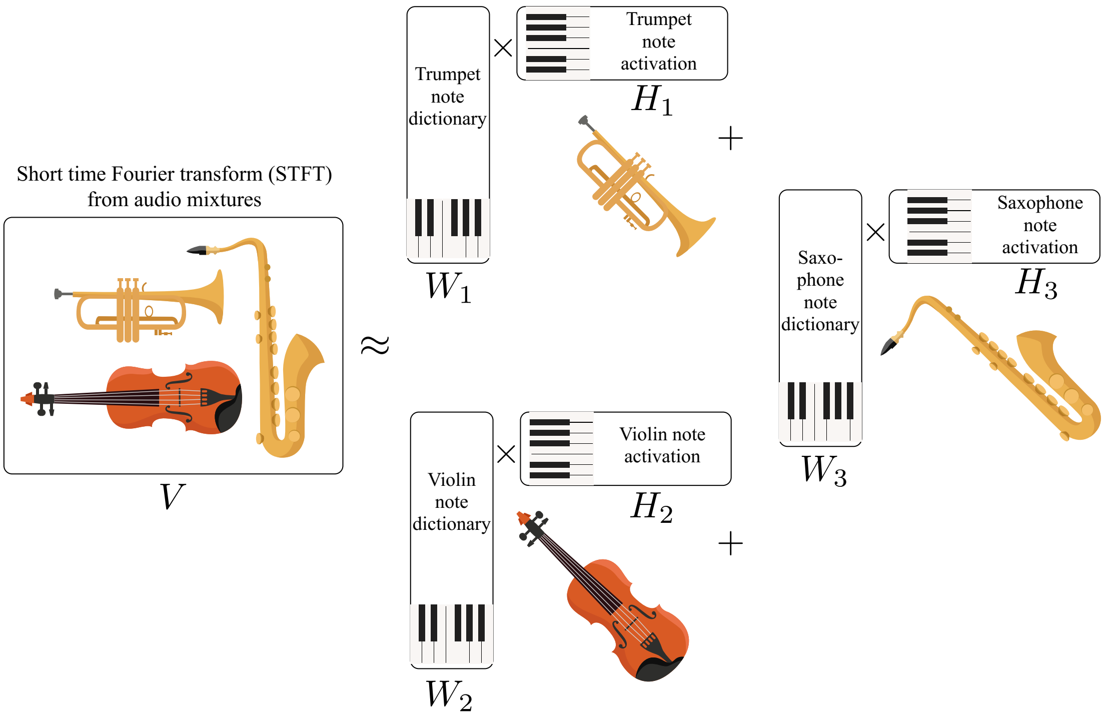


Figure 2.3: Source separation illustration

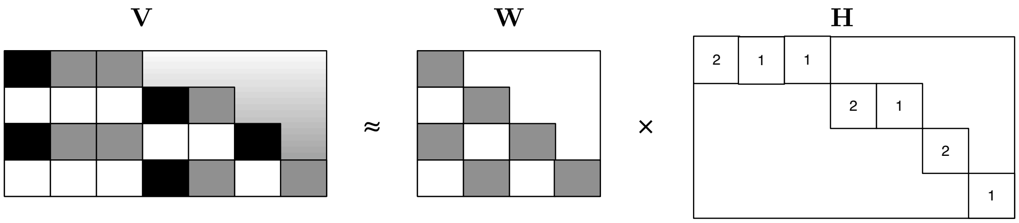


Figure 2.4: Basic sketch of NMF

For the instrument separation scenario in Fig. 2.5, **V** is the short time Fourier transform (STFT) of the observed signal. After several iteration, the **V** is factored into the note basis matrix **W** and its corresponding weight matrix **H** with non-negative values. The frequency bin is related to the notation f in Hz. The label t is related to time in seconds and k is related to the potential number of notes. In this simple example in Fig. 2.5, K is supposed to be five notes which are determined by prior knowledge. Matrix **W** contains the note frequency profiles in column vectors. Matrix **H** is their related activations described as attack and decay profiles in the row vectors of time. For the instrument separation scenario in Fig. 2.6, each source STFT

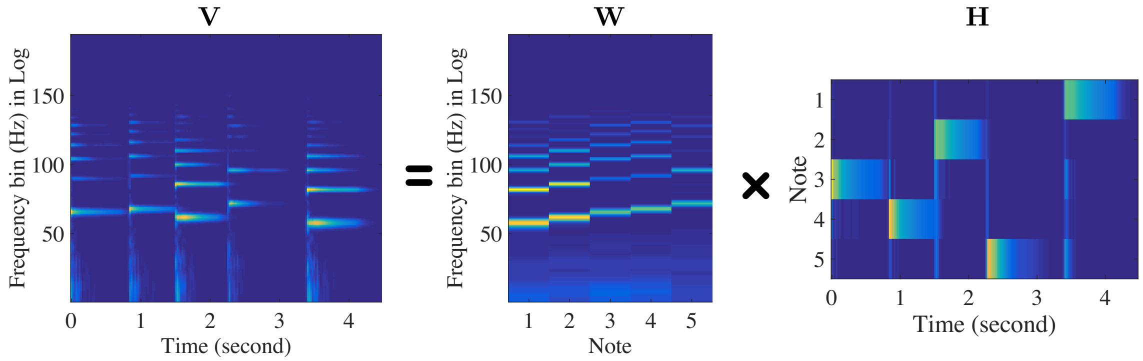
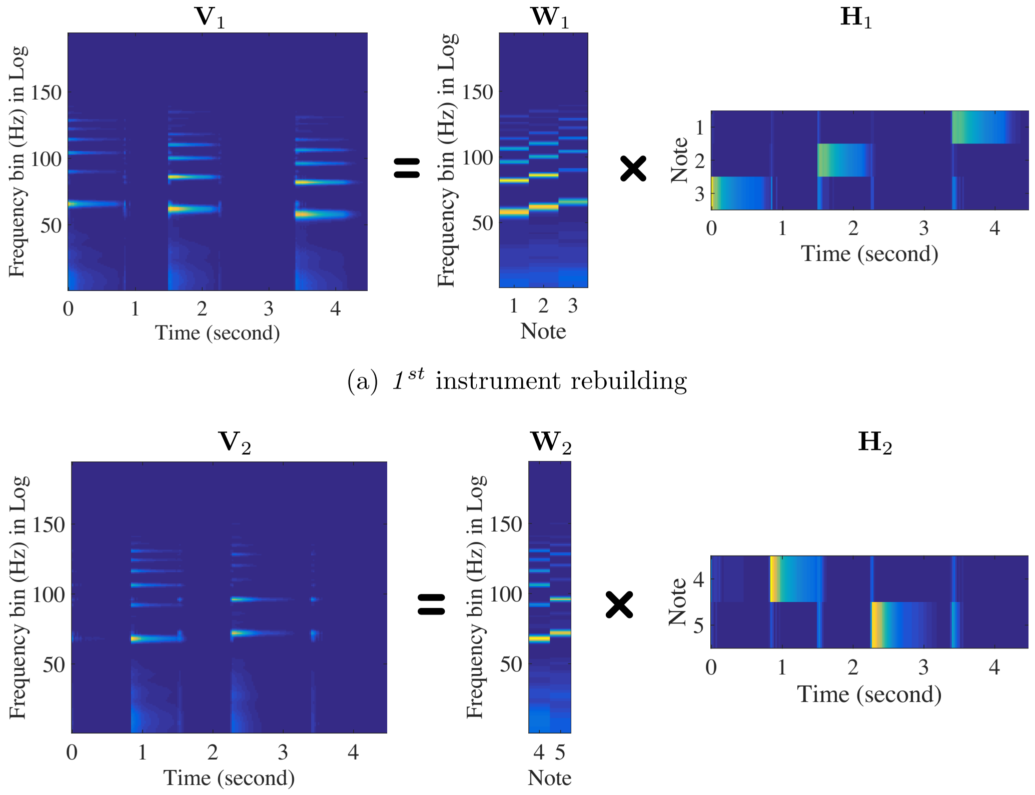


Figure 2.5: STFT non-negative matrix factorization

is rebuilt by masking **W** and **H**. For example in Fig. 2.6(a) and 2.6(b), if informed by side information that notes 1, 2, and 3 are played by the *1*st instrument and notes 4 and 5 belong to the *2*nd instrument, the rebuilt STFT of the *1*st source can be obtained as $\mathbf{V}_1$ by multiplying the matrices $\mathbf{W}_1$ and $\mathbf{H}_1$ whose components of notes, 4 and 5, are removed in Fig. 2.6(a). In the same way, the rebuilt STFT of the *2*nd source are worked out as $\mathbf{V}_2$ in Fig. 2.6(b). Obviously, we have $\mathbf{V} \approx \mathbf{V}_1 + \mathbf{V}_2$ in the frequency domain and their time domain signals, $s_1(t)$ and $s_2(t)$, can be estimated through an inverse STFT. For the simple condition, the factored $\mathbf{H}_1$ and $\mathbf{H}_2$ still contain the note triggered and decay information from another source.

2.3 Literature Review

To be a tool for MSSS, NMF has an ability for data separating and clustering [53–55]. The main research is that among the many solutions in non-convex space, is how to avoid being trapped in a local minimum [56] that does not correspond to a realistic source separation. In other words, if matrices $\hat{\mathbf{W}}$ and $\hat{\mathbf{H}}$ are particular minimum solutions, the general solutions can be given by countless pairs of solutions $\hat{\mathbf{W}} = \mathbf{W}\mathbf{C}$ and $\hat{\mathbf{H}} = \mathbf{C}^{-1}\mathbf{H}$ for any non-negative invertible $\mathbf{C}$. To improve the uniqueness convergence, a constraint in the form of sparsity has proven to be useful and merged with other different levels of constraints together for optimized searching guidance [57]. The following sections discuss a few common constraints.



(a) *1*st instrument rebuilding

(b) *2*nd instrument rebuilding

Figure 2.6: Instrument separation based on NMF

2.3.1 Constraint on W and H

In the work [23,52,58–61], the sparseness was measured and employed within the cost function. The generalized error residual of sparse NMF becomes

$$\mathcal{E}(e) = \boldsymbol{D}(\mathbf{V}|| \sum_{k} W_{f,k} \cdot H_{k,t}) + \eta_w ||\mathbf{W}||_p + \eta_h ||\mathbf{H}||_p, \tag{2.2}$$

where $\boldsymbol{D}(\cdot||\cdot)$ would be any distance function such as Euclidean distance, Kullback−Leibler distance or $\beta-$divergence distance. η_w and η_h are Lagrange multipliers and p represents the L_p-norm of a matrix. According to the definition of sparsity, these regularizer constraints relate to the work of L_0-norm measurement [62]. But in the work of Tao et al. [63], under the restricted isometry property condition, the solutions from L_0-norm are proved to equal the solutions from L_1-norm. So these regularizer measurements mainly emphasize L_1-norm.

Furthermore, the work of complex NMF [64] by Kameoka, Ono, Kashino and

Sagayama is denoted in the complex-spectrum domain to recover the phase for the separated signals also with regularizer sparsity constraints on the code matrix **H** to guide to a unique iterated solution. The form of its cost function is

$$\mathcal{E}(e) = \sum_{f,t} | V_{f,t} - \hat{V}_{f,t}|^2 + 2\eta \sum_{k,t} ||H_{k,t}||_p, \tag{2.3}$$

where

$$\begin{aligned} \hat{V}_{f,t} &= \sum_{d} \hat{V}_{d,f,t} \\ &= \sum_{d} \sum_{k_d} W_{f,k_d} \cdot H_{k_d,t} \cdot e^{j\phi_{k_d,f,t}}, \end{aligned} \tag{2.4}$$

and η is a Lagrange multiplier and p refers to L_p-norm. d is defined as the source identification and k_d refers to the k^{th} component which belongs to the d^{th} source. This method was improved by Brian King's work [65, 66] with the constraint of complex probabilistic latent component analysis. With a similar idea of forcing the constraint on $\boldsymbol{W}$, in Hualiang Li's work [67, 68] an orthogonality measure is defined and orthogonality constraints are forced upon the standard NMF to get a flexible subsequent detection. In the work of Seungjin [69], the orthogonality constraints are imposed on the basis or encoding matrix.

2.3.2 Shift Constraints on W

If the L_p-norm regularizer can be accepted as a kind of external constraint, the relation among the basis vectors in W can be treated as a kind of inner bound constraint. In 2006, Mikkel and Morten [5, 6] developed the NMF model into 2-D deconvolution NMF which extends the matrix factorization into tensor factorization with the parameters τ and ϕ which indicate shifted operations along the time and frequency domain. So the original NMF was modified as

$$\begin{aligned} \mathbf{V} \approx \hat{\mathbf{V}} &= \sum_{\tau,\phi} \overset{\downarrow\phi}{\mathbf{W}}{}^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi} \\ \hat{V}_{\omega,t} &= \sum_{\tau,\phi,d} \mathbf{W}^{\tau}_{\omega-\phi,d} \mathbf{H}^{\phi}_{d,t-\tau}, \end{aligned} \tag{2.5}$$

and the error residual is written as

$$\mathcal{E}(e) = \boldsymbol{D}\left(\mathbf{V}||\hat{\mathbf{V}}\right) + \eta f(\mathbf{H}), \tag{2.6}$$

where $f(\mathbf{H})$ is a sparsity term with a positive derivative. Additionally, Paris also set a deconvolution rule [7, 8] to the underlying factor code matrix. Similarly, Cichocki used this method in 3-D NTF [70]. In the work of Fitzgerald et al. [9], their method is based on a simplified approximation of the real condition to overcome the problem of dealing with multiple notes belonging to a single source. This assumes that these notes are translated versions of a single frequency basis function. They developed their theory in [10, 22, 71] in 2011 and 2013 with the constant Q technique in order to limit the shifted step size to a semi or quarter note. The matrix operand in factorization is replaced by a tensor operand. So the expression of NTF is

$$\mathbf{V} \approx \langle\langle\mathcal{T}\mathcal{W}\rangle_{(3,1)}\mathcal{H}\rangle_{(2:3,1:2)}, \tag{2.7}$$

where $\mathcal{T}$ is a translation tensor containing the shift rules of each basis tensor $\mathcal{W}$. To strengthen the sparseness in the solution by a group sparsity constraint, the error residual is regarded as

$$\mathcal{E}(e) = \min_{\langle\mathcal{R}\mathcal{D}\rangle_{(3,1)},\mathcal{H}\geq 0}\left(\boldsymbol{D}(\mathbf{V}||\langle\langle\mathcal{T}\mathcal{W}\rangle_{(3,1)}\mathcal{H}\rangle_{(2:3,1:2)}) + \eta\Phi(\boldsymbol{\mathcal{H}})\right), \tag{2.8}$$

where the term $\Phi(\mathcal{H})$ is defined as the group sparsity constraint. For all shift constrained NMF and NTF, the observed STFT $\mathbf{V}$ must be transferred into $log(f)$ or the constant Q scale to force the minimum shift step size of a basis vector to equal a semitone. The work of 2-D deconvolution NMF, Kirbiz and Gunsel [72] proposed to improve the perceptual qualities of separated STFT sources by means of a defined perceptual evaluation of audio quality (PEAQ) auditory model [73].

Based on the source-filter theory, Tomohiko and Hirokazu [74] 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 to reduce the separation error caused by using a shifted copy of a spectrum template to represent the spectra of different fundamental frequency F_0s. They developed the shifted NMF model in the form of

$$X_{l,m} = \sum_{k,r}\sum_{p,\tau} F_{r,l,\tau}S_{k,l-p}U_{k,r,p,m-\tau}, \tag{2.9}$$

where $F_{r,l,\tau}$ is the filter spectrum templates related to the number r, the indices of log-frequency l and the time shift index τ. Also, m indicates the indices of time; k denotes the source excitation and p is an element within the set of possible frequency shifts.

2.3.3 Score Informed Constraints on H

The use of constraints on **W** may help to reduce the problem of partial interference, but this situation is going to remain in **H** when applying the algorithm in separation. So in the work of R.Hennequin [1], J. Fritsch [75], and S.Ewert [29, 76] , the synchronized MIDI information is taken out as a constraint on code matrix **H** to restrict the note attack and decayed profile in time. Their expression is in the form of

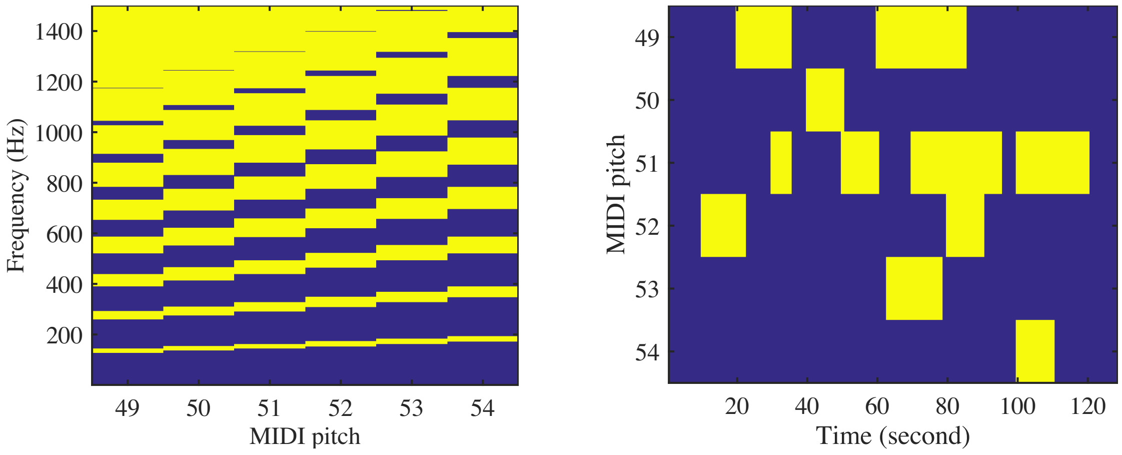
$$\mathbf{V} \approx (\mathbf{W} \odot \mathbf{W}_{tmp}) \cdot (\mathbf{H} \odot \mathbf{H}_{tmp}), \tag{2.10}$$

where $\odot$ represents elements-wise multiplication. $\mathbf{W}_{tmp}$ and $\mathbf{H}_{tmp}$ are defined by prior information from the note spectrogram envelops or synchronized MIDI files. In Fig. 2.7, the $\mathbf{W}_{tmp}$ and $\mathbf{H}_{tmp}$ are given in simple examples. The values in lightened places are set to 1 while in dark are 0. More exactly, the area of the n^{th} harmonic pitch p corresponds to the frequency range $(n \cdot f(p-\phi), n \cdot f(p+\phi))$, where ϕ is a parameter in semitone to control the size of these areas lighted in Fig. 2.7(a) and selected to 1 [76]. Here, $f : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ is defined by

$$f(p) := 2^{(p-69)/12} \cdot 440. \tag{2.11}$$

Details of Fig. 2.7(b) are provided by the synchronized MIDI file which are manually aligned one by one with Sonic Visualizer [75, 77]. The matrix $\mathbf{H}_{tmp}$ is similar to a piano roll representation. The size of lightened places along the time axes represents the note attack and decay action. The initialization of **W** and **H** are randomly set to non-negative values.

Apart from using a synchronized MIDI file, Ning and Roger [78] described a method that aligns polyphonic audio recordings of music to symbolic score information in the standard MIDI file. For matching notes in a musical performance to the corresponding notes in a score, Pedro and Alex implemented a score-performance matching algorithm based on hidden Markov models (HMM) [79]. The other existing



(a) An example of $\mathbf{W}_{temp}$

(b) An example of $\mathbf{H}_{temp}$

Figure 2.7: Examples of $\mathbf{W}_{tmp}$ and $\mathbf{H}_{tmp}$ in score informed algorithm

audio-score alignment work includes [80, 81] using the dynamic time warping (DTW) method and [2, 82] using the hidden Markov model method. The model structure of SoundPrism [2] is illustrated in Fig. 2.8 where x_n is the note position in beats, v_n is the tempo and the n^{th} frame y_n is associated with a 2-D hidden state vector defined as

$$s_n = \begin{pmatrix} x_n \\ v_n \end{pmatrix}. \tag{2.12}$$

But v_n in their process model is based on the real-time music tempo estimation and y_n is associated with each audio frame in their observation model. The SoudPrism score follower is built on multi-pitch estimation work [31] which is a maximum-likelihood-based algorithm and trained on thousands of isolated musical chords.

2.4 Optimization Algorithm

When the p changes in L_p-norm regularizer constraints and the NMF or NTF model structure becomes more complicated, the multiplicative iteration rules, auxiliary function algorithm and EM algorithm are adopted to get the minimization of univariate functions.

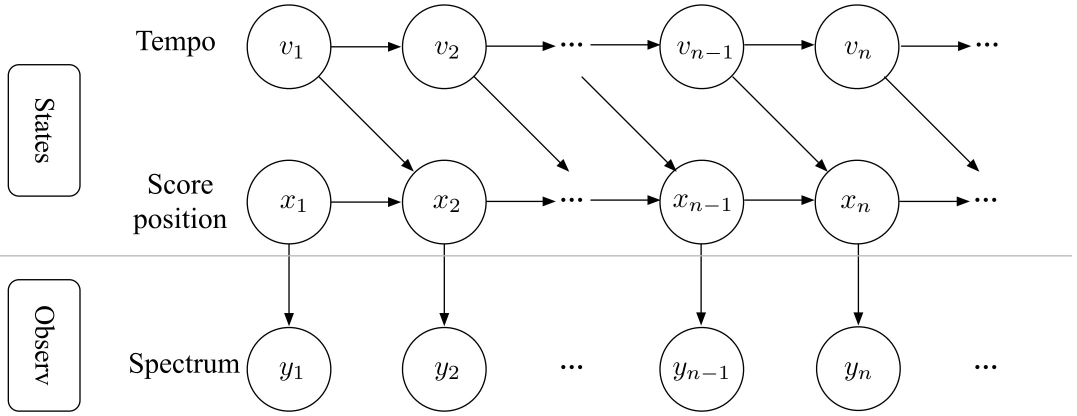


Figure 2.8: Illustration of the SoundPrism model structure [2]

2.4.1 Multiplicative Update Algorithm

The multiplicative update rules are developed from the gradient descent algorithm [83] which is often used to deal with the optimization of convex or concave functions such as the exponential function, *log* function and parts of power functions. In 2001, Lee provided the derivation for multiplicative iteration rules [21] for NMF and in 2007 Lin focused on the convergence of this algorithm [26]. For the classic NMF model, to choose a specific iteration step size, the iterated results can inherit the positive character from the initialization values. The expressions for multiplicative update are

$$\overset{new}{H}_{k,t} = \overset{old}{H}_{k,t} \frac{(\overset{old}{W}_{f,k})^T \cdot V_{f,t}}{(\overset{old}{W}_{f,k})^T \cdot (\overset{old}{W}_{f,k} \cdot \overset{old}{H}_{k,t})} \qquad \forall k, t, \tag{2.13}$$

and

$$\overset{new}{W}_{f,k} = \overset{old}{W}_{f,k} \frac{V_{f,t} \cdot (\overset{old}{H}_{k,t})^T}{(\overset{old}{W}_{f,k} \cdot \overset{old}{H}_{k,t}) \cdot (\overset{old}{H}_{k,t})^T} \qquad \forall f, k. \tag{2.14}$$

If the changes between the *new* and the *old* factors meet a stopping criterion the iteration loop is stopped. Given this work, Lin provided his projected gradient algorithm [84] for NMF optimization.

2.4.2 Auxiliary Function Iterative Algorithm

The multiplicative update is not the only approach in solving the iterative optimization problem of NMF; the auxiliary-function method [85] is an alternative method adopted in work [64, 74, 86, 87]. Meanwhile, the distance function $\mathcal{D}(\cdot)$ is extended into Kullback-Leibler divergence [88–90], Itakura-Saito divergence [91], the developing β-divergence [92–94], α-β-divergence [95] and α-β-γ-divergence [96]. With the parameters α, β or γ, the NMF optimization problem falls into a non-convex and non-smooth problem. The auxiliary function method was only discussed when the distance function is a convex or a concave function in different ranges of parameter values dependent on the values of α, β. Unlike the fast convergence speed of the multiplicative update rule, the auxiliary-function method evolves the underlying matrix in a slower but more stable way. In general, the auxiliary function is described as a delicate artificial function which complies with the following requirement for $\forall \boldsymbol{h}$ belonging to a $K-$dimensional non-negative real domain $\mathbb{R}_+^K$

$$F_e(\boldsymbol{h}^{i+1}) \leq G(\boldsymbol{h}^{i+1}|\boldsymbol{h}^i) \leq G(\boldsymbol{h}^i|\boldsymbol{h}^i) = F_e(\boldsymbol{h}^i), \tag{2.15}$$

where $F_e(\cdot)$ is a convex function and $G(\cdot|\cdot)$ is an auxiliary function. A simple example of their relation is given in Figure 2.9. As drawn, the value of function $G(\hat{\boldsymbol{h}}|\boldsymbol{h})$ is determined by two parameters $\hat{\boldsymbol{h}}$ and $\boldsymbol{h}$. If the i^{th} iteration of $\boldsymbol{h}$ is defined as $\boldsymbol{h}^i$, the only overlap between these functions is $G(\boldsymbol{h}^i|\boldsymbol{h}^i) = F_e(\boldsymbol{h}^i)$ and we also have $\hat{\boldsymbol{h}} = \boldsymbol{h}^i$. When $\hat{\boldsymbol{h}}$ moves toward the new point $\boldsymbol{h}^{i+1}$ to let the auxiliary function reach its minimum value, the value of $G(\boldsymbol{h}^{i+1}|\boldsymbol{h}^i)$ is smaller than $G(\boldsymbol{h}^i|\boldsymbol{h}^i)$ but still greater than $F_e(\boldsymbol{h}^i)$. Alternatively, the new point $\boldsymbol{h}^{i+1}$ is viewed as the beginning of the $(i+1)^{th}$ iteration and set to $\boldsymbol{h}$. Then we look for a new value of $\hat{\boldsymbol{h}}$ to let $G(\cdot|\boldsymbol{h}^{i+1})$ go toward its new minimum position. We repeat this operation n times while the point $\boldsymbol{h}^n$ moves closer to $\boldsymbol{h}^{min}$ step by step.

Thanks to the following lemma and referring to Jensen's inequality, we have a way to create an auxiliary function.

Lemma 1. Let F:$\mathbb{R} \mapsto \mathbb{R}$ be a convex function. If λ_k with $k \in \mathbb{Z}^+$ satisfies $\forall k$, $\lambda_k \geq 0$ and $\sum_k \lambda_k = 1$, then for $x_k \in \mathbb{R}$

$$F\left(\sum_k x_k\right) \leq \sum_k \lambda_k F\left(\frac{x_k}{\lambda_k}\right) \tag{2.16}$$

Equality occurs when $\lambda_k = x_k / \sum_k x_k$.

Lemma 2. Let f:$\mathbb{R} \mapsto \mathbb{R}$ be a continuously differentiable and concave function. Then, for any point z

$$F\left(x\right) \leq F^{'}(z)(x - z) + F(z). \tag{2.17}$$

According to these two lemma, if a continuously differentiable distance function $F_e(\cdot)$ can be separated into convex and concave sections, its auxiliary function can be constructed and proved.

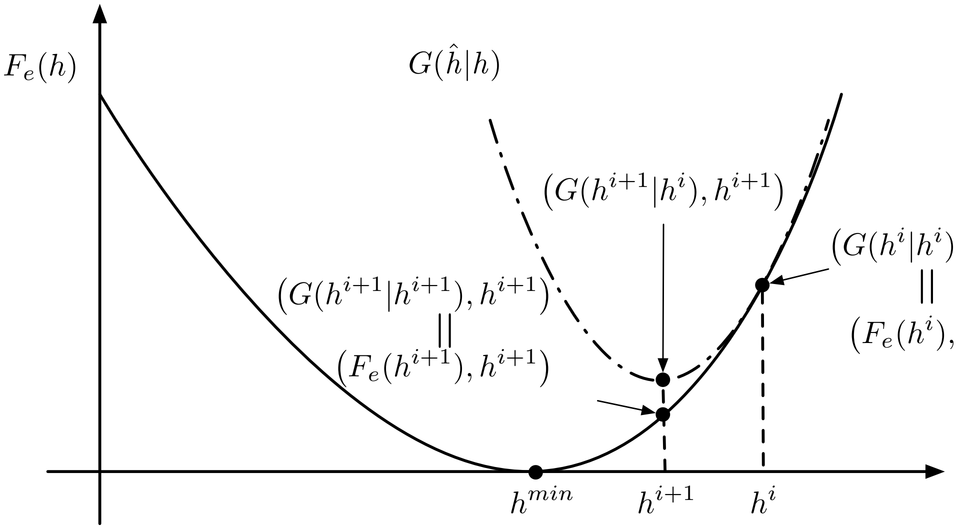

$F_e(h)$
$G(\hat{h}|h)$
$\big(G(h^{i+1}|h^i), h^{i+1}\big)$
$\big(G(h^i|h^i), h^i\big)$
$||$
$\big(F_e(h^i), h^i\big)$
$\big(G(h^{i+1}|h^{i+1}), h^{i+1}\big)$
$||$
$\big(F_e(h^{i+1}), h^{i+1}\big)$
h^{min}
h^{i+1}
h^i

Figure 2.9: A convex function and its auxiliary function in 2-D space.

2.4.3 Expectation Maximization Algorithm

Since Paris and his student Madhusudana brought the latent probability concept [97–100] into the NMF model, the expectation maximization (EM) algorithm [101] is employed to get an iterative rule for getting the convergence distribution value of the probabilistic latent component and related parameters. This concept explains the NMF framework in a probability context based on Bayes' theorem [102, 103]. The basic concept of probabilistic latent component analysis (PLCA) is if the observed non-negative entries $V_{f,t}$ are generated by an underlying distribution $p(f, t)$, the latent class $z_{i\in[1,2,\ldots,k]} \in \mathbb{Z}$ characterizes the underlying distribution $p(f,t)$ as

$$p(f,t) = \sum_{z_i} p(z_i)p(f|z_i)p(t|z_i). \tag{2.18}$$

In matrix form, the expression of the latent variable model is

$$p_{f,t} = \sum_{i=1}^{k} S_{z_i,z_i} W_{f,z_i} H_{z_i,t}, \tag{2.19}$$

and

$$\mathbf{P} = \mathbf{S} \cdot \mathbf{W} \cdot \mathbf{H}, \tag{2.20}$$

where $\mathbf{P}$ represents the probability distribution of the observed data and $\mathbf{S}$ is defined as the probability of latent components. Basis matrix $\mathbf{W}$ and activation matrix $\mathbf{H}$ are normalized in the form of $\frac{W_{f,z_i}}{\sum_i W_{f,z_i}}$ and $\frac{H_{z_i,t}}{\sum_i H_{z_i,t}}$. The Kullback−Leibler ($KL$) divergence between the model of $p(V_{f,t};\mathit{\Lambda})$ with $\Lambda = \{p(z_i), p(z_i|t), p(z_i|f)\}$ and the true distribution model $p(V_{f,t})$ is

$$D(p(V_{f,t})||p(V_{f,t};\Lambda)) = -E_{V_{f,t}}\{log[p(V_{f,t};\Lambda)]\} - Entropy(V_{f,t}). \tag{2.21}$$

Using the EM algorithm, the iterative update rule is

$$\overset{new}{\Lambda} = \arg\max_{\Lambda} Q(\Lambda, \overset{old}{\Lambda}), \tag{2.22}$$

and

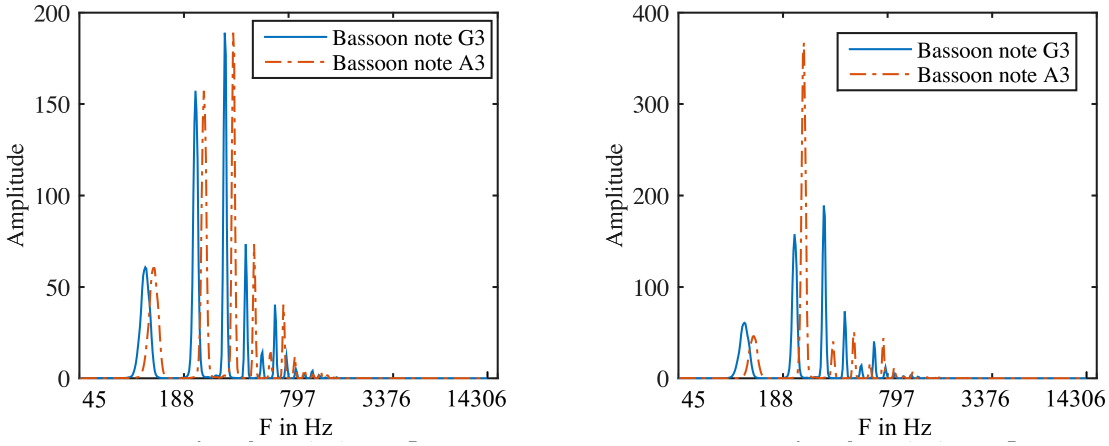
$$Q(\Lambda, \overset{old}{\Lambda}) = E_{V_{f,t}}\{E_{(z_i|V_{f,t};\overset{old}{\Lambda})}\{log[p(V_{f,t}, z_i; \Lambda)]\}\}. \tag{2.23}$$

Chapter 3

Core Work on Proposed Model

3.1 Motivation

According to the previous research mentioned, there are still a few questions that need to be resolved and our proposed source template NMF algorithm has been developed with advantages in solving such problems. For example, the work of shift-invariant operands [5–10, 22, 71] are based on the assumption that the sound source consists of translations of a single envelope basis function and the magnification is constant.

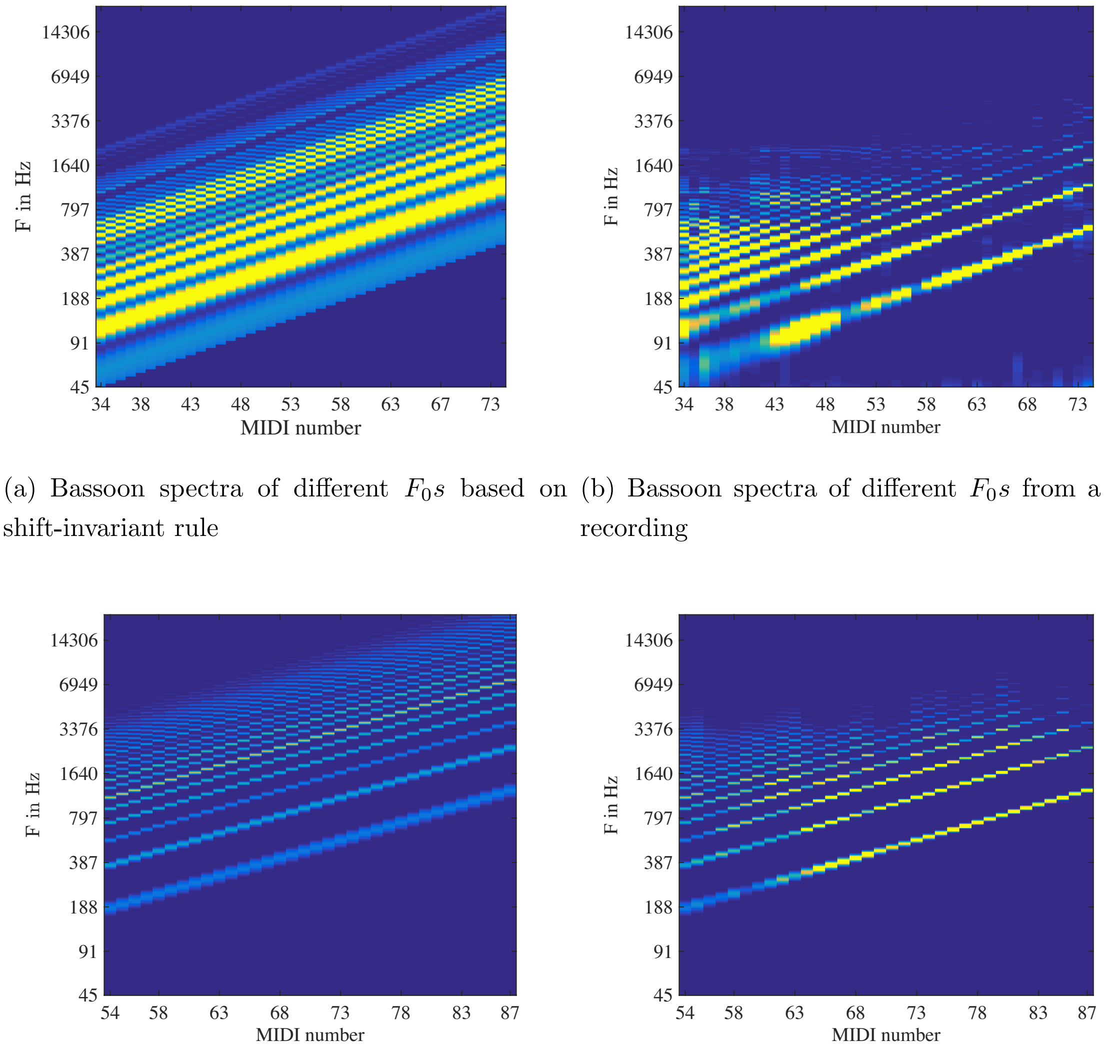


(a) Two notes' spectra from the same instrument based on the shift-invariant theory

(b) Two notes' spectra from the same instrument from a real recording

Figure 3.1: Note spectra comparison between the shift-invariant theory and reality condition ($f_s = 44100$ Hz, 8 frequency bins per note).

A simple example of this assumption is illustrated in Fig. 3.1(a) which describes a bassoon note $G3$ in the solid line and $A3$ in the dotted line. According to the assumption of shift-invariant operands, the envelope of the note $A3$ is only a unit constant magnification shift of note $G3$.



(a) Bassoon spectra of different F_0s based on shift-invariant rule

(b) Bassoon spectra of different F_0s from a real recording

(c) Trumpet spectra of different F_0s based on shift-invariant rule

(d) Trumpet spectra of different F_0s from a real recording

Figure 3.2: Comparisons of instrument spectra of different fundamental frequency F_0s between the shift-invariant rules and reality ($f_s = 44100$ Hz and 8 frequency bins per note). The note range of bassoon is $A1^{\#} \sim D5$ and trumpet is $E3 \sim D6$).

According to reality, the spectrum of these two notes $A3$ and $G3$ are different in the frequency profile envelope illustrated in Fig. 3.1(b). The comparison in Fig. 3.1 demonstrates that the shift-invariant operand method does not work well enough to deal with real conditions because its shift rules are too rigid. Moreover, this phenomenon becomes obvious in comparison of the spectra of different fundamental frequencies denoted as F_0s by instrument in Fig. 3.2. For a specific source like the bassoon, there is a "missing fundamental" problem [104–106] in Fig. 3.2(b) where though for some notes there is no apparent source or component of their pitch, they still can be heard. If the shift-invariant rule is still employed to solve the source separation problem, it is hard to build the note spectra of actual recorded instruments like in Fig.3.2(b) and 3.2(d).

The second problem exists in score informed separation methods [1,2,29,75], which get their results with the necessary help of synchronized time or score alignment information from a source like a MIDI file. Most of the time, it is hard to get the synchronized note trigger-decay information, and sometimes the score alignment information. This situation makes these algorithms limited in practice.

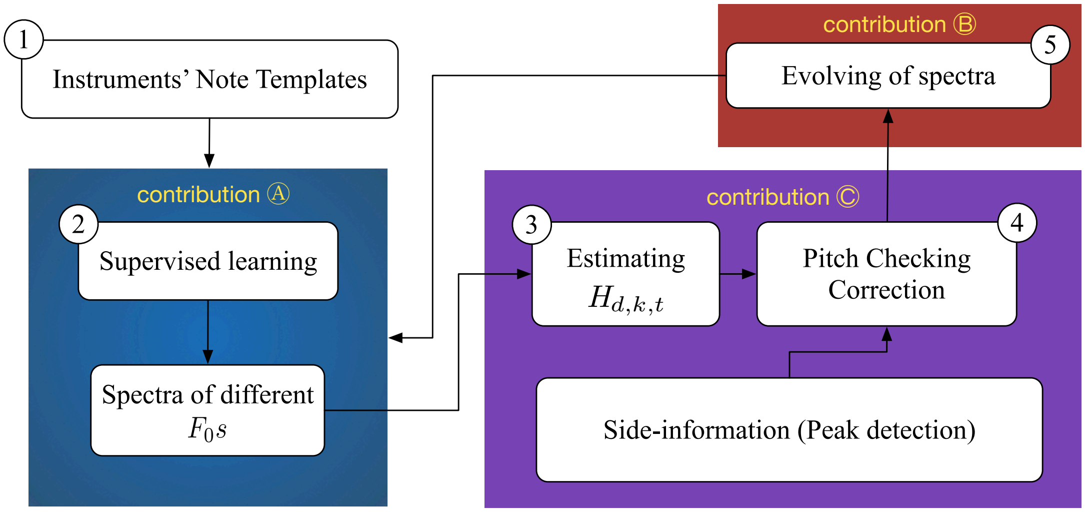

1
Instruments' Note Templates
contribution Ⓐ
2
Supervised learning
Spectra of different F_0s
contribution Ⓑ
5
Evolving of spectra
contribution Ⓒ
3
Estimating $H_{d,k,t}$
4
Pitch Checking Correction
Side-information (Peak detection)

Figure 3.3: Diagram of proposed algorithm.

In this chapter, the core model of the proposed source template NMF is introduced. It contains the supervised learning algorithm or called shift-variant operand to solve the problem of shift-invariant theory. Besides this, it estimates the note onset-offset-like features and supervises them under the pitch-check algorithm to abandon the need of a MIDI file. The diagram of our proposed algorithm is summarised in Fig. 3.3. Stage 1 and 2 are explained in Sec. 3.3. Stage 3 and 4 are introduced in Sec. 3.4.1-3.4.2 and the iterative update rule of the final stage 5 is derived in Sec. 3.4.3.

3.2 Mathematical Model

3.2.1 Mathematical Model of NMF

The work of proposed source template NMF starts from a statement of two-dimensional convention. If W is defined as a column vector $[w_1, w_2, \dots, w_n]'$ and H is defined as a row vectors $[h_1, h_2, \dots, h_n]$, the convention of these two matrices is denoted as

$$\begin{aligned} \hat{V} &= W \cdot H \\ &= \sum_{n=1}^{N} w_n \cdot h_n \\ &= w_1 \cdot h_1 + w_2 \cdot h_2 + \ldots + w_N \cdot h_N, \end{aligned} \tag{3.1}$$

where $'$ is the transpose operand and $n \in \mathbb{N}$. The subscripts are index of coordinates. Therefore if we let $\hat{V}_{f,t}$ be seen as one of the entries of the reconstructed audio STFT matrix $\hat{\boldsymbol{V}}$ with size of $F \times T$,

$$\hat{\boldsymbol{V}} = \begin{pmatrix} \hat{V}_{1,1} & \hat{V}_{1,2} & \cdots & \hat{V}_{1,T} \\ \hat{V}_{2,1} & \hat{V}_{2,2} & \cdots & \hat{V}_{2,T} \\ \vdots & \vdots & \hat{V}_{f,t} & \vdots \\ \hat{V}_{F,1} & \hat{V}_{F,2} & \cdots & \hat{V}_{F,T} \end{pmatrix}, \tag{3.2}$$

then the basic expression of NMF is

$$\hat{V}_{f,t} = \sum_{\phi=1}^{N} W_{f,\phi} \cdot H_{\phi,t}, \tag{3.3}$$

where $\mathbf{W}$ has been extended to be an $F \times \Phi$ matrix $[W_1, W_2, \ldots, W_\Phi]$ of column basis vectors $W_\phi = [w_{1,\phi}, w_{2,\phi}, \ldots, w_{F,\phi}]'$ and $\mathbf{H}$ has been extended as a $\Phi \times T$ matrix $[H_1, H_2, \ldots, H_\Phi]'$ of row activation vectors $H_\phi = [h_{\phi,1}, h_{\phi,2}, \ldots, h_{\phi,T}]$. f and t denote frequency and time domains. ϕ denotes the number of potential notes whose frequency template and attack decay profiles are described by the corresponding columns of $\mathbf{W}$ and rows of $\mathbf{H}$. In the mono source separation (MSSS) problem, the basic NMF model is expanded in dimensions to indicate a number of sources and written as

$$\hat{V}_{f,t} = \sum_{d,\phi} W_{f,\phi,d} \cdot H_{d,\phi,t}, \tag{3.4}$$

where d represents the d^{th} sound source. $\mathbf{W}$ is beyond the meaning of basis vectors and denoted as spectra of different fundamental frequency F_0s, and in reality, the note's profile of f is unique for each note ϕ and source d. But, this kind of model will not be efficient in the MSSS problem because it has an absence of constraints on these spectra of different fundamental freequency F_0s to cluster the data into its target group.

3.2.2 Basic Expression of Proposed Algorithm

Considering the translation tensor $T_{f,\phi,f}$ in the shift-invariant operand [5,6,10,22,71] is only determined by parameters of the frequency and the note, one of the contributions of our proposed algorithm is extending the translation matrix from 3-D to 4-D by assuming that each instrument source has its unique characteristics of the shift operand and has been recorded as d, the extended dimension, in the proposed translation matrix $T_{d,f,\phi,\hat{f}}$. Moreover, the proposed $T_{d,f,\phi,\hat{f}}$ is updatable along the d dimension by the way of the iteration update rule to form the shift characteristics of the instrument templates. With the adaptive 4-D translation $T_{d,f,\phi,\hat{f}}$, our proposed source template NMF model based on the shift-variant operand is written as

$$\hat{V}_{f,t} = \sum_{d,\phi,\hat{f}} \mathbf{T}_{d,f,\phi,\hat{f}} \cdot \mathbf{\Omega}_{\hat{f},d} \cdot \mathbf{H}_{d,\phi,t}, \tag{3.5}$$

$$\hat{W}_{f,\phi,d} = \sum_{\hat{f}} \mathbf{T}_{d,f,\phi,\hat{f}} \cdot \mathbf{\Omega}_{\hat{f},d}, \tag{3.6}$$

where the translation matrix $\mathbf{T}$ is set up with a size $D \times F \times \Phi \times \hat{F}$ and is responsible for translating the note of an instrument into other various notes. Φ represents the number of possible translations or potential notes played in musical clips. $\hat{F}$ represents the potential component domain of the basis notes. In the shift-variant operand, $\hat{F}$ equals F, but in our proposed algorithm $\hat{F} \geq F$ as long as $T_{d,f,\phi,\hat{f}}$ and $W_{\hat{f},d}$ are satisfied with the 2-D matrix production rule in Fig. 3.4. Matrix $\hat{V}_{f,t}$, with size of $F \times T$, denotes the estimation of the STFT of the observed signal $V_{f,t}$. We defined the basic expression of adaptive shifted NMF model as

$$\begin{aligned} V_{f,t} &= \hat{V}_{f,t} + e_{f,t} \\ &= \sum_{d,\phi,\hat{f}} T_{d,f,\phi,\hat{f}} \cdot \Omega_{\hat{f},d} \cdot H_{d,\phi,t} + e_{f,t}. \end{aligned} \tag{3.7}$$

Under a maximum *a posteriori* (MAP) assumption [107, 108], when the residual $e_{f,t}$ has a normal distribution $\mathcal{N}((f,t);0,1)$, the cost function of our proposed algorithm becomes

$$\mathcal{D}_{\mathcal{E}}(\mathbf{V}||\hat{\mathbf{V}}) = \frac{1}{2}\sum_{f,t}\left(V_{f,t} - \hat{V}_{f,t}\right)^2, \tag{3.8}$$

and if the residual $e_{f,t}$ has a Poisson distribution, the cost function of our proposed algorithm becomes the Kullback–Leibler (K-L) divergence in the form of

$$\mathcal{D}_{\mathcal{E}}(\mathbf{V}||\hat{\mathbf{V}}) = \sum_{f,t}\left(V_{f,t} \cdot log\frac{V_{f,t}}{\hat{V}_{f,t}} - V_{f,t} + \hat{V}_{f,t}\right). \tag{3.9}$$

In Fig. 3.4, the flow chart of the proposed source template NMF is illustrated and divided into three levels of summations to reconstruct the targeted mixtures signal. The first level is the summation in the potential component domain $\hat{f}$. It allows the translation matrix $\mathbf{T}$ to combine with the basis fundamental frequency F_0 template $\boldsymbol{\Omega}$ to calculate the spectra of different fundamental frequency F_0s $\mathbf{W}$. In the next level of summation of the score domain Φ, the spectra within $\mathbf{W}$ are controlled by their activation in $\mathbf{H}$ to construct the estimated STFT by the source. Finally, these specified STFTs are summed along the D dimension to estimate the spectrogram of the observed the audio mixture.

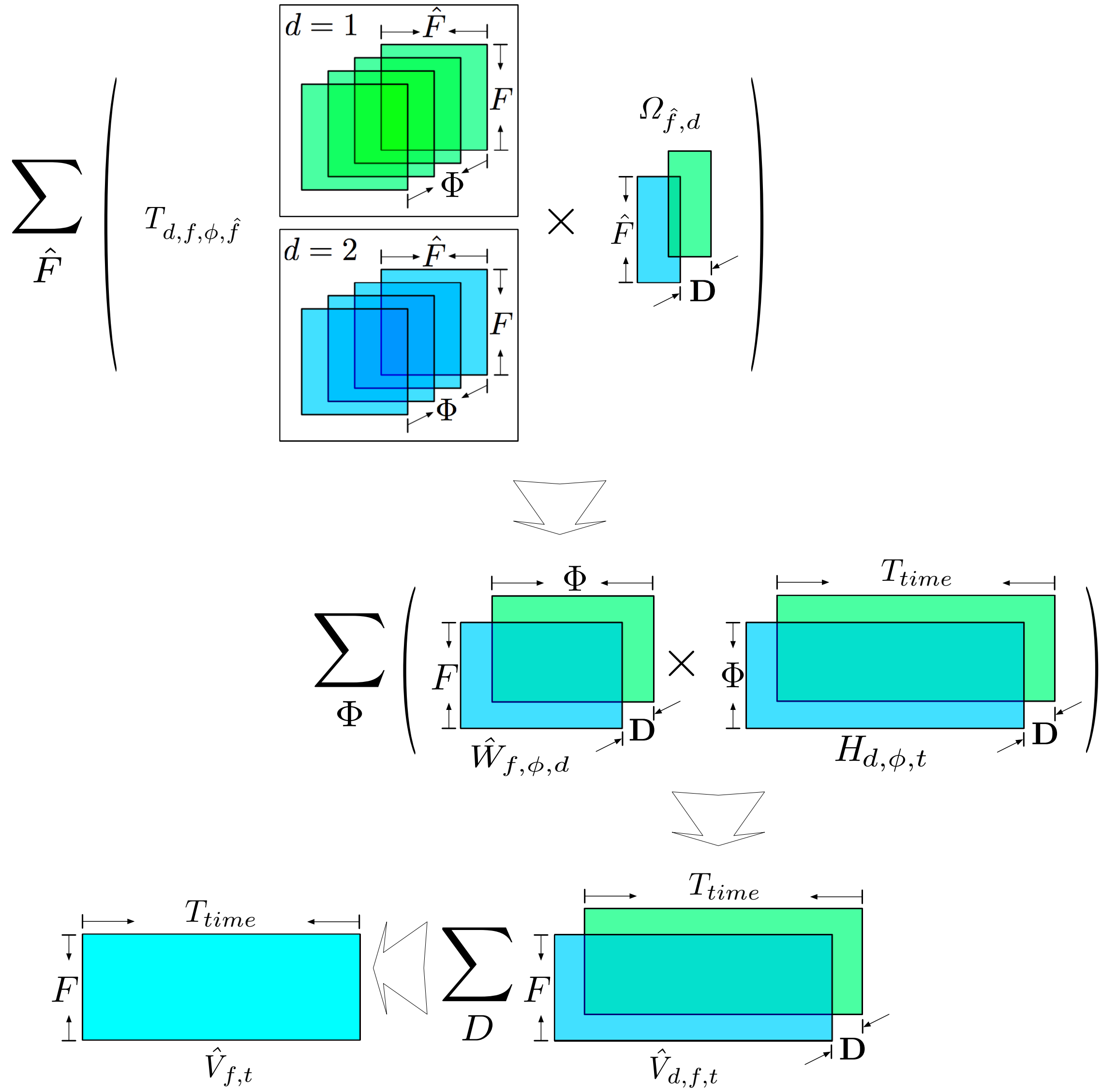


Figure 3.4: Sketch of proposed source template NMF

3.3 Instrument Labeled Note Templates and Supervised Learning Algorithm

The distinction in this thesis is made merely to introduce the new concept of notes' spectrum template, $\mathbf{W}$, which is actually the spectra of different $F_0 s$ by assuming that it consists of the resonance responses of different $F_0 s$, $\mathbf{T} \times \mathbf{\Omega}$, from a basis F_0 impulse excitation $\mathbf{\Omega}$. The new symbols of $T_{d,f,\phi,\hat{f}} \in \mathbb{R}^{D \times F \times \Phi \times \hat{F}}$ and $\Omega_{\hat{f},d} \in \mathbb{R}^{\hat{F} \times D}$ are employed here to be the 4-D resonance response translation matrix and the 2-D basis impulse excitation matrix where $T_{d,f,\phi,\hat{f}}$ represents the resonance coefficient of specific instrument and note at specific frequency f, and $W_{\hat{f},d}$ represents the impulse excitation of specific instrument. In addition, $\hat{f}$ forms an additional frequency dimension that allows the selection of a different frequency resolution over what $\hat{f}$ may offer. If we denote $\hat{W}_{f,\phi,d}$ as the reconstructed notes' spectrum template by instrument, the relationship among $\hat{\mathbf{W}}$, $\mathbf{T}$ and $\mathbf{\Omega}$ would be

$$\begin{aligned} W_{f,\phi,d} &= \sum_{\hat{f}} T_{d,f,\phi,\hat{f}} \cdot \mathit{\Omega}_{\hat{f},d} + \mathcal{E}^{W}_{f,\phi,d} \\ &= \hat{W}_{f,\phi,d} + \mathcal{E}^{W}_{f,\phi,d}, \end{aligned} \tag{3.10}$$

where $\mathcal{E}^{W}_{f,\phi,d}$ is an error/residual term under a maximum *a posteriori* (MAP) assumption [107] [108]. When the residual has a normal distribution $\mathcal{N}(f,t;\mu=0,\sigma=1)$, the cost function of our proposed algorithm is based on Euclidean distance and becomes

$$D_{Eu}(\mathbf{W}||\hat{\mathbf{W}}) = \frac{1}{2}\sum_{d,f,\phi}\big(W_{f,\phi,d} - \sum_{\hat{f}} T_{d,f,\phi,\hat{f}} \cdot \mathit{\Omega}_{\hat{f},d}\big)^2. \tag{3.11}$$

When the residual obeys the rule of Poisson distribution [107], the cost function of our proposed algorithm is based on Kullback-Leibler (K-L) distance and becomes

$$\begin{aligned} D_{K-L}(\mathbf{W}||\hat{\mathbf{W}}) = \sum_{d,f,\phi}\big(&W_{f,\phi,d} \cdot log\frac{W_{f,\phi,d}}{\sum_{\hat{f}} \mathbf{T}_{d,f,\phi,\hat{f}} \cdot \mathit{\Omega}_{\hat{f},d}} \\ &- W_{f,\phi,d} + \sum_{\hat{f}} T_{d,f,\phi,\hat{f}} \cdot \mathit{\Omega}_{\hat{f},d}\big). \end{aligned} \tag{3.12}$$

Considering $\mathbf{T}$ and $\boldsymbol{\Omega}$ works under the framework of (3.7), we substitute (3.10) into it to get our proposed updatable shift-variant NMF model as

$$\begin{aligned} V_{f,t} &= \sum_{d,\phi} \Big(\big(\sum_{\hat{f}} T_{d,f,\phi,\hat{f}} \cdot \mathit{\Omega}_{\hat{f},d} + \mathcal{E}^W_{f,\phi,d} \big) \cdot H_{\phi,t,d} \Big) + \mathcal{E}^V_{f,t} \\ &= \hat{V}_{f,t} + \big(\sum_{d,\phi} \mathcal{E}^W_{f,\phi,d} \cdot H_{\phi,t,d} \big) + \mathcal{E}^V_{f,t} \\ &= \hat{V}_{f,t} + \mathcal{E}^W_{f,t} + \mathcal{E}^V_{f,t}, \end{aligned} \tag{3.13}$$

where $\hat{V}_{f,t}$ is the reconstructed spectrogram of the audio mixture. $\mathcal{E}^W_{f,t}$ is the residual coming from the process of the notes' spectrum template reconstruction and $\mathcal{E}^V_{f,t}$ is the residual coming from the procedure of the spectrogram reconstruction. If they obey the same rules as the statistical distribution, the overall residual of the proposed algorithm can be written as

$$\mathcal{E}_{f,t} = \mathcal{E}^W_{f,t} + \mathcal{E}^V_{f,t}. \tag{3.14}$$

A sketch of the source template model (without $\mathcal{E}_{f,t}$) is also shown in Fig. 3.4 for the multi-source separation problem and indicated in the first level of sums. The task of this level is to form the reconstructed source template spectra by $T_{d,f,\phi,\hat{f}} \cdot \mathit{\Omega}_{\hat{f},d}$ throughout the index $\hat{f}$ and does note clustering.

3.3.1 Supervised Learning Operand

The core idea of the proposed source template NMF is to find the notes' spectrum templates of the actual recorded instruments and their activation. By masking the specified template and the corresponding trigger-decay profiles, the designated channel audio can be reconstructed. Then the procedure of finding the notes' spectrum templates of the actual recorded instruments starts at the initializations of $\mathbf{T}$ and $\boldsymbol{\Omega}$. Though the spectra of different F_0s of the actual recorded instrument are hard to obtain prior to the separation, the initialization procedure can be a supervised learning to use side-information of the individual note audio of the similar instruments from standard datasets [3, 4, 109]. When we obtained the different F_0s spectra, $\mathbf{W}$, from the side-information, the following work is to factor it into underlying factors of $\mathbf{T}$ and $\boldsymbol{\Omega}$, which play a role of a translation matrix providing the shift-variant rule and restoring the rule's weights, source template coefficients, and role of a basis matrix representing the basis F_0 impulse excitation individually.

Construction of side-information based $\mathbf{W}$

The beginning of this supervised learning starts from the prior set up of $\mathbf{W}$, the spectra of different notes, obtained from a set of instrument sounds' spectrograms. A simple example of making three notes' spectrum template is drawn in Fig. 3.5 with notes D_3, $D_3^\#$ and E_3. Each note has an audio waveform at sampling frequency 44100 Hz. A windowing function ω with the width of 4096 point is adopted to pick up a spectrum from the centre of the score audio spectrogram and make this selected spectrum as the corresponding score spectrum in $\mathbf{W}$. This operand gives two aspects of information: the first is the resonance response character of each note; the second is the potential candidates of notes by an instrument in a audio mixture. When a side-information based $\mathbf{W}$ is given, the following work is to get the optimized $\mathbf{T}$ and $\mathbf{\Omega}$.

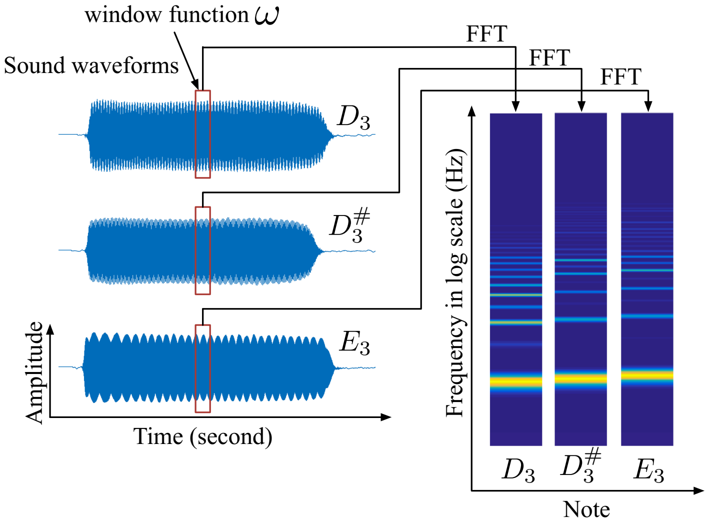


Figure 3.5: Making a notes' spectrum template from the RWC music database [3,4]

Convergence method by means of multiplicative iteration rules

In this part, multiplicative update rules [26–28] are developed to iteratively update the proposed adaptive translation matrix $\mathbf{T}$. The iterative update rules are derived from a gradient descent optimization algorithm [110] and the key issue is the selection of step size η to maintain non-negativity and zero keeping results. The basic expression of the gradient descent optimization algorithm of $\mathbf{T}$ is

$$\overset{n+1}{T}_{d,f,\phi,\hat{f}} = \overset{n}{T}_{d,f,\phi,\hat{f}} - \overset{n}{\eta}_T \cdot \frac{\partial D_{Eu}(\mathbf{W}||\overset{n}{\hat{\mathbf{W}}})}{\partial \overset{n}{T}_{d,f,\phi,\hat{f}}}, \tag{3.15}$$

where n and $n+1$ represent the n^{th} and the $(n+1)^{th}$ iteration. $\overset{n}{\eta}_T$ represents the step size in the gradient descent algorithm. To develop the update rules of $\mathbf{T}$, first consider the partial derivatives of $\hat{W}_{f,\phi,d}$ with respect to the adaptive translation matrix given as

$$\frac{\partial \hat{W}_{f,\phi,d}}{\partial T_{d,f,\phi,\hat{f}}} = \Omega_{\hat{f},d}. \tag{3.16}$$

Combining (3.11) with (3.16), the partial derivative of $D_{Eu}(\mathbf{W}||\overset{n}{\hat{\mathbf{W}}})$ with respect to $\overset{n}{T}_{d,f,\phi,\hat{f}}$ is

$$\frac{\partial D_{Eu}(\mathbf{W}||\overset{n}{\hat{\mathbf{W}}})}{\partial \overset{n}{T}_{d,f,\phi,\hat{f}}} = -\left(W_{f,\phi,d} - \overset{n}{\hat{W}}_{f,\phi,d}\right) \cdot \overset{n}{\Omega}_{\hat{f},d}. \tag{3.17}$$

If we set $\overset{n}{\eta}_T$ to be the value of

$$\overset{n}{\eta}_T = \frac{\overset{n}{T}_{d,f,\phi,\hat{f}}}{\overset{n}{\hat{W}}_{f,\phi,d} \cdot \overset{n}{\Omega}_{\hat{f},d}}, \tag{3.18}$$

the update rule for $T_{d,f,\phi,\hat{f}}$ in (3.15) becomes the multiplicative update rule

$$\begin{aligned} \overset{n+1}{T}_{d,f,\phi,\hat{f}} &= \overset{n}{T}_{d,f,\phi,\hat{f}} + \frac{\overset{n}{T}_{d,f,\phi,\hat{f}}}{\overset{n}{\hat{W}}_{f,\phi,d} \cdot \overset{n}{\Omega}_{\hat{f},d}} \cdot \left(W_{f,\phi,d} - \overset{n}{\hat{W}}_{f,\phi,d}\right) \cdot \overset{n}{\Omega}_{\hat{f},d} \\ &= \overset{n}{T}_{d,f,\phi,\hat{f}} \cdot \left(\frac{W_{f,\phi,d} \cdot \overset{n}{\Omega}_{\hat{f},d}}{\left(\sum_{\hat{f}} \overset{n}{T}_{d,f,\phi,\hat{f}} \cdot \overset{n}{\Omega}_{\hat{f},d}\right) \cdot \overset{n}{\Omega}_{\hat{f},d}}\right). \end{aligned} \tag{3.19}$$

Using the same approach, the update rule of $\mathbf{\Omega}$ is denoted as

$$\overset{n+1}{\Omega}_{\hat{f},d} = \overset{n}{\Omega}_{\hat{f},d} \frac{\sum_{f,\phi} W_{f,\phi,d} \cdot \overset{n}{T}_{d,f,\phi,\hat{f}}}{\sum_{f,\phi}\left(\sum_{\hat{f}} \overset{n}{T}_{d,f,\phi,\hat{f}}\right) \cdot \overset{n}{T}_{d,f,\phi,\hat{f}}}. \tag{3.20}$$

Additionally, based upon the K-L distance, the multiplicative update rules of the supervised learning are

$$\overset{n+1}{T}_{d,f\phi,\hat{f}} = \overset{n}{T}_{d,f\phi,\hat{f}} \frac{W_{f,\phi,d} \sum_{\hat{f}} (\overset{n}{T}_{d,f,\phi,\hat{f}} \cdot \overset{n}{\Omega}_{\hat{f},d}) \overset{n}{\Omega}_{\hat{f},d} \overset{n}{T}_{d,f,\phi,\hat{f}}}{\overset{n}{\Omega}_{\hat{f},d}}, \tag{3.21}$$

and

$$\overset{n+1}{\Omega}_{\hat{f},d} = \overset{n}{\Omega}_{\hat{f},d} \sum_{f,\phi} \Big(\sum_{\hat{f}} \overset{n}{T}_{d,f,\phi,\hat{f}} \cdot \overset{n}{\Omega}_{\hat{f},d} \Big). \tag{3.22}$$

Initialization of supervised learning operand

The initialization of the translation matrix $T_{d,f,\phi,l}$ is informed according to the prior information. For example, if we have sheet music of a combined clip of the bassoon and trumpet in Fig. 3.6, it is easy to conclude that the basis note of the bassoon is C_4 and the trumpet is $D4$. Considering the scale on the basis note to others, for the bassoon translation matrix, it needs the shift operand to include $C4$ to $D4$, $C4$ to $D4\#$, and $C4$ to $E4$. But for the trumpet translation matrix, it only requires the shift operand for $D4$ to $E4$. In Fig. 3.7, the **T** is initialized separately for the bassoon and the trumpet.

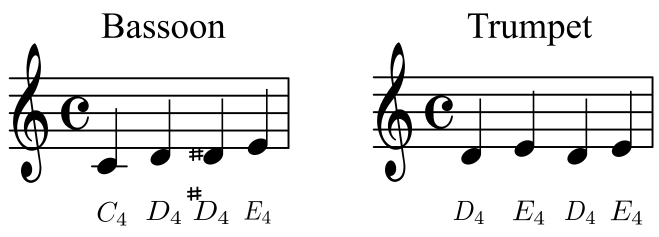
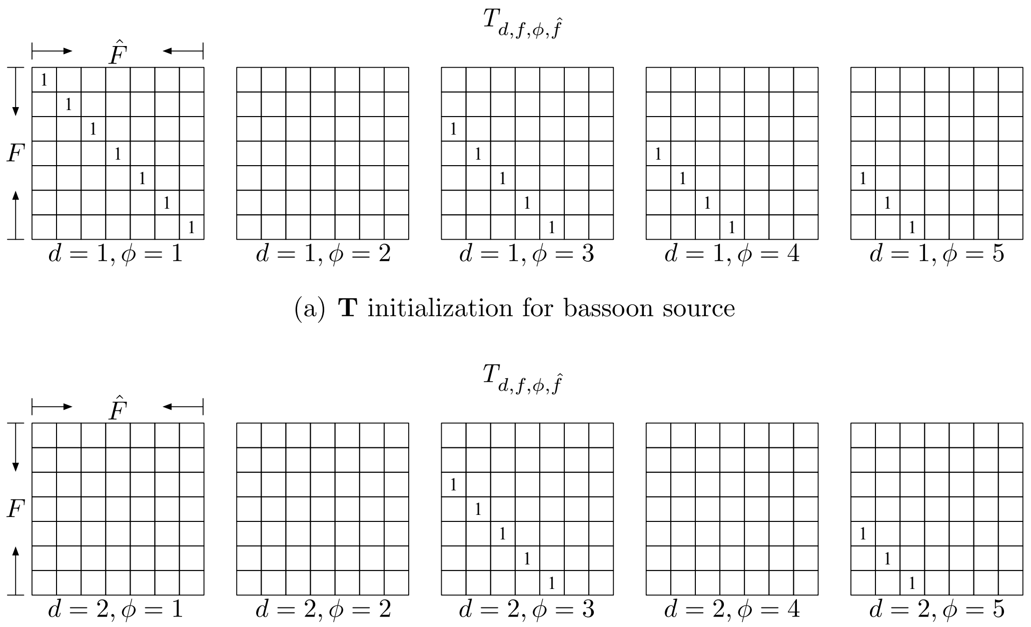


Figure 3.6: Sheet music of a mixture clip

When $d = 1$, it corresponds to the bassoon translation matrix initialization and is represented as $T_{d=1,f,\phi,\hat{f}}$. The $\phi = 1$ situation is equivalent to the shift operand from $C4$ to $C4$. However, $D4$ is not included in the bassoon playing list, so $T_{d=1,f,\phi=2,\hat{f}}$ is filled with zeros. Then all situations described in Fig 3.7(a) of $\phi = 3$, $\phi = 4$, and $\phi = 5$ represents the shift operands $C4$ to $D4$, $D4\#$, and $E4$ individually. Alternatively, Fig. 3.7(b) demonstrates how the trumpet translation matrix initialization is figured out. Since there are two notes, $D4$ and $E4$, on the trumpet playing list, the magnification



(a) **T** initialization for bassoon source

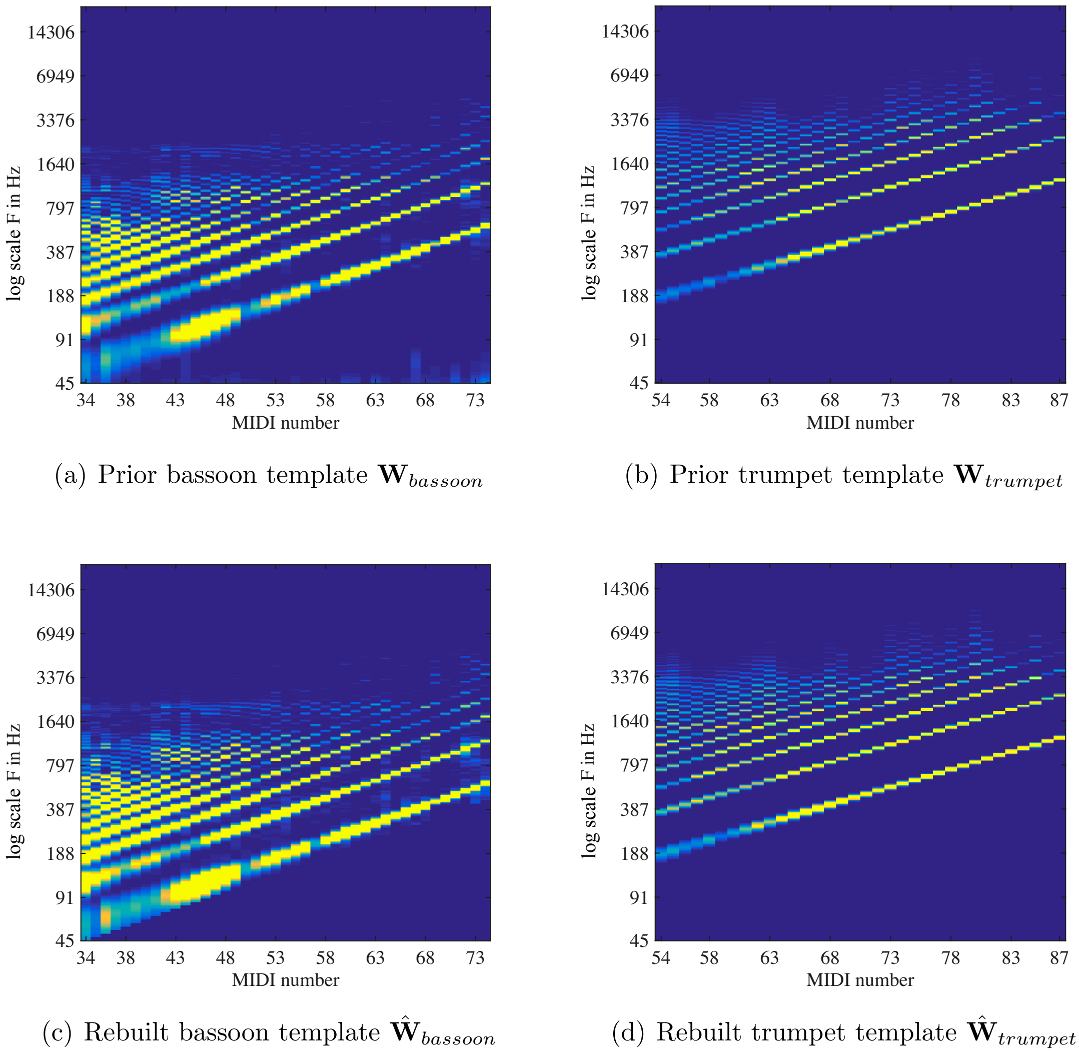
(b) **T** initialization for trumpet source

Figure 3.7: Prior set-up of proposed shift-variant translation matrix **T** in supervised learning operand.

entries within the trumpet translation matrix $T_{d=2,f,\phi,\hat{f}}$ are initialized to 1 only at $\phi = 3$ and $\phi = 5$. In this step of the proposed work, the magnification entries within the translation matrix are set to 1, but during the iterations, the magnification entries will be adjusted adaptively to meet the demands of instrument separation. For improving the separation results, we need the proposed source template NMF to do a supervised learning from prior information. It will have several advantages such as discovering clues for notes and timbres of instruments, supporting the prediction of the unknown values of underlying matrices, and self-upgrading to accommodate the instruments' individual differences in separation. The prior information is the spectra of different F_0s of the specific instrument, $\mathbf{W}$, which is introduced in Sec. 3.3.1. The audio of each individual note is provided by the database in [3, 4, 109]. Essentially, the proposed supervised learning is the process of factoring the $W_{f,\phi,d}$ into the non-negative matrices $T_{d,f,\phi,\hat{f}}$ and $\Omega_{\hat{f},d}$. The supervised learning factored results are regarded to be the initialization of **T** and $\boldsymbol{\Omega}$, which involve the prior knowledge of the instrument range and notes' frequency profiles into our proposed model.

3.3.2 Exampler results of supervised learning operand

Based on the iteration update rules introduced in Sec. 3.3.1 and datasets [1], the supervised learning results of a variety of instruments are given in the following paragraphs. The target instrument templates $\mathbf{W}_{bassoon}$ and $\mathbf{W}_{trumpet}$ are presented in Fig 3.8(a) and 3.8(b), meanwhile their rebuilt results are given in Fig. 3.8(c) and 3.8(d). The note recordings are all provided from Iowa University Electronic Music Studios Lab [109] and McGill University [3, 4]. The target instrument templates of trumpet



(a) Prior bassoon template $\mathbf{W}_{bassoon}$

(b) Prior trumpet template $\mathbf{W}_{trumpet}$

(c) Rebuilt bassoon template $\hat{\mathbf{W}}_{bassoon}$

(d) Rebuilt trumpet template $\hat{\mathbf{W}}_{trumpet}$

Figure 3.8: Prior and rebuilt instrument templates. (Fs = 44100 Hz and the bin in log frequency scale is $\frac{1}{4}$ note.)

from MIDI numbers 54 to 87 and bassoon from 34 to 73 are built by their note audio. The residue between the target and rebuilt templates are shown in Table 3.1. The reconstructed $\mathbf{W}$ are calculated by the product of $\mathbf{T}$ and $\boldsymbol{\Omega}$. The residual $\mathcal{E}$ between the target and rebuilt instrument templates are used to evaluate the supervised learning performances. The rebuilt accuracy of templates is nearly perfect.

Table 3.1: Supervised learning results

Instrument	$\mathcal{E}_{start}$	$\mathcal{E}_{end}$	**Iterations**	$\frac{W}{\hat{W}}\%$
Bassoon	31555066.31	11665.78	1	99.98
Trumpet	3920484.98	88.36	1	100

Note: $\mathcal{E} = \mathbf{D}_{Eucild}(W||\hat{W})$. W is notes' template by source and $\hat{W}$ is the rebuilt notes' template by our supervised learning algorithm.

The supervised learning algorithm is performed based on the shift-variant operand and based on the results in Table 3.1, $\hat{\mathbf{W}}$ is a high accuracy reconstructed copy of the instrument notes' template $\mathbf{W}$. In the shift-variant operand, the factorized $\mathbf{T}$ and $\boldsymbol{\Omega}$ restores the translation coefficients and the basis note spectrum separately. The supervised learnt $T_{d,f,\phi,\hat{f}}$ and $\mathit{\Omega}_{\hat{f},d}$ are provided in Fig. 3.10. The barrier for a convenient viewing is the 4 dimensional factor $T_{d,f,\phi,\hat{f}}$. It can not be observed in 2-D plane directly and need a special angle to show the details. If we use a specific $\tilde{d}$ to exhibit $T_{\tilde{d},f,\phi,\hat{f}}$, the shift-variant translation matrix $\mathbf{T}$ can be depicted in 3-D space as drawn illustrated in Fig. 3.9. For any note ϕ, the shift weights are arranged in the form of a diagonal matrix. If we take a look at $\mathbf{T}$ along the "ϕ direction" in Fig. 3.9, all the shift weights will be viewed easily in Fig. 3.10.

Present the shift weights of the bassoon $\mathit{\Phi}$ notes within $\mathbf{T}_{bassoon}$ together in Fig. 3.10(a). Similarly, the shift weights of trumpet $\mathit{\Phi}$ notes within $\mathbf{T}_{trumpet}$ is presented in Fig. 3.10(b). Compared with the shift-invariant operand, these shift weights are not kept at unity anymore. With the increase of $\mathit{\Phi}$, the length of a group of shift weights becomes shorter, which corresponds to a note and arranged in a line parallel to the diagonal. So it lets us arrange the shift weight of one instrument into a lower triangular matrix for convenient viewing. The supervised learning results of the basis

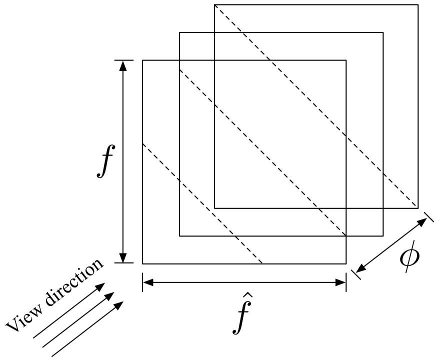


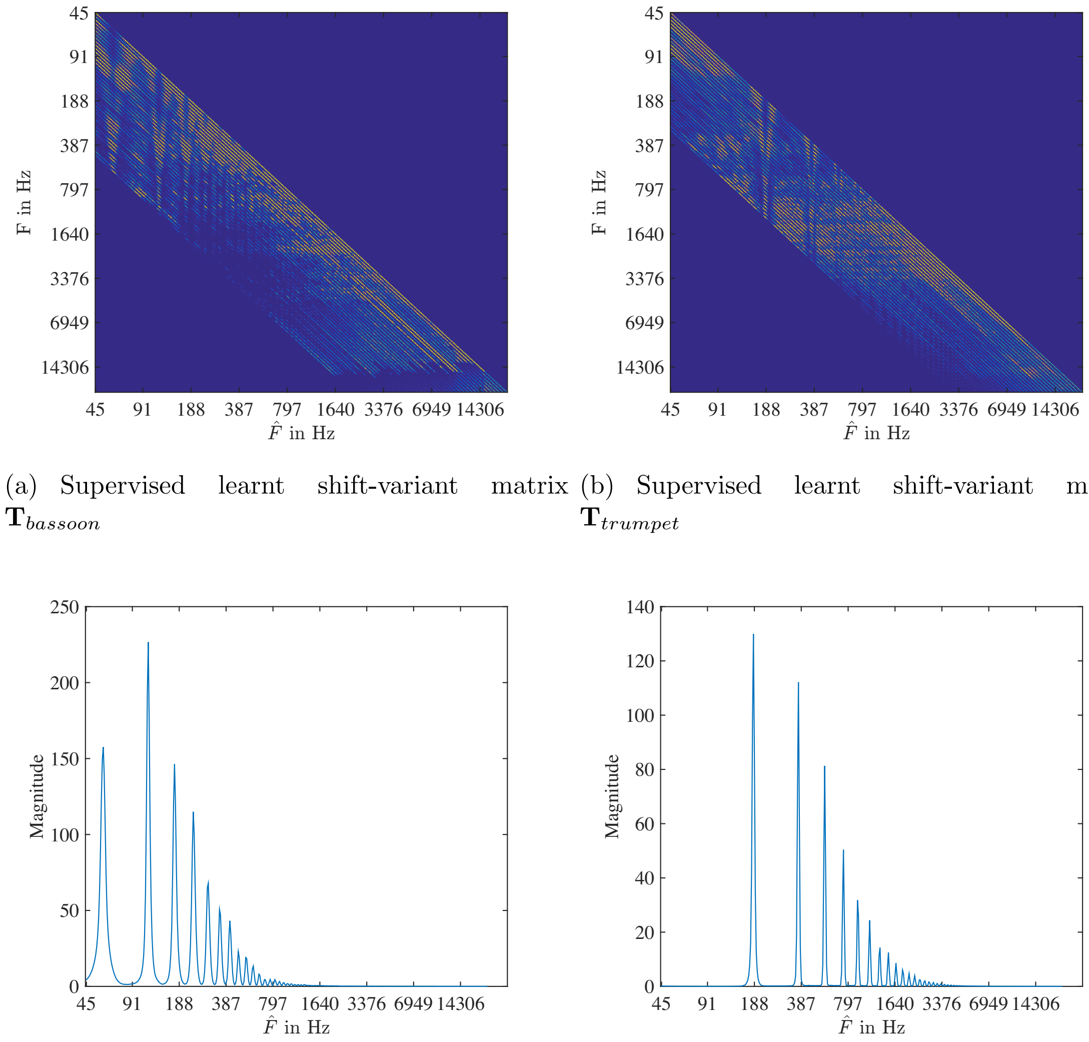
Figure 3.9: Convenient view of shift-variant translation matrix $\mathbf{T}$

matrix $\mathbf{\Omega}$ are plotted in Fig. 3.10(c) and 3.10(d). The entries of the basis F_0 template represent the impulse excitation by the instrument. The difference between the learnt $\mathbf{\Omega}_{bassoon}$ and $\mathbf{\Omega}_{trumpet}$ is the ratio of each frequency component weight to their sum. The components are observed as fundamental frequency and its harmonics which are formed in peaks in Fig. 3.10(c) and 3.10(d). This can be calculated by the work in [111] and considered the pattern of the instrument. Therefore, it will be a discussion of the basis F_0 template $\mathit{\Omega}_{\hat{f},d}$ and the shift-variant matrix $T_{d,f,\phi,\hat{f}}$ later to assume their functions in the impulse excitation and the resonance response at different F_0s.

3.3.3 Comparing the Supervised Learning Algorithm with Previous Research

Proposed shift-variant operand VS shift-invariant operand

To further understand the procedures of using the shift-invariant or variant copies from the basis F_0 template to form the spectra of different scores, a structural comparison is made. In [5–8] presented diagrammatically in Fig 3.11, the shift operand is denoted as a downward notation $\downarrow \phi$ on the top of the basis matrix $\mathbf{W}$ to indicate the operand of moving the entries $W_{f,d}$ downward along the frequency ordinate, where the entries beyond the bottom vanish and their vacancies on the top are replaced by zeroes. ϕ is the parameter representing the score pitch and corresponds to the shift-invariant downward operand. In [9, 10] given diagrammatically in Fig 3.12, a shift-invariant translation tensor $\boldsymbol{\mathcal{T}}$ filled with unit shift-weights is introduced. Its



(a) Supervised learnt shift-variant matrix $\mathbf{T}_{bassoon}$

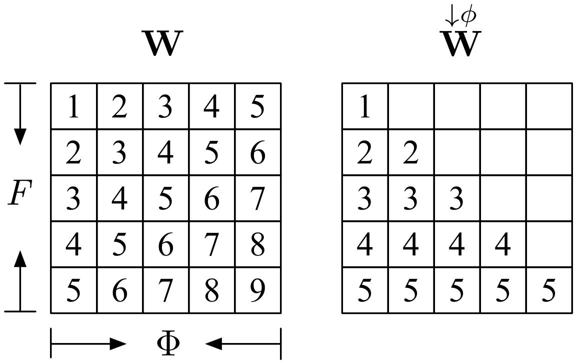
(b) Supervised learnt shift-variant matrix $\mathbf{T}_{trumpet}$

(c) Supervised learnt basis F_0 template $\boldsymbol{\Omega}_{bassoon}$

(d) Supervised learnt basis F_0 template $\boldsymbol{\Omega}_{trumpet}$

Figure 3.10: Supervised learnt $\mathbf{T}$ and $\boldsymbol{\Omega}$. (Fs = 44100 Hz and the bin in log frequency scale is $\frac{1}{4}$ note.)

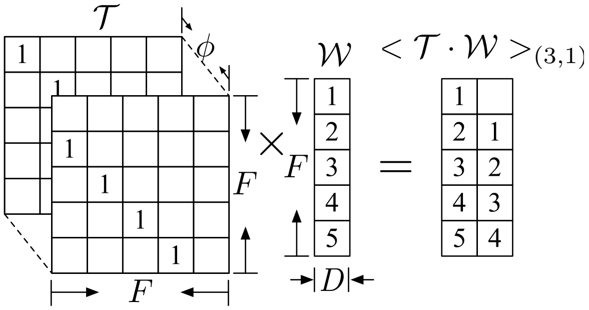
shift-invariant operand is demonstrated by the product of $\boldsymbol{\mathcal{T}} \times \boldsymbol{\mathcal{W}}$. The characteristics of the shift-invariant translation tensor $\boldsymbol{\mathcal{T}}$ are non-updatable, non-prior setting by the source and its shift-weights keep the unit form during the separation process. For comparison, our proposed shift-variant operand is shown in Fig 3.13 for performing the supervised learning algorithm and extended the concept from the shift-weights and the basis vector to resonance coefficients and impulse excitation matrices. In our proposed algorithm, the product $T_{d,f,\phi,\hat{f}} \times \Omega_{\hat{f},d}$ will move its pitch value toward a high position along the log scale f axis by increasing ϕ in a fixed d. Theoretically, the parameter $\hat{f}$, the fourth subscript of $T_{d,f,\phi,\hat{f}}$ and the first subscript of $W_{\hat{f},d}$, can adopt any value to fit the matrix multiplication rule, but for the simplified calculation, we let $\hat{f}$ equal to f in this thesis. In the scope of the proposed model, the matrix $\mathbf{W}$ appears as the notes' spectrum template and the matrix $\mathbf{\Omega}$ is denoted as the basis F_0 excitation impulse. The resonance translation matrix $\mathbf{T}$ is designed as an underlying factor and can be evolved in the separation as well as $\mathbf{\Omega}$ and $\mathbf{H}$.



1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8
5	6	7	8	9

1				
2	2			
3	3	3		
4	4	4	4	
5	5	5	5	5

Figure 3.11: Examples of shift operand in [5–8].



1
2
3
4
5

1	
2	1
3	2
4	3
5	4

Figure 3.12: Examples of shift operand in [9, 10].

Supervised Learning Operand Versus Score Training Session

The novelty of our proposed supervised learning procedure is to deconvolve the prior $\mathbf{W}$ into $\mathbf{T}$ and $\mathbf{\Omega}$. In Fig. 3.14, the supervised learnt $\mathbf{T}$ is assumed to be a container

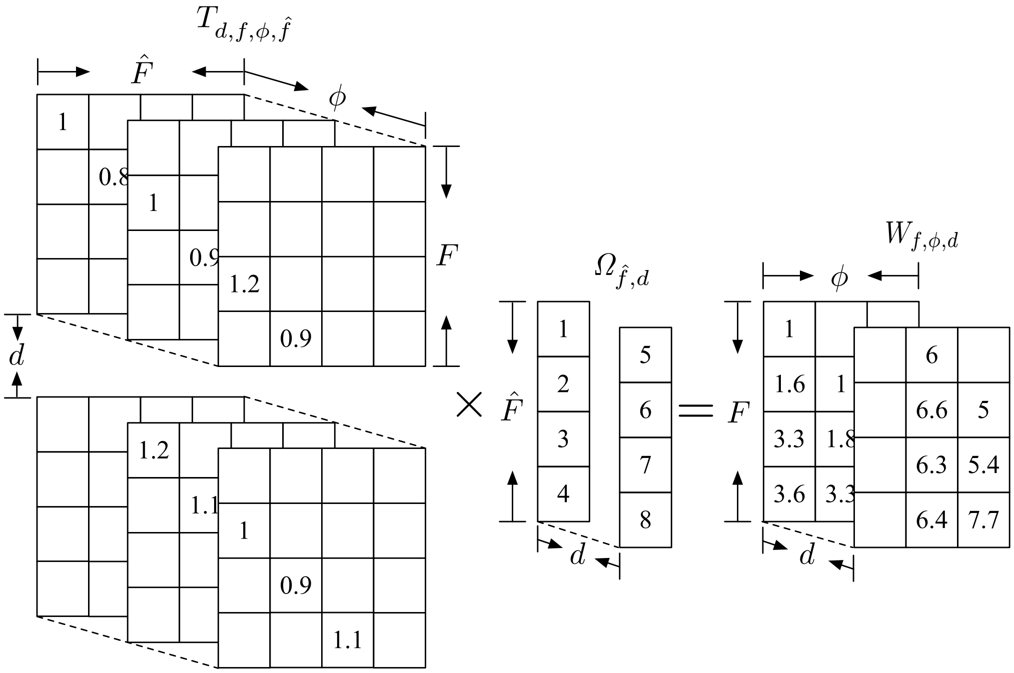


Figure 3.13: Examples of shift operand in proposed shift-variant operand.

of F_0 variant resonance response coefficients which amplify or compress the amplitude of the basis F_0 impulse excitation $\mathbf{\Omega}$ by note and source. Fig. 3.15 figures out the training procedure in previous work [11]. In this training, the prior knowledge of the notes' spectrogram generated from an instrument sound database was adopted and factored into a notes' spectrum template matrix $\mathbf{W}$ and an activation matrix $\mathbf{H}$. This training session cannot work without the help of prior synchronized note trigger and decay information in the activation matrix $\mathbf{H}$. Because the translation matrix $\mathbf{T}$ is an underlying factor and can be updated as well as the basis matrix $\mathbf{\Omega}$, the proposed supervised learning performance is determined by the constraint of shift-variant rule, while in [11] the performance of note training session is only concerned by the prior constraint of a note's onset-offset information in the matrix $\mathbf{H}$.

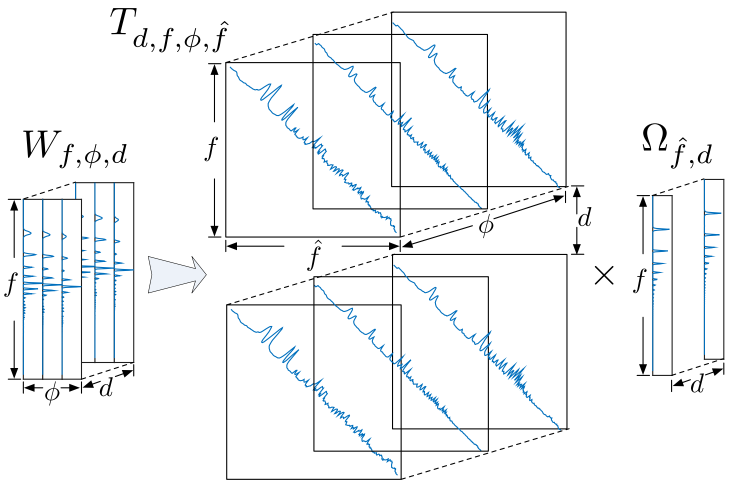
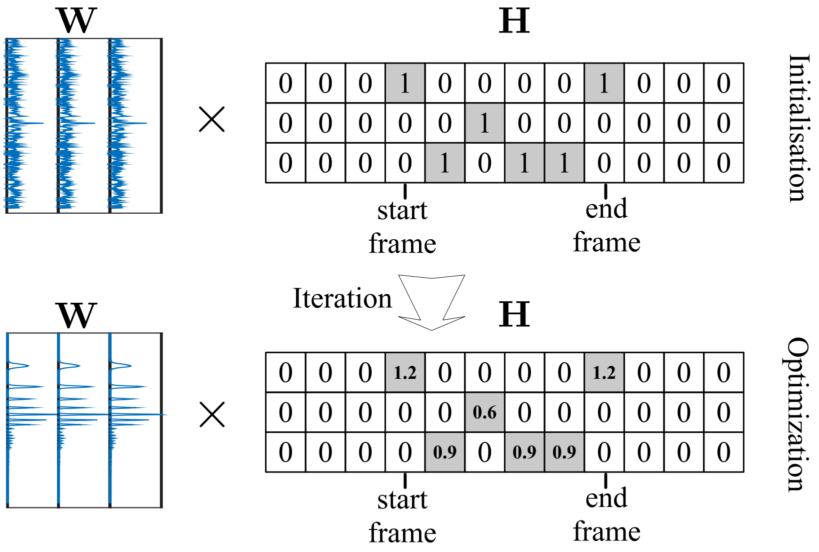


Figure 3.14: Supervised learning procedure.



0	0	0	1	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	1	1	0	0	0	0

0	0	0	1.2	0	0	0	0	1.2	0	0	0
0	0	0	0	0	0.6	0	0	0	0	0	0
0	0	0	0	0.9	0	0.9	0.9	0	0	0	0

Figure 3.15: Note training session in previous work [11].

3.4 Estimating the Onset-Offset-Like Features Within H

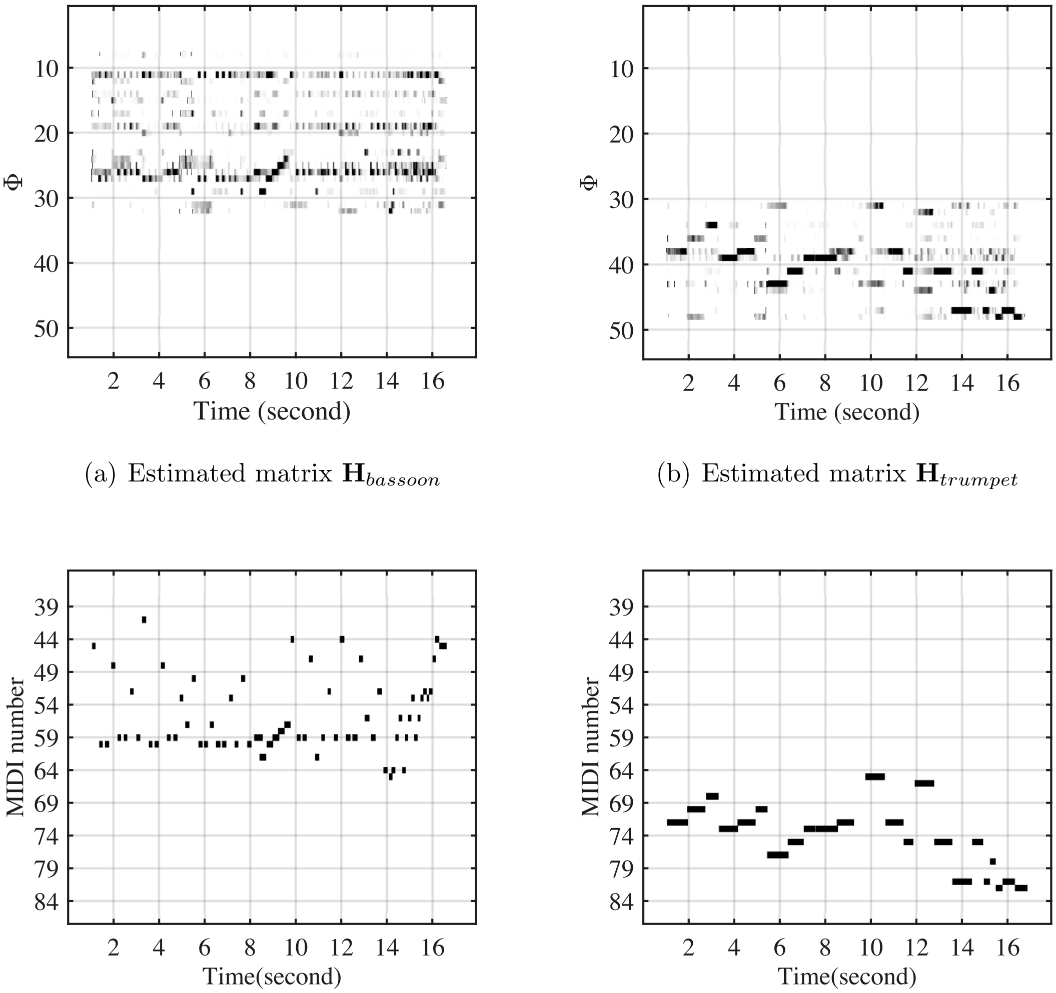
In practical situations, it is hard to determine the entries $H_{d,\phi,t}$ within a matrix $\mathbf{H}$ with an absence of synchronized information of the notes' attack and decay process. However, with the support of prior knowledge, we can obtain a rough estimate of a matrix $\mathbf{H}$ to narrow down its possible values. This estimate can be regarded as one kind of music transcription [11]. Though the note spectra of different F_0s extracted from datasets are not the same from recorded instruments, the supervised learnt $\mathbf{T}$ and $\mathbf{\Omega}$ are integrated with the note information which is similar to the actual recorded one. Combined with the results from [99, 100], the estimated values are considered to be probabilities of notes' triggered and decayed activation and also defined as the onset-offset-like features. Based on the concepts in (3.8), when the supervised learnt $\hat{\mathbf{W}}$ is obtained, based on multiplicative update rules [26–28], we can estimate the values within $\mathbf{H}$ by means of

$$\overset{n+1}{H}_{\phi,t,d} = \overset{n}{H}_{\phi,t,d}\frac{\sum_f V_{f,t}\overset{n}{\hat{W}}_{f,\phi,d}}{\sum_f \overset{n}{\hat{V}}_{f,t}\overset{n}{\hat{W}}_{f,\phi,d}}, \tag{3.23}$$

where $\overset{n}{\hat{V}}_{f,t}$ is the rebuilt $V_{f,t}$ by using the n^{th} iterated underlying factors and based on the concept in (3.9), the iterative estimation becomes

$$\overset{n+1}{H}_{\phi,t,d} = \overset{n}{H}_{\phi,t,d}\frac{\sum_{\hat{f}} \frac{V_{f,t}\overset{n}{\hat{W}}_{f,\phi,d}}{\overset{n}{\hat{V}}_{f,t}}}{\sum_{\hat{f}} \overset{n}{\hat{W}}_{f,\phi,d}}. \tag{3.24}$$

In this case, a music sample named “Lusser” provided by the TRIOS dataset [75] is used in estimating the onset-offset-like features from the activation matrix $\mathbf{H}$. This sample is an audio mixture of the bassoon and the trumpet. The estimated $\mathbf{H}_{bassoon}$ is presented in Fig. 3.16(a) and $\mathbf{H}_{trumpet}$ in Fig. 3.16(b). For a visual evaluation of this estimate, the prior knowledge of the notes' trigger and decay profile informed by the synchronized MIDI are also shown in 3.16(c) and 3.16(d). It is important to know that the onset-offset-like features of different F_0s are represented in ϕ over time



(a) Estimated matrix $\mathbf{H}_{bassoon}$

(b) Estimated matrix $\mathbf{H}_{trumpet}$

(c) Bassoon synchronized MIDI events of notes' attack and decay information in piano-roll form

(d) Trumpet synchronized MIDI events of notes' attack and decay information in piano-roll form

Figure 3.16: Comparison of the predicted matrix $\mathbf{H}$ and synchronized MIDI events. (Fs = 44100 Hz.)

but the synchronized MIDI informed note's trigger and decay profiles are organized in piano-roll style, and labelled by MIDI numbers. In the estimate results, parts of $H_{d,\phi,t}$ at the higher value represented in dark colour are located in the same areas as those informed by the synchronized MIDI, while other parts of the lower value represented in light colour are located in unexpected areas according to the synchronized MIDI. Ideally, if the prior instrument note templates and the actual recorded templates are exactly the same, the estimated values of $\mathbf{H}$ should completely overlap with the MIDI piano roll information. These conflicts are concluded from two reasons: first is that spectrograms within the prior instrument note templates are different from the actual recorded instrument note templates; the second is the misleading from the gradient descent oriented iteration rules, the multiplicative update rule in (3.23) and (3.24). The plan for solving the first problem comes forward in Sec. 3.4.3 and the plan for solving the second problem is raised in Sec. 3.4.1.

3.4.1 Onset-Offset Components Extraction

The estimated results obtained in Fig. 3.16(a) and 3.16(b) are explained in two aspects: from the first point of view, the high and low values within the estimated **H** represent possibilities of the note being activated or suppressed; from the second point of view, the estimated values are composed of useful information of onset-offset-like features and noise. Because it is hard to obtain the synchronized clue of a note being triggered and decayed, it is a challenge for us to determine which value is useful and which value is misleading. It is necessary to set up a preprocessing operand of estimated elements of $H_{d,\phi,t}$ for data de-noising, clustering and recognizing, so the next following steps of Gaussian blur, boundary detection and peak selection are developed to fit the needs.

Gaussian Smoothing for De-noising

We adopt the Gaussian blur method [112, 113] to smooth or reduce the noise and detail in the estimated $H_{d,\phi,t}$. The fact in Fig. 3.16(a) and 3.16(b) is that the estimated elements $H_{d,\phi,t}$ with high amplitude have higher probability of being onset-offset-features than the elements with low amplitude. Because of the wide use of the Gaussian blur algorithm in image processing [114–119], it is expected to enlarge the approximation and compress the detail in **H**. The Gaussian blur function is expressed as

$$G(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}}, \tag{3.25}$$

where t is the pixel involved in the smoothing operand. σ is the standard deviation and controls the width to contain the majority of data in a Gaussian distribution. If σ is too big, excessive noise would be clustered together to generate a fake peak. If σ is too small, the Gaussian blur function will lose the smoothing effect. In consideration of audio frequency sampling at 44100 Hz and step-size is 512 points in the spectrogram analysis, one bin of $H_{d,\phi,t}$ represents 11.61 ms. Based on experience, when σ reaches 50, the smoothing effect is satisfied by experiments. Because the column of $\mathbf{H}$ is determined by the step-size of STFT analysis for audio mixtures, the time of each column bin equals a time of samples in a step-size window. So the value of σ is various with the changes of step-size of STFT analysis. For easy viewing, one row of an estimated $\mathbf{H}$ at $d = bassoon$, $\phi = 8$ is picked up to demonstrate the results of a Gaussian smoothing operand. The calculation of smoothed $H_{d=bas,\phi=8,t}$ is determined by the formula of

$$H^{smoothed}_{d,\phi,t} = H^{estimated}_{d,\phi,t} \otimes G(t), \tag{3.26}$$

where $\otimes$ is to a convolution operation and output $H^{smoothed}_{d,\phi,t}$ is drawn in Fig. 3.17. The estimated result from Sec. 3.4 is given in Fig. 3.17(a) to illustrate the condition of a "noisy" estimate, where it is hard to distinguish onset-offset components from onset-offset-like features. The Gaussian smoothing results are shown out in Fig. 3.17(b) to show the approximation of the estimated $H_{d=bas,\phi=8,t}$ enveloped by a series of peak regions.

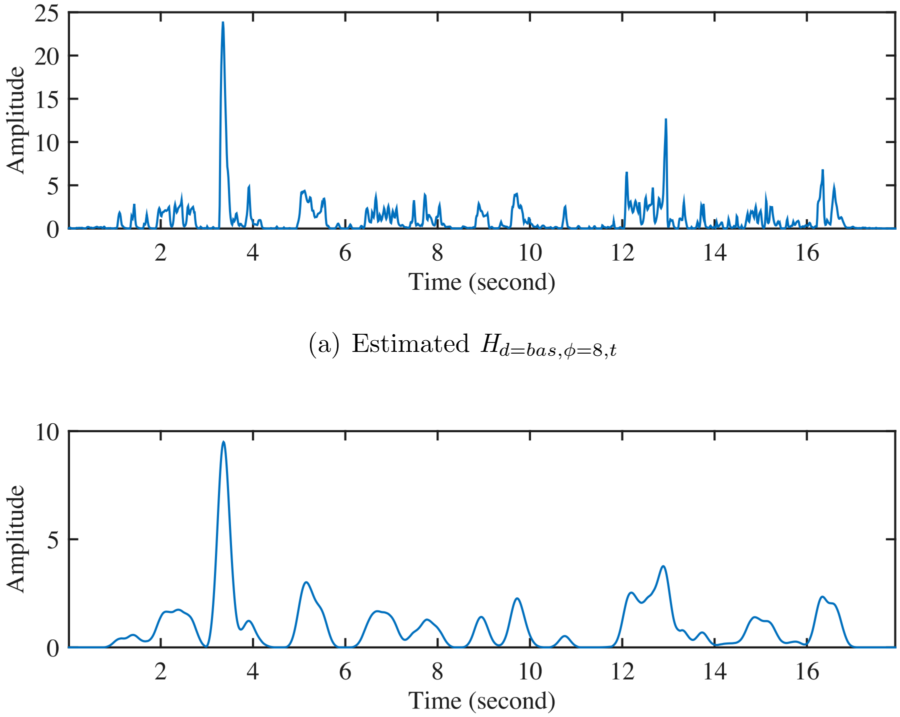
Boundary Detection for Regions

Though the Gaussian smoothed $H_{d=bas,\phi=8,t}$ gives an approximation of the notes' trigger-decay profile, it is seen as a temporary and not final solution of the separation problem. In Fig. 3.17(b), the smoothed profiles reflect that the note activation is composed of a series of peak and each peak region is assumed as a completed onset-offset process. If the approximation can be decomposed in regions, it would be easy for us to examine the estimated values and find misleading ones. For this purpose, the Laplace-of-Gaussian (LoG) factor [112,120–122] algorithm is employed for a boundary detection. Its definition as a 1-dimensional function starts from a Gaussian kernel with width of σ_{LOG},

$$G_{\sigma_{LoG}}(t) = \frac{1}{\sqrt{2\pi\sigma^2_{LoG}}} exp\left(-\frac{t^2}{2\sigma^2_{LoG}}\right), \tag{3.27}$$

and to further suppress the noise before using LoG for boundary detection

$$\begin{aligned}\Delta[G_{\sigma_{LoG}}(t) \otimes H^{smoothed}_{d,\phi,t}] &= [\Delta G_{\sigma_{LoG}}] \otimes H^{smoothed}_{d,\phi,t} \\ &= LoG \otimes H^{smoothed}_{d,\phi,t}.\end{aligned} \tag{3.28}$$



(a) Estimated $H_{d=bas,\phi=8,t}$

(b) Smoothed $H_{d=bas,\phi=8,t}$

Figure 3.17: Gaussian smoothing results. (Fs = 44100 Hz.)

The tricky way of getting results from (3.28) is that we can calculate the partial differential of $G_{\sigma_{LoG}}$ instead of calculating the partial differential of the convolution $G_{\sigma_{LoG}} \otimes H^{smoothed}_{d,\phi,t}$. So, firstly we consider that

$$\frac{\partial}{\partial t} G_{\sigma_{LoG}}(t) = \frac{-t}{\sqrt{2\pi\sigma^2}\sigma^2} exp\left(-\frac{t^2}{2\sigma^2}\right), \tag{3.29}$$

and then

$$\frac{\partial^2}{\partial^2 t} G_{\sigma_{LoG}}(t) = \frac{-t^2}{\sqrt{2\pi\sigma^2}\sigma^4} exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.30}$$

From now on, we get the LoG as an operator defined as

$$LoG \triangleq \frac{\partial^2}{\partial^2 t} G_{\sigma_{LoG}}(t). \tag{3.31}$$

When LoG is convolved with the smoothed $H_{d,\phi,t}$, after zero-crossing detection, the detected boundaries are marked by vertical lines and drawn together with the smoothed $H_{d,\phi,t}$ in Fig. 3.17. The effects of LoG decomposes the Gaussian blurred $H_{d,\phi,t}$ into several peak regions which are limited by a pair of vertical lines and defined as an independent trigger-decay process. For the example in Fig. 3.17, the smoothed $H_{d,\phi,t}$ is divided into 17 regions indicating the individual onset-offset-like feature. It is still hard for us to determine which estimated elements in $H_{d,\phi,t}$ are actual onset-offset-features, so a further step is needed for judging the note alignment which is clustered by estimated elements in $H_{d,\phi,t}$.

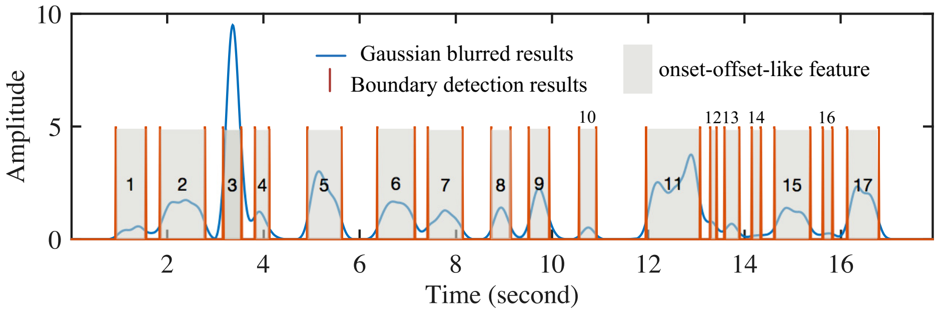


Figure 3.18: Boundary detection results. (Fs = 44100 Hz.)

Peak Detection in Each Region

Before the work of making a judgment on the peak alignment, peak detection is necessarily employed here to determine an "attack" peak within a peak's alignment. A peak value exhibits the time when a note is activated at its maximum level. If a representative peak is determined, its occurance time is denoted as the top in the "attack" period of an isolated note [123]. The reason for finding the "attack" peak is that its moment in the audio mixture gives evidence to prove whether this alignment is a misleading estimate. But for the playing of some wind and string instruments, there would be several peaks in an isolated note attack-decay profile during its activation. The prominence becomes a peak measurement in "attack" peak selection, which is a local maximum and depends on how much the peak exceeds others because of its inherent height and its relative position [124]. The main task of the peak detection operation is to find a local maximum peak in an independent alignment. Fig. 3.19 illustrates the final peak finding results and marks their top three peaks. These top three peaks are the local maxima peak in 3^{rd}, 11^{th} and 17^{th} regions separately. Their occurrance time will be the key issue and highlighted in the audio mixture STFT spectrogram to help the work of bias correction of multiplicative update rules.

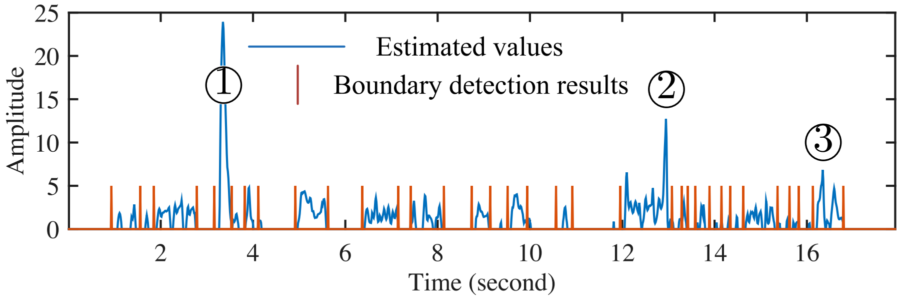


Figure 3.19: Local maxima peak detection results. (Fs = 44100 Hz.)

3.4.2 Pitch-checking for Onset-offset-like Features

For estimated $H_{d,\phi,t}$, the Gaussian blur, the *LoG* boundary detection and the local maxima peak selection algorithms are the preprocessing operations for onset-offset-like features recognition. Though they all describe the activation of the notes in the time line, there is still a difference between onset-offset-like and onset-offset features. From a mathematical point-of-view, these estimated trigger-decay profiles are the maximum likelihood estimation of note activation and still have the uncertainty of being onset-offset features in reality. If the onset-offset-like features can be further checked and corrected based on some evidence, they will become closer to the onset-offset features of actual recorded condition.

The examination and correction work begins at the results of local maxima peak detection in Fig. 3.19. For a simplified example, the top three local maxima peaks are marked and their corresponding time locations are also labelled by t_1, t_2 and t_3 with highlight vertical lines in Fig. 3.20. In an audio mixture spectrogram, each column represents a spectrum at a window of time. If the spectrogram can be denoted as $V_{f,t}$, the symbols of V_{f,t_1}, V_{f,t_2} and V_{f,t_3} are used to indicate the three spectra in time of t_1, t_2 and t_3. Based on the basic definition of NMF in (3.4), for the first spectrum in Fig. 3.20, we have

$$V_{f,t_1} \approx W_{f,\phi,d=bas} \cdot H_{d=bas,\phi,t} + W_{f,\phi,d=trp} \cdot H_{\phi,t,d=trp}. \tag{3.32}$$

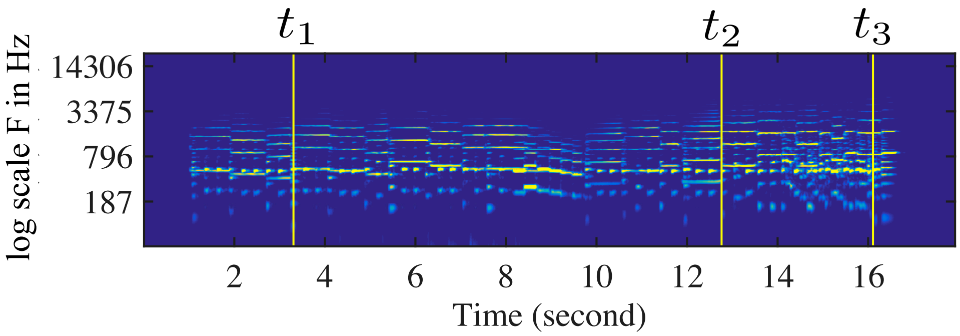


Figure 3.20: Spectrogram of mixture audio with labels of t_1, t_2 and t_3. (Fs = 44100 Hz.)

If $W_{f,\phi,d=trp}$ or $H_{\phi,t,d=trp}$ equals zero, the relationship in (3.32) becomes

$$V_{f,t_1} \approx W_{f,\phi,d=bas} \cdot H_{d=bas,\phi,t}. \tag{3.33}$$

If we only chose $\phi = 8$, (3.33) turns out to be

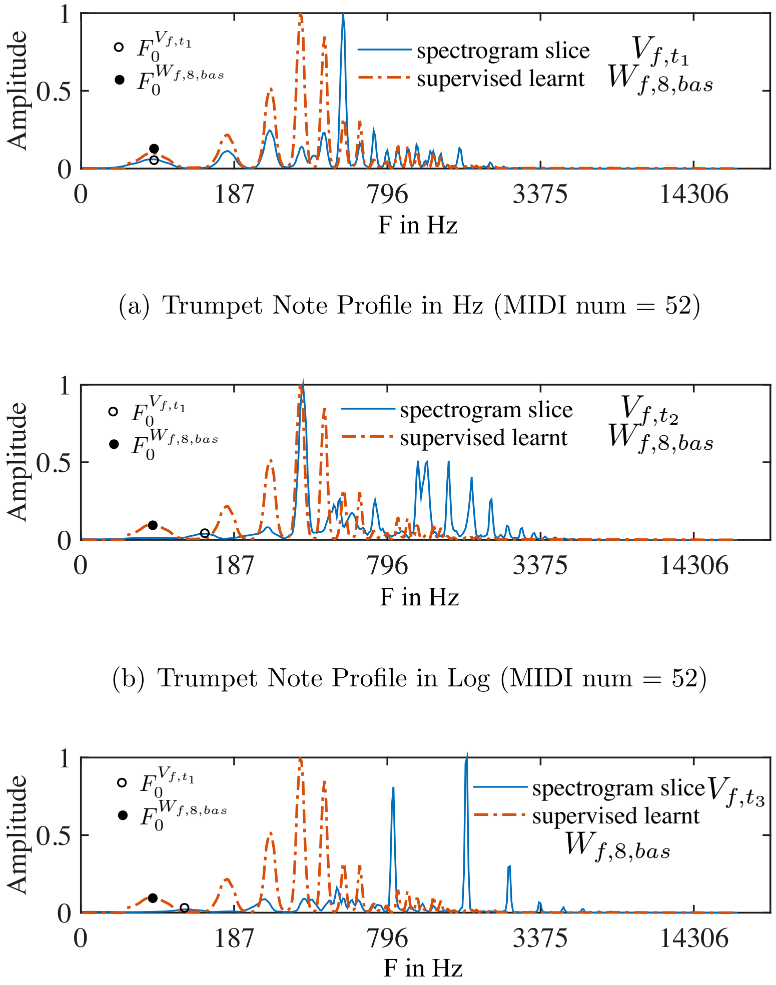
$$V_{f,t_1} \supseteq W_{f,\phi=8,d=bas} \cdot H_{d=bas,\phi=8,t}. \tag{3.34}$$

If $H_{d=bas,\phi=8,t}$ is a scalar to amplify or compress, the amplitude of $W_{f,\phi=8,d=bas}$, (3.34) is further derived into

$$V_{f,t_1} \supseteq W_{f,\phi=8,d=bas}. \tag{3.35}$$

A lesson from the derivation of (3.35) says that the envelop of V_{f,t_1} should contain $W_{f,\phi=8,d=bas}$ and this $\supseteq$ relation is demonstrated in Fig. 3.21, where the fundamental frequency of V_{f,t_1} is denoted as $F_0^{V_{f,t_1}}$ and the fundamental frequency of $W_{f,\phi=8,d=bas}$ is denoted as $F_0^{W_{f,8,bas}}$. The definition of $\supseteq$ means a containing or an involving. In Fig. 3.21(a), the peaks of components within $W_{f,\phi=8,d=bas}$ are coherently overlapped by the peaks in V_{f,t_1}, especially, their pitches $F_0^{V_{f,t_1}}$ and $F_0^{W_{f,8,bas}}$ are equal in frequency value. But when $t = t_2$ and $t = t_3$, the $\supseteq$ relation is broken and these phenomenons are illustrated in Fig. 3.21(b) and 3.21(c). In these situations, we can find partial areas of overlap between these two kinds of spectrum, but prominently their in-equivalent F_0s state that V_{f,t_2} is not a superset of $W_{f,\phi=8,d=bas}$ as well as the condition of V_{f,t_3}. The relationship between $\supseteq$ and $\not\supseteq$ suggest that $W_{f,\phi=8,d=bas}$'s activation $H_{d=bas,\phi=8,t_1}$ happening at time t_1 is more reliable than another $H_{d=bas,\phi=8,t}$ happening at time t_2

and t_3, though they are all acceptable from a mathematical point of view. The pitch-check based examination has the characteristic that it cannot verify the correct $H_{d,\phi,t}$ but it can falsify the misleading parts. Meanwhile, the pitch-check based examination operations can be adopted as evidence to force the misleading parts into zeros and supervise the evolution of $\mathbf{H}$.



(a) Trumpet Note Profile in Hz (MIDI num = 52)

(b) Trumpet Note Profile in Log (MIDI num = 52)

(c) Trumpet Note Profile in Log (MIDI num = 52)

Figure 3.21: Pitch-check examination (Fs = 44100 Hz, MIDI NO. = 52.)

Why might the estimated $H_{d,\phi,t}$ given by the multiplicative algorithm [26–28] not provide the expected solution in separation? One reason is that in convex optimization, the gradient descent algorithm working on the negative direction of gradient increments will not consider the rules of (3.34) and (3.35). If there exists a large area of overlap with $V_{f,t}$, $W_{f,\phi,d}$, it has a high probability of being marked as the activated state in $H_{d,\phi,t}$ without any consideration of the superset relationship of frequency components between $V_{f,t}$ and $W_{f,\phi,d}$. So we need an examination algorithm to check on this situation and the simplest way is to detect whether the pitch component of $W_{f,\phi,d}$ is included by $V_{f,t}$. Therefore, the pitch-checking examination algorithm introduced here does not have a capability to prove that the selected estimated $H_{d,\phi,t}$ is correct but has the ability to show it is incorrect and can be eliminated. According to the pitch-checking examination theory, for the example results in Fig. 3.21(a), after the pitch-checking examination, the corrected $H_{d,\phi,t}$ is illustrated in Fig. 3.22. Only the peaks-alignment region around $H_{d=bas,\phi=8,t_1}$ remains and the rest of the parts of $H_{d=bas,\phi=8,t\neq t_1}$ are set to zeros. Based on the multiplicative iteration theory, only the information expressed by non-zero values can be relayed by evolving the other underlying factors.

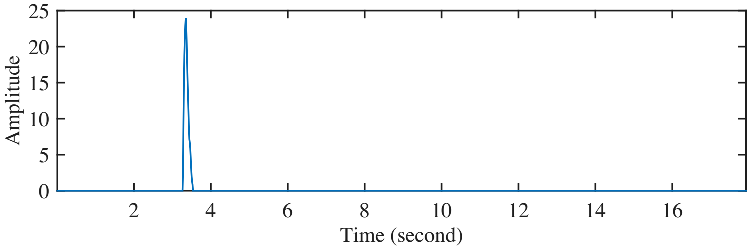


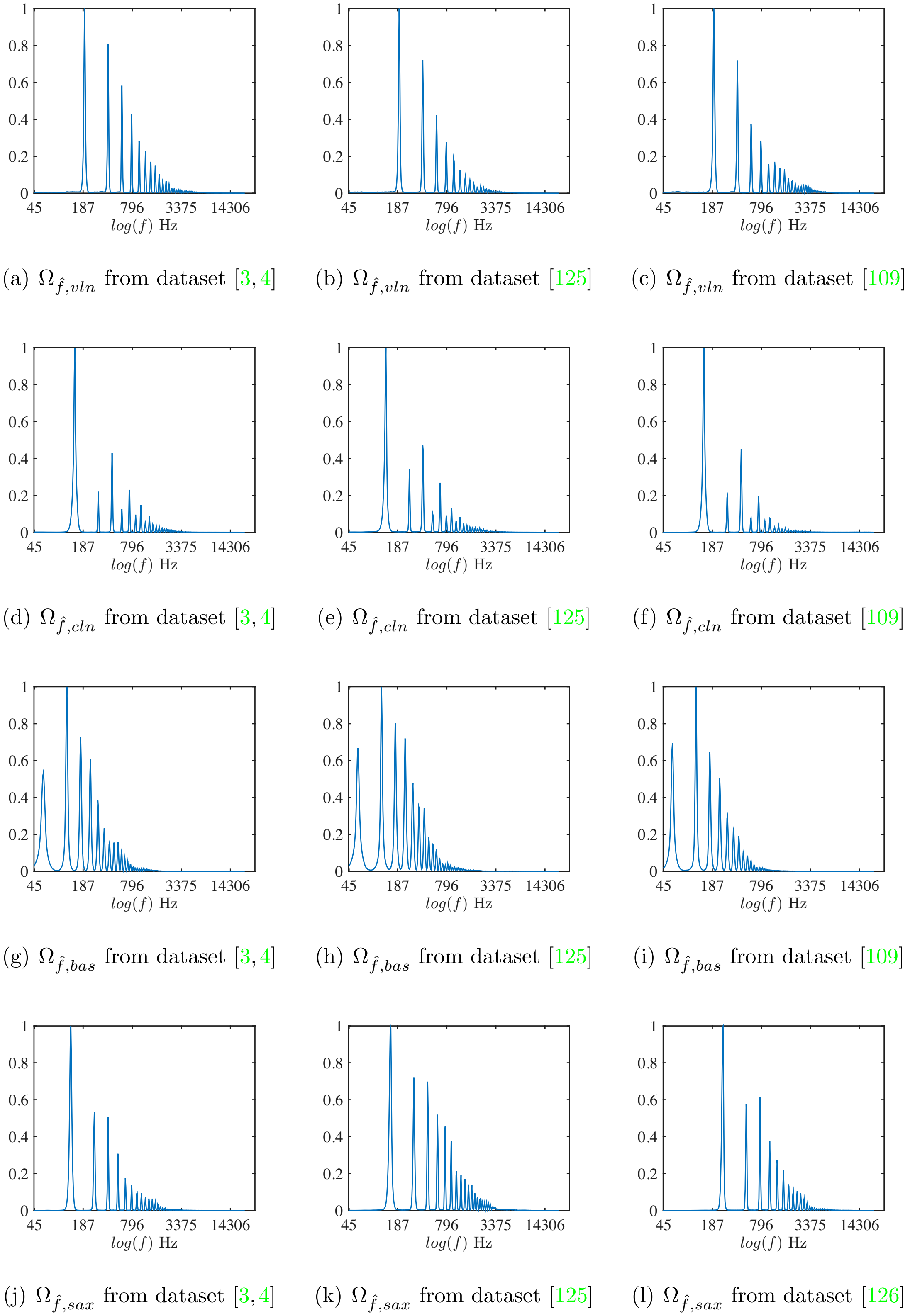
Figure 3.22: Corrected $H_{d,\phi,t}$ after pitch-checking examination (Fs = 44100 Hz, MIDI NO. = 52.)

3.4.3 Evolving of Supervised Learnt Notes Spectra

The main goal of our proposed algorithm is to try to find the actual recorded instruments' note spectrum template of different F_0s. It needs to evolve from the prior informed note spectrum template. This evolving operation is different from the previous iterative updates in supervised learning operation in (3.19), (3.20), (3.21) and (3.22), which is a plan of pursuing the optimized $T_{d,f,\phi,\hat{f}}$ and $\Omega_{\hat{f},d}$ to get the reconstruction of a prior $W_{f,\phi,d}$. But in this section, the reconstructed target is the spectrogram $V_{f,t}$ of the observed audio mixture and the reconstructed underlying factors include $T_{d,f,\phi,\hat{f}}$, $\Omega_{\hat{f},d}$ and $H_{d,\phi,t}$. Referring to the iterative update rules in the supervised learning operation, if we had a fixed $V_{f,t}$ and a corrected $H_{d,\phi,t}$, the rest are updating $T_{d,f,\phi,\hat{f}}$ or $\Omega_{\hat{f},d}$.

The evolution of $W_{f,\phi,d}$ is determined by updating the underlying factors $T_{d,f,\phi,\hat{f}}$ and $\Omega_{\hat{f},d}$. Three methods can be adopted for the updating. The first way is to update both $T_{d,f,\phi,\hat{f}}$ and $\Omega_{\hat{f},d}$; the second way is to update $T_{d,f,\phi,\hat{f}}$ only; and the final way is presented as the updating of $\Omega_{\hat{f},d}$. For the first plan, if $T_{d,f,\phi,\hat{f}}$ and $\Omega_{\hat{f},d}$ are both changed, the prior information from the dataset based on the note spectrum template will be discarded and the shift-variant problem becomes an unconstrained optimization procedure. So we should make a selection from $T_{d,f,\phi,\hat{f}}$ or $\Omega_{\hat{f},d}$ to update. In our proposed algorithm, we assume that in shift-variant rule, $W_{f,\phi,d}$, the score spectra of different F_0s is generated from the shift-variant copy of the basis F_0 template $\Omega_{\hat{f},d}$ and also assume that combining with the shift-variant translation matrix $T_{d,f,\phi,\hat{f}}$ they construct a source template model (3.10) for the instrument. If the $T_{d,f,\phi,\hat{f}}$ represents the resonance coefficients and $\Omega_{\hat{f},d}$ refers to the excitation impulses, that means $\Omega_{\hat{f},d}$ should have a general form for other cases. By analyzing the individual note audio labelled by instrument from datasets [3, 4, 109, 125, 126], we present the forms of $\Omega_{\hat{f},d}$ in Fig. 3.23. Though the isolated note audio comes from different datasets and definitely distinct source, if the audio from the same category instrument, their impulse excitation $\Omega_{\hat{f},d}$ should have similarities in shape. For example, Fig. 3.23(a), 3.23(b) and 3.23(c) are note audio of violins from three different datasets, but they belong to instruments of violins and their $\Omega_{\hat{f},d}$ have similarities in frequency envelop and are easily distinguished from the instruments of the bassoon, clarinet and saxophone. These characteristics are summarized in that the $\Omega_{\hat{f},d}$ from the instruments of one kind spans their own space which can be distinguished by the $\Omega_{\hat{f},d}$ from the instruments of other kinds. We verify our assumption using principal component analysis (PCA) [127–131]. If the relative weights of n harmonics in $\Omega_{\hat{f},d}$ is formed as a column vector, the m different $\Omega_{\hat{f},d}$s are arranged together as a matrix $\mathbf{X}$ with size of $m \times n$ and each vector per column. After its data adjusted operations deduct the mean value of each vector, its covariance matrix can be calculated as

$$\mathbf{\Sigma} = \begin{pmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_n) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_n) \\ \cdots & \cdots & \cdots & \cdots \\ cov(x_n, x_1) & cov(x_n, x_2) & \cdots & cov(x_n, x_n) \end{pmatrix}. \tag{3.36}$$



(a) $\Omega_{\hat{f},vln}$ from dataset [3, 4] (b) $\Omega_{\hat{f},vln}$ from dataset [125] (c) $\Omega_{\hat{f},vln}$ from dataset [109]

(d) $\Omega_{\hat{f},cln}$ from dataset [3, 4] (e) $\Omega_{\hat{f},cln}$ from dataset [125] (f) $\Omega_{\hat{f},cln}$ from dataset [109]

(g) $\Omega_{\hat{f},bas}$ from dataset [3, 4] (h) $\Omega_{\hat{f},bas}$ from dataset [125] (i) $\Omega_{\hat{f},bas}$ from dataset [109]

(j) $\Omega_{\hat{f},sax}$ from dataset [3, 4] (k) $\Omega_{\hat{f},sax}$ from dataset [125] (l) $\Omega_{\hat{f},sax}$ from dataset [126]

Figure 3.23: $\Omega_{\hat{f},d}$ forms from different datasets by instrument (vln = violin, cln = clarinet, bas = bassoon, sax = saxophone)

In (3.36), if the expectation of x_i is denoted as $\mu_i = E(x_i)$, then $cov(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)]$. Based on the values of $\mathbf{\Sigma}$, its eigenvalues are labelled as λ_1 through λ_p and ordered from largest to smallest in the form of

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p. \tag{3.37}$$

Then their corresponding eigenvectors should be arranged as

$$\mathbf{e}_1 > \mathbf{e}_2 > \cdots > \mathbf{e}_p. \tag{3.38}$$

We can calculate the i^{th} PCA coefficient $coef_i$ as

$$coef_i = \mathbf{X} \times \mathbf{e}_i^{'}, \tag{3.39}$$

where $'$ is the matrix transposition. By obtaining the top 3 $coef_i$, we mapped the $\Omega_{\hat{f},d}$s in Fig. 3.23 into the points in Fig. 3.24. In this figure, the points that belong to the instrument of the same kind have a trend of being clustered together as one group. That means that the $\Omega_{\hat{f},d}$s labelled by the same kind of source will have similarities in forms and keeps its shape stable within the group.

Evolution Rule of Supervised Learnt Notes Spectra

Since the samples from the standard datasets are individual note recordings of limited kinds of instruments, we can only get research on $\Omega_{\hat{f},d}$ of a few instruments. For example, the $\Omega_{\hat{f},d}$ in Fig. 3.23(j), Fig. 3.23(k), and Fig. 3.23(l) comes from different types of saxophones, and have little difference in shape of frequency profiles. However, the illustrations in Fig. 3.23 and 3.24 provide preliminary support for our assumption that $T_{d,f,\phi,\hat{f}}$ acts as a resonance character and $\mathit{\Omega}_{\hat{f},d}$ acts as an impulse excitation which keeps its form during the separation procedure. Therefore, the evolution of $\hat{W}_{f,\phi,d}$ is only involved in the iterative update of $T_{d,f,\phi,\hat{f}}$. To accomplish this goal, we suppose a two procedure structure in our proposed source template NMF algorithm in Fig. 3.25, which are supervised learning and separation procedures. In the first procedure, we need to evolve the underlying factors $T_{d,f,\phi,\hat{f}}$ and $\Omega_{\hat{f},d}$ to reconstruct the prior setting $W_{f,\phi,d}$. After the $\Omega_{\hat{f},d}$ inherits partial side-information from the supervised learning operand, it switches to update underlying factors $T_{d,f,\phi,\hat{f}}$ and $H_{d,\phi,t}$ to reconstruct the

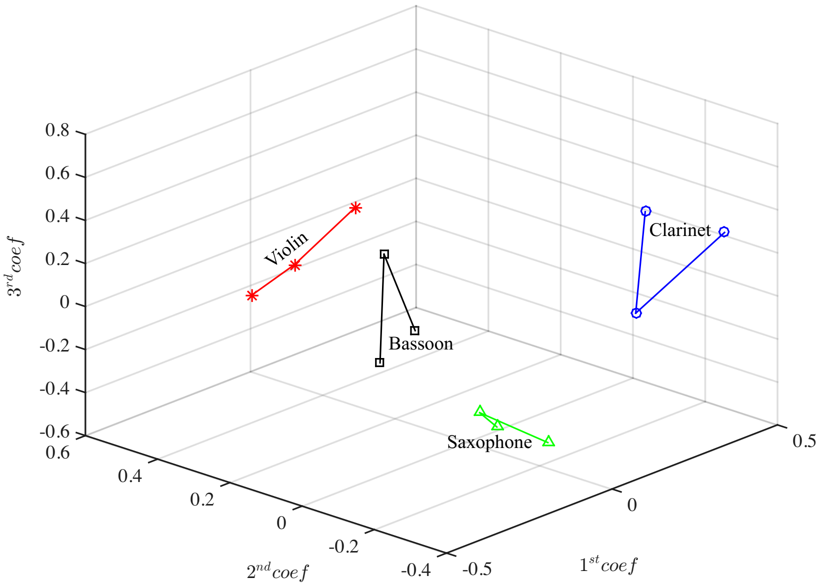


Figure 3.24: Principal component analysis (PCA) on $\Omega_{\hat{f},d}$ for different instruments from different datasets

targeted $V_{f,t}$.

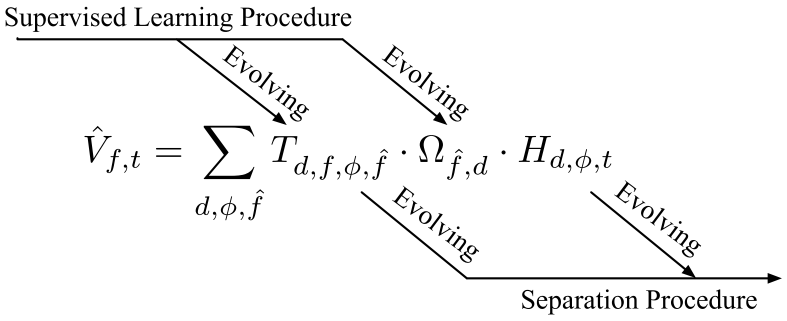

Supervised Learning Procedure
Evolving
Evolving
$\hat{V}_{f,t} = \sum_{d,\phi,\hat{f}} T_{d,f,\phi,\hat{f}} \cdot \Omega_{\hat{f},d} \cdot H_{d,\phi,t}$
Evolving
Evolving
Separation Procedure

Figure 3.25: Illustration of underlying factors in different procedures

On the basis of (3.10), (3.8) and (3.9), the multiplicative update rules of shift-variant translation matrix $T_{d,f,\phi,\hat{f}}$ in the separation procedure should be

$$\overset{n+1}{T}_{d,f,\phi,\hat{f}} = \overset{n}{T}_{d,f,\phi,\hat{f}} \cdot \frac{\sum_t V_{f,t} \cdot \overset{n}{\Omega}_{\hat{f},d} \cdot \overset{n}{H}_{\phi,t,d}}{\sum_t \overset{n}{\hat{V}}_{f,t} \overset{n}{\Omega}_{\hat{f},d} \cdot \overset{n}{H}_{\phi,t,d}}, \tag{3.40}$$

and

$$\overset{n+1}{T}_{d,f,\phi,\hat{f}} = \overset{n}{T}_{d,f,\phi,\hat{f}} \cdot \frac{\sum_t V_{f,t} \frac{\overset{n}{\Omega}_{\hat{f},d} \cdot \overset{n}{H}_{\phi,t,d}}{\overset{n}{\hat{V}}_{f,t}}}{\sum_t \overset{n}{\hat{V}}_{f,t} \overset{n}{\Omega}_{\hat{f},d} \cdot \overset{n}{H}_{\phi,t,d}}. \tag{3.41}$$

3.4.4 Specific Channel Audio Reconstruction

For the i^{th} source, the specific STFT is obtained by applying the source template NMF on

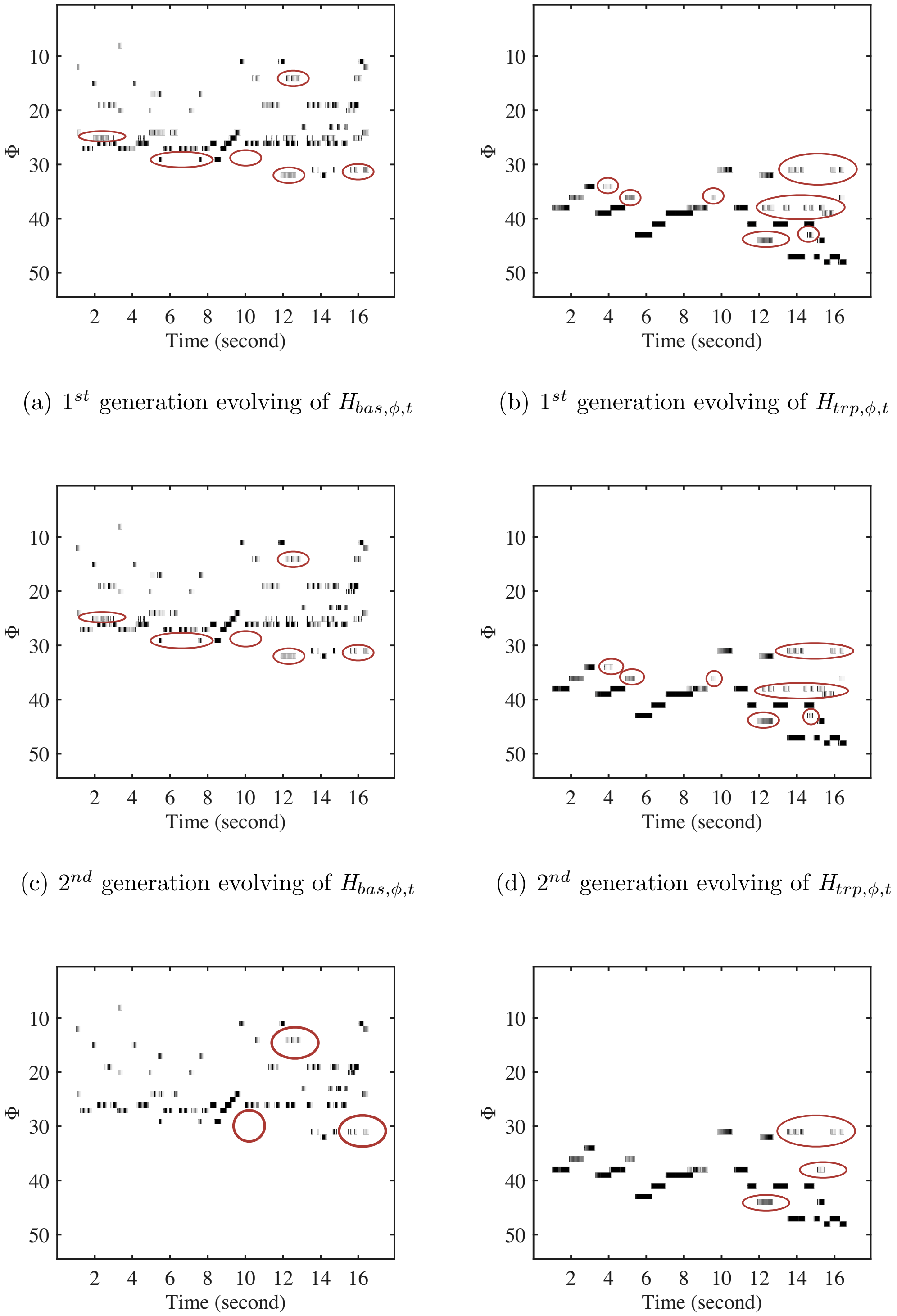
$$\hat{V}_{f,t,d} = \sum_{\hat{f},\phi} T_{d,f,\phi,\hat{f}} \cdot \Omega_{\hat{f},d} \cdot H_{d,\phi,t}. \tag{3.42}$$

With (3.42), the reconstruction of the specific channel audio magnitude spectrogram needs the optimized $T_{d,f,\phi,\hat{f}}$ and $H_{d,\phi,t}$ calculated by the alternative iterations in the separation procedure. The algorithm of this part is given in Algorithm 1.

Algorithm 1: $T_{d,f,\phi,\hat{f}}$ and $H_{d,\phi,t}$ Optimization in Separation

```
Data: magnitude spectrogram of the audio mixture $V_{f,t}$, supervised learnt
      $T_{d,f,\phi,\hat{f}}$ and pitch-checking corrected $H_{d,\phi,t}$
Result: Optimized $T_{d,f,\phi,\hat{f}}$ and $H_{d,\phi,t}$
$\hat{V}_{f,t} \leftarrow \sum_{\hat{f},\phi} T_{d,f,\phi,\hat{f}} \cdot \Omega_{\hat{f},d} \cdot H_{d,\phi,t}$;
$D_{\mathcal{E}} \leftarrow \hat{V}_{f,t}, V_{f,t}$;
while $D_{\mathcal{E}}$ does not reach the criterion do
    while $T_{d,f,\phi,\hat{f}}$ does not go convergence in this sub-while do
        $\overset{n+1}{T}_{d,f,\phi,\hat{f}} \leftarrow \overset{n}{T}_{d,f,\phi,\hat{f}} \cdot \frac{\sum_t V_{f,t} \cdot \overset{n}{\Omega}_{\hat{f},d} \cdot \overset{n}{H}_{\phi,t,d}}{\sum_t \overset{n}{\hat{V}}_{f,t} \overset{n}{\Omega}_{\hat{f},d} \cdot \overset{n}{H}_{\phi,t,d}}$;
    while $H_{d,\phi,t}$ does not go convergence in this sub-while do
        $\overset{n+1}{H}_{\phi,t,d} \leftarrow \overset{n}{H}_{\phi,t,d} \frac{\sum_f V_{f,t} \overset{n}{\hat{W}}_{f,\phi,d}}{\sum_f \overset{n}{\hat{V}}_{f,t} \overset{n}{\hat{W}}_{f,\phi,d}}$;
        $\overset{n+1}{H}_{\phi,t,d} \leftarrow pitch.checking\left(\overset{n+1}{H}_{\phi,t,d}\right)$
    $\hat{V}_{f,t} \leftarrow \sum_{\hat{f},\phi} T_{d,f,\phi,\hat{f}} \cdot \Omega_{\hat{f},d} \cdot H_{d,\phi,t}$;
    $D_{\mathcal{E}} \leftarrow \hat{V}_{f,t}, V_{f,t}$;
```

Evolving $H_{d,f,\phi,\hat{f}}$



(a) 1^{st} generation evolving of $H_{bas,\phi,t}$

(b) 1^{st} generation evolving of $H_{trp,\phi,t}$

(c) 2^{nd} generation evolving of $H_{bas,\phi,t}$

(d) 2^{nd} generation evolving of $H_{trp,\phi,t}$

(e) 3^{rd} generation evolving of $H_{bas,\phi,t}$

(f) 3^{rd} generation evolving of $H_{trp,\phi,t}$

Figure 3.26: Evolution of **H** in proposed algorithm

It is hard for us to verify the evolution of $T_{d,f,\phi,\hat{f}}$ with the absence of the notes spectra of the actual recorded instrument. However, it is easy to obtain the prior knowledge of the synchronized scores event from dataset [75], we give the verification of evolving $H_{d,\phi,t}$ of each sub-while loop in Fig. 3.26.

Following the estimated $H_{d,\phi,t}$ in Fig 3.16(a) and 3.16(b), the $\overset{n}{H}_{d,\phi,t}$ are corrected by the pitch-checking method and involved in evolving $\overset{n+1}{T}_{d,f,\phi,\hat{f}}$. Alternatively, the evolved $\overset{n+1}{T}_{d,f,\phi,\hat{f}}$ is employed to determine the first generation $\overset{n+1}{H}_{d,\phi,t}$ in Fig. 3.26(a) and 3.26(b). Compared with the prior synchronized score events, the misleading parts are circled in solid line for easy observation. In the next step, the corrected $\overset{n+1}{H}_{d,\phi,t}$ gives its contributions in an evolved $\overset{n+1}{T}_{d,f,\phi,\hat{f}}$ and repeats the alternative update rules, the second and third generations $\overset{n+2}{H}_{d,\phi,t}$ and $\overset{n+3}{H}_{d,\phi,t}$ are represented in Fig. 3.26(c)-3.26(f). With generations increasing, the misleading parts are reduced and the evolved $H_{d,\phi,t}$ becomes close to prior synchronized score events.

3.4.5 Audio Reconstruction

Channel Masking Template

In the proposed algorithm, the typical way of time-frequency decomposition, the spectrogram of the audio mixture is the sum of multi spectrogram of individual audio source based on their magnitude which are the absolute results from complex values. This implies that factoring a magnitude spectrogram into the sum of additive independent components leads to a partially missing phase of the information of the targeted spectrogram [64–66, 132]. This inaccurate reconstruction on phase information is acceptable here by using a masking template M_d with the form of

$$M_d = \frac{\hat{V}_{f,t,d}}{V_{f,t} + \epsilon}, \tag{3.43}$$

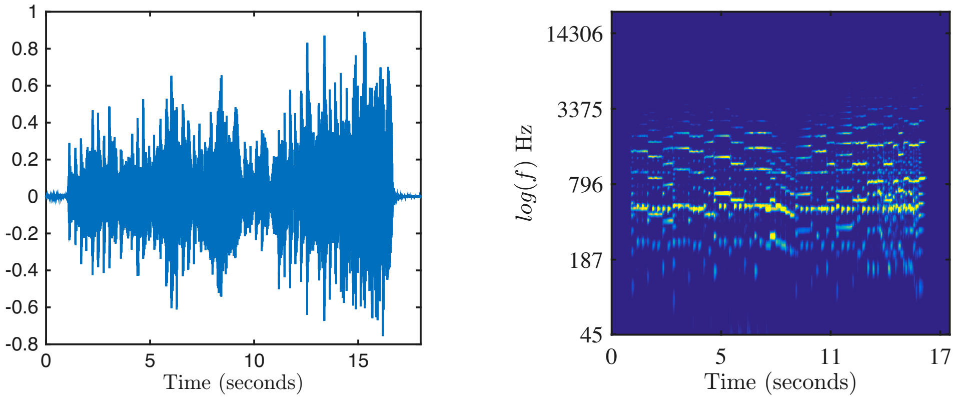
where ϵ is a small positive number to avoid a zero value in denominator. Then the reconstructed complex spectrogram of a specific channel can be obtained by

$$\dot{\hat{V}}_{f,t,d} = M_d \cdot \dot{V}_{f,t}, \tag{3.44}$$

where $\dot{V}_{f,t}$ is the observed complex spectrogram of the audio mixture and $\dot{\hat{V}}_{f,t,d}$ is the reconstructed complex spectrogram of the specific channel.

Separated Magnitude Spectrogram Reconstruction

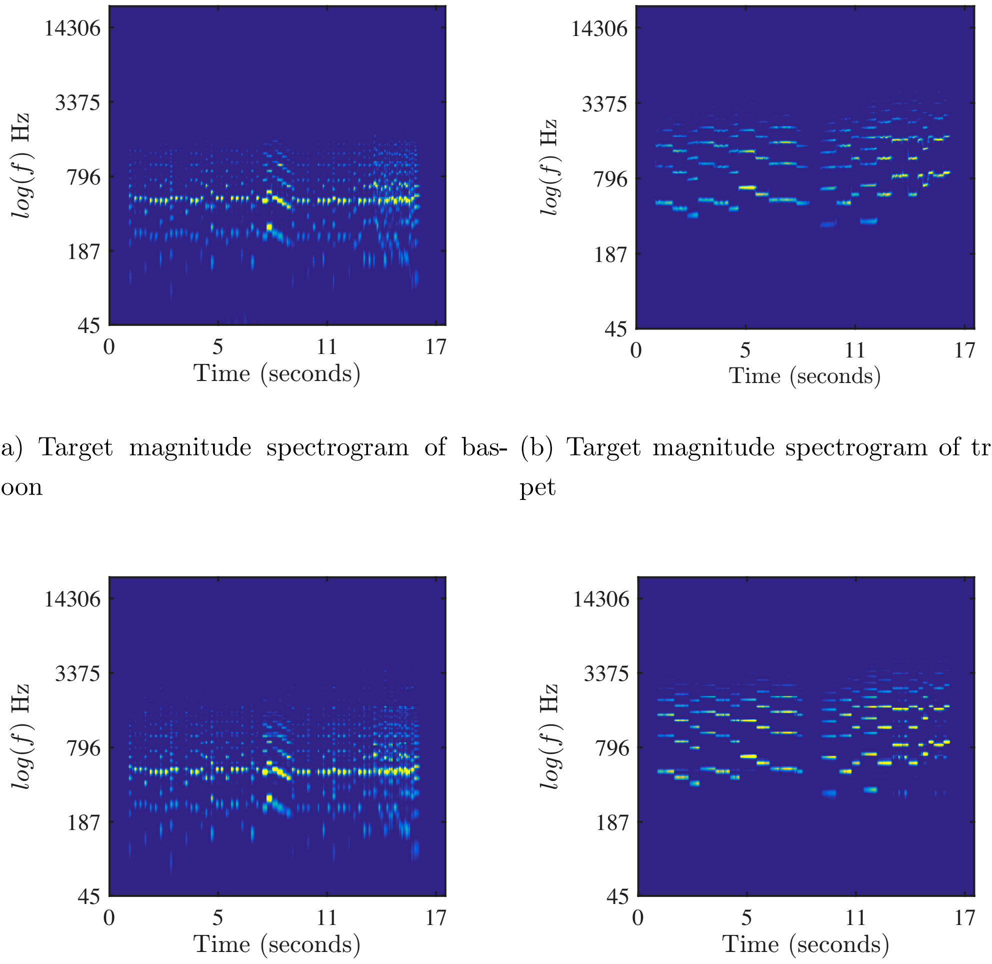
The music sample named “Lussier” and the audio mixture of the bassoon and trumpet from the dataset [75] is selected to perform the separation procedure of the proposed source template NMF. Its waveform and magnitude spectrogram are both plotted in Fig. 3.27. Fig. 3.27(a) shows that it is a mixed audio lasting for over 17 seconds at a sampling frequency of 44100 Hz and Fig. 3.27(b) illustrates its magnitude spectrogram by using 4096-points per frame for a window, 4096-points per frame for fast Fourier transform (FFT), and 512-points for step-size. In particular, each bin in $log(f)$ means $2^{\frac{1}{48}}$ per octave.



(a) Audio mixture of bassoon and trumpet (b) Target magnitude spectrogram of audio mixture

Figure 3.27: Waveform and magnitude spectrogram of the audio mixture “Lussier”

The separated magnitude spectrogram of the bassoon and trumpet channel are given in Fig. 3.28(c) and 3.28(d). We also give the target magnitude spectrogram of the bassoon and trumpet channels in Fig. 3.28(a) and 3.28(b) for esay comparison. In overview, the separated magnitude spectrograms give a high accuracy reconstruction of the target one. But in some details, the intensity of the separated track is a little different from the target track. This phenomenon is the result of the linear separation operation on a additive sound mixture.



(a) Target magnitude spectrogram of bassoon

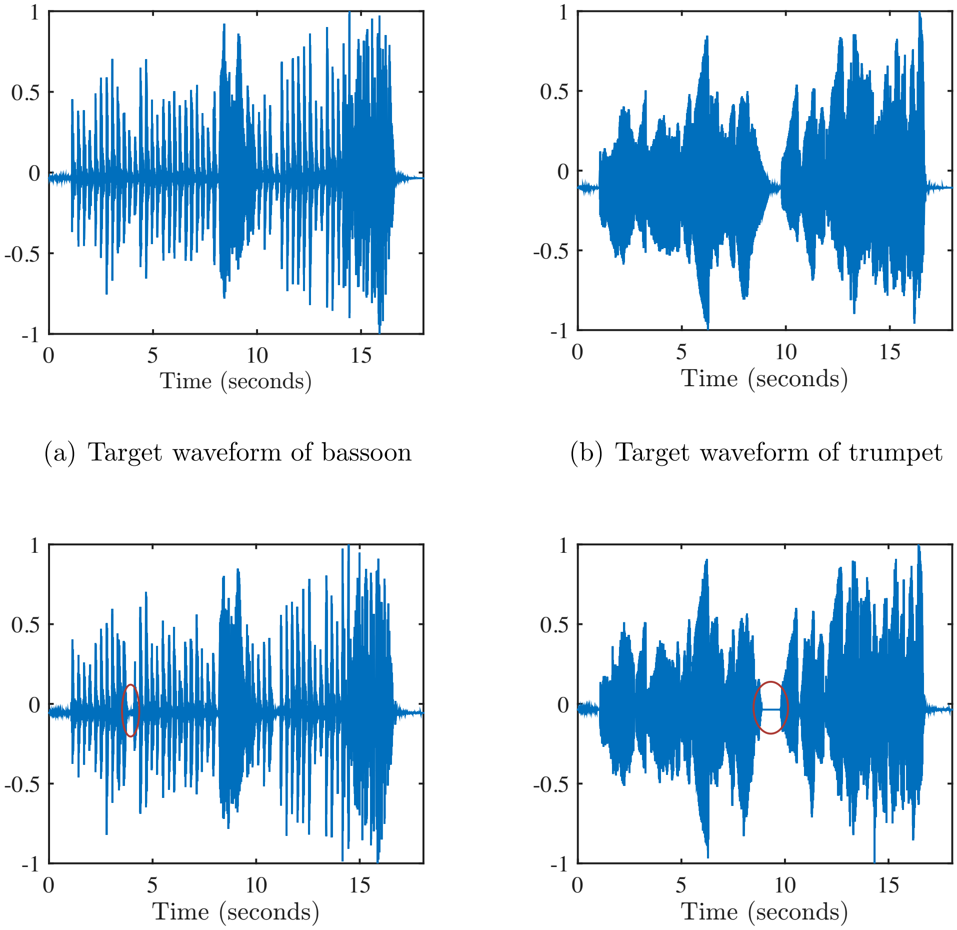
(b) Target magnitude spectrogram of trumpet

(c) Reconstructed magnitude spectrogram of bassoon

(d) Reconstructed magnitude spectrogram of trumpet

Figure 3.28: Target and reconstructed magnitude spectrograms bassoon and trumpet

To further demonstrate the performance of the proposed source template NMF, the target and the separation waveforms are all illustrated in Fig. 3.29. Referring to the target waveforms in Fig. 3.29(a) and 3.29(b), some obvious areas with inaccurate reconstructions are marked with a solid circle in Fig. 3.29(c) and 3.29(d).



(a) Target waveform of bassoon

(b) Target waveform of trumpet

(c) Reconstructed waveform of bassoon

(d) Reconstructed waveform of trumpet

Figure 3.29: Target and reconstructed waveforms of bassoon and trumpet

Except for the note profile results given above, the performance of the proposed algorithm was evaluated by using popular separation quality measures, such as the signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR) and the signal-to-artifact ratio (SAR) [133] [71]. From the basic notation in their work, the observed input signal S_{input} is considered to be the sum of

$$S_{input} = S_{target} + E_{interf} + E_{noise} + E_{artif}, \tag{3.45}$$

where S_{target}, E_{interf}, E_{noise} and E_{artif} mean the reconstructed track interference, noise and artifacts error terms individually. The SDR, SIR and SAR are determined in dB as

$$SDR = 20 \cdot log_{10}\left(\frac{S_{target}}{E_{interf} + E_{noise} + E_{artif}}\right), \tag{3.46}$$

$$SIR = 20 \cdot log_{10}\left(\frac{S_{target}}{E_{interf}}\right), \tag{3.47}$$

$$SAR = 20 \cdot log_{10}\left(\frac{S_{target} + E_{interf} + E_{noise}}{E_{artif}}\right). \tag{3.48}$$

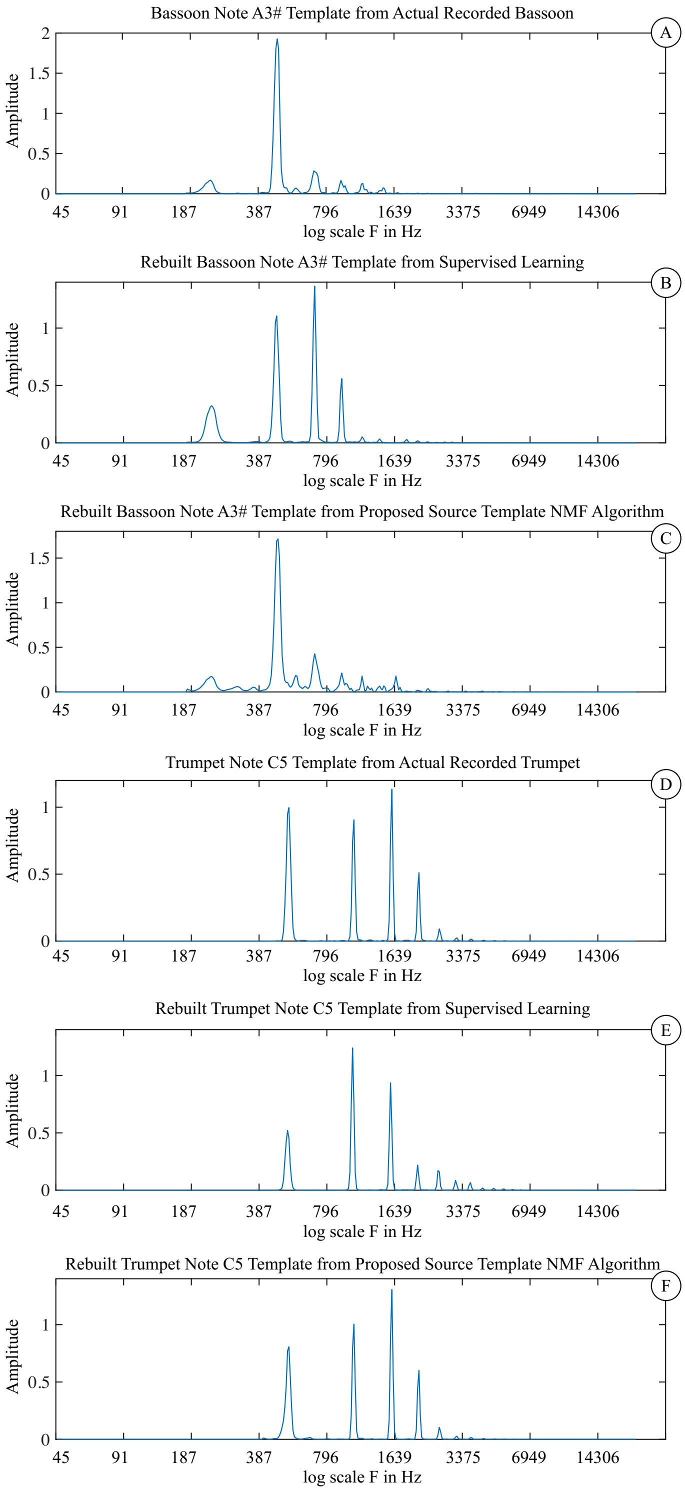
Table 3.2: Separation performance from proposed source template NMF algorithm

Instrument	**SAR** (dB)	**SDR** (dB)	**SIR** (dB)
Bassoon	14.39	11.06	13.93
Trumpet	16.28	15.90	26.71

Table 3.3: Separation performance from SoundPrism algorithm

Instrument	**SAR** (dB)	**SDR** (dB)	**SIR** (dB)
Bassoon	8.55	6.28	10.75
Trumpet	8.14	4.43	7.45

The separation performance is evaluated using SAR, SDR and SIR parameters in Table. 3.2 and 3.3. Because the SoundPrism [2] algorithm is more efficient than the other state-of-the-art algorithms in source separation [2], we employed it as a comparing algorithm in this dissertation. We also present the evolution of note templates in separation from beginning to the end in Fig. 3.30. We select the bassoon note $A3\#$ and trumpet note $C5$ as an example. At first, we give the note templates of $A3\#$ and $C5$ from their recorded instrument in Fig. 3.30 (A) and (D). This part is provided as the targets. At the beginning of separation, the initialized note templates of $A3\#$ and $C5$ in Fig. 3.30 (B) and (E) are obtained from the dataset [3, 4, 109] to build the constraints of instruments resonance characteristics and notes impulse excitation through the proposed supervised learning algorithm. Under the operation of the proposed source template NMF algorithm, the constraints of instruments resonance characteristics are updated and evolve a new form of the underling note templates in



Bassoon Note A3# Template from Actual Recorded Bassoon

Rebuilt Bassoon Note A3# Template from Supervised Learning

Rebuilt Bassoon Note A3# Template from Proposed Source Template NMF Algorithm

Trumpet Note C5 Template from Actual Recorded Trumpet

Rebuilt Trumpet Note C5 Template from Supervised Learning

Rebuilt Trumpet Note C5 Template from Proposed Source Template NMF Algorithm

Figure 3.30: Comparisons of note templates in separation

Fig. 3.30 (C) and (F) of the proposed model. In Fig. 3.30, it is easy to find that the evolved note templates of $A3\#$ of bassoon in sub-fig (C)and $C5$ of trumpet in sub-fig (F) are similar to the forms of note templates given by the actual recorded bassoon in sub-fig (A) and trumpet in sub-fig (D).

Chapter 4

Performance Evaluation and Comparison

The aim of the proposed source template NMF is to make a wide application on a multi-instrument separation scenario. Matlab toolbox [41] is adopted to provide the SAR, SDR, and SIR calculation between the separated signal and target source. A surprising number of datasets are created for music transcription and source separation testing such as MedleyDB [134], RWC music database [135], structural segmentation of the multitrack dataset (SSMD) [136], Woodwind Quintet [137], MAPS [138], LabROSA piano [139], TRIOS [1], Bach10 [31] and so on. However, considering the further performance evaluations compared with the other state-of-the-art algorithms [2, 30], the contents of alternative datasets should contain the mixed audio which is played by the multiple instruments, the targeted isolated sound source of each instrument, and the MIDI transcription which is necessary for the sound-prism algorithm [2] but not for the proposed source template NMF. For the dataset, based on these considerations above, the database TRIOS [1] and Bach10 [31] are both competent selections in this thesis. In terms of algorithms, the proposed source template NMF algorithm, the sound-prism method [2] and the oracle toolbox [30] were also selected as comparing algorithms. The sound-prism algorithm is a prevalent algorithm which has been proved with higher separation accuracies than many previous source separation algorithms [31, 140–142] and tested by Bach 10 database containing more complex and longer audio mixtures than ever used in source separation work. Meanwhile for the oracle toolbox, it is believed theoretically the best source separation method is based on time frequency masking methods and the analysis filter bank used on the separation system. Its calculation requires isolated sound sources. The mixed signals are filtered with the analysis filter bank. After that, the isolated sources, which are also filtered by the analysis filter bank, are used to obtain the ideal masks, which are the best mark that can be obtained with the given frequency resolution. Then these masks are applied to the mixed signal and the oracle separated signal is obtained. This process gives the best possible separation with the system set-up. It sets an upper bound of all the configurations of the proposed method. The runtime environment of each algorithm is listed in Table 4.1.

Table 4.1: Runtime environment of each algorithm

Side information	**Test algorithms**		
	Pitch-variant NMF	Sound-prism	Oracle toolbox
Score alignment info		✓	
Score list of each recording	✓	✓	
Multi-pitch estimation*		✓	
Note recordings	✓	✓	
Target audio			✓

* The "universal" likelihood model is trained on thousands of isolated musical chords generated by different combinations of notes from 16 kinds of instruments.

4.1 Separation Performance Based on TRIOS Dataset

In this TRIOS dataset [1], except the fifth percussion music, it has four melodies: Brahms's "Horn Trio in Eb major, op. 40," Lussier's "Bacchanale pour trompette, bassoon, et piano," Mozart's "Trio in Eb major Kegelstatt, K.498," and Schubert's "Piano Trio in Eb major, D.929." Each piece includes synchronized audio recordings created from the MIDI score by an instrument and a manually played isolated audio recording while the player is listening to the mix of the synthesized audio of other parts strictly synchronized in beat through a headphone. Though each isolated track is recorded individually, when we use Logic Pro software [143] for audio mixing, all pieces have a steady tempo. But it has errors in synchronized MIDI files of Brahms and Mozart pieces and these errors have effects on the application of the sound-prism algorithm. The details of audio mixture, isolated track sounds and synchronized MIDI

Table 4.2: TRIOS dataset [1]

Melody	**Solo**	**Mixture audio**				**MIDI**	**Time**	**Amount**
		2 source			3 source			
Brahms	horn	✓		✓	✓	✓	43 s	4 audio mixtures
	piano	✓	✓		✓	✓		
	violin		✓	✓	✓	✓		
Lussier	bassoon	✓		✓	✓	✓	18 s	4 audio mixtures
	piano	✓	✓		✓	✓		
	trumpet		✓	✓	✓	✓		
Mozart	clarinet	✓		✓	✓	✓	33 s	4 audio mixtures
	piano	✓	✓		✓	✓		
	viola		✓	✓	✓	✓		
Schubert	cello	✓		✓	✓	✓	53 s	4 audio mixtures
	piano	✓	✓		✓	✓		
	violin		✓	✓	✓	✓		

files are listed in Table 4.2. In one melody piece, the involved instruments are listed down a column, when being used for playing an audio mixture, they are marked by a checkmark in the column. Each two instruments can get one polyphony-2 audio mixtures. By means of it, we obtain twelve 2-source and four 3-source audio mixtures from this dataset. The mean length of all pieces is 36.75 seconds, and 8 kinds of instruments are used for playing melodies. For the track reconstruction, 24 samples are obtained from 2-source separation and 8 samples from 3-source separation. As it shows in Table 4.1, each melody is composed of different kinds of instruments, the numbers of reconstructed audio channels were not equal and calculated in Table 4.3.

4.1.1 Results under Proposed Source Template NMF Separation System

In this section, we use box-whisker plots to demonstrate our separation results in a descriptive statistical way. The box-whisker plot has an advantage in depicting the

Table 4.3: The amount of rebuilt channel audio

Rebuilt channels	**From mixture audio**		**Total NO.**
	2-source	3-source	
bassoon	2	1	3
cello	2	1	3
clarinet	2	1	3
horn	2	1	3
piano	8	4	12
trumpet	2	1	3
violin	4	2	6
viola	2	1	3

numbers of numerical data through its quartiles graphically, which is shown in Fig. 4.1 [12] to make a comparison between a box-whisker plot and a normal $N(0, 1\sigma^2)$ probability density function (pdf). Q_1 and Q_3 are the first and the third quartiles. The lowest values are located within 1.5 interquartile range (IQR) of the lowest quartile, and the same theory for the highest part. By these means, in Fig. 4.2, we present multi-source separation results using the proposed source template NMF. The mean values of SAR, SDR and SIR in 2-source separation are tagged around 10.5 dB, 5.8 dB and 9.8 dB respectively while in 3-source separation, they are 7 dB, 0 dB, and 2.1 dB. One reason is that when the number of sources increases, the number of parameters in the proposed model gets higher, but the useful information of indicating the notes and their sources are still very limited. This leads to the lower performances in 3-source separation. From the other respect, the separation work is based on the estimated process of the notes' onset-offset like features whose accuracy is also influenced by the number of sources. If the misleading estimations gets higher, the multi-source separation accuracy gets lower. To show a further separation performance evaluation based on the proposed source template NMF, Fig. 4.3 focuses on the SAR, SDR, and SIR parameters by rebuilt track. Each reconstructed audio channel has different performances of SAR values in Fig. 4.3(a), SDR values in Fig. 4.3(b) and SIR values

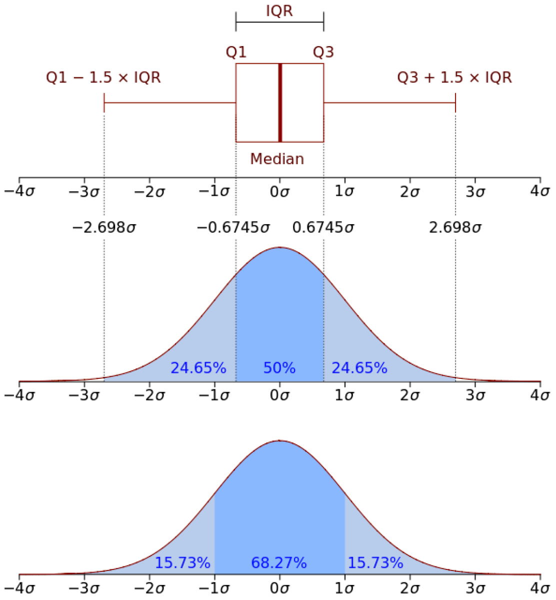


Figure 4.1: Box-whisker plot and a normal $N(0, 1\sigma^2)$ distribution [12]

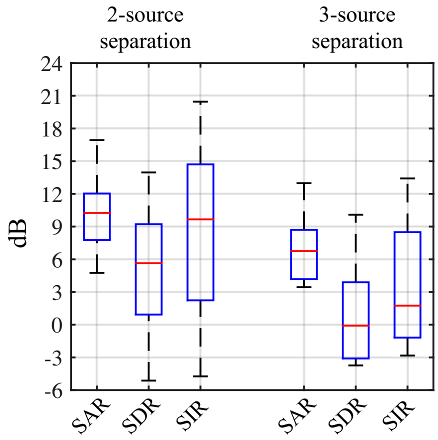
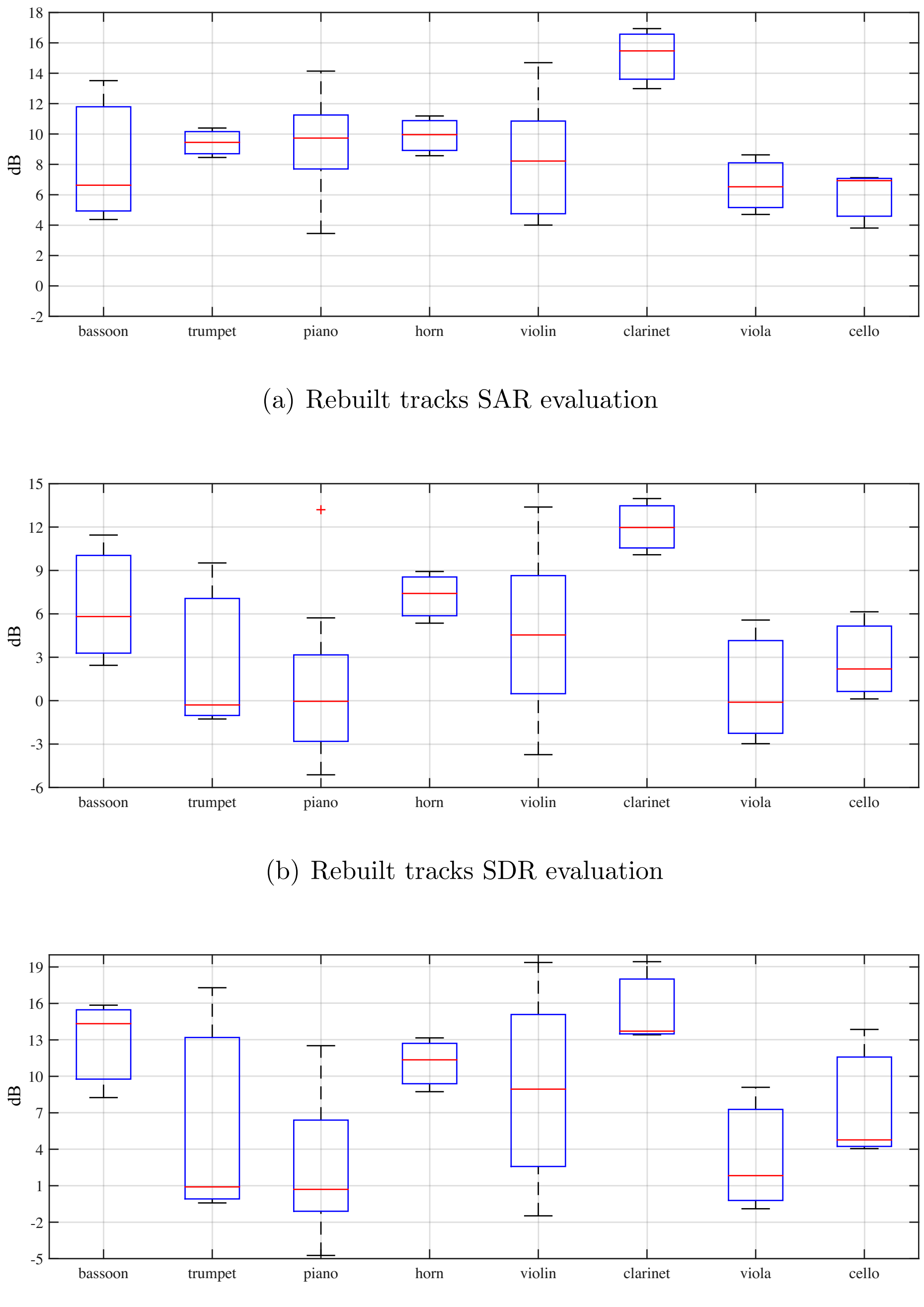


Figure 4.2: Multi-source separation on TRIOS dataset under the proposed source template NMF



(a) Rebuilt tracks SAR evaluation

(b) Rebuilt tracks SDR evaluation

(c) Rebuilt tracks SIR evaluation

Figure 4.3: Rebuilt track evaluation on TRIOS dataset under the proposed source template NMF by instrument

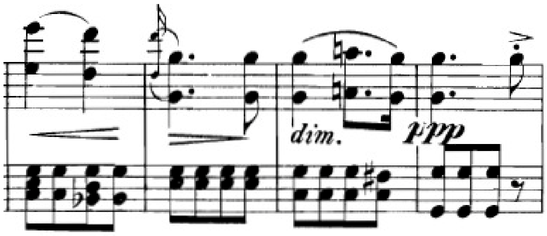


Figure 4.4: Piano chord in an audio mixture

in Fig. 4.3(c). We count on the 2-source separation rebuilt and 3-source separation rebuilt samples together to get enough samples and give a statistical analysis. The rebuilt samples of piano and violin are 6 and 12 which are greater than other instruments in the TRIOS dataset. However, the piano is a very complicated device with plenty of chords which cannot be recognized by human ears without any prior training practice. Fig. 4.4 gives one part of the sheet music of Schubert's piece in TRIOS database. Its low performance in 3-source separation may be a crucial reason to get the negative values in Fig. 4.3. It appears that one piano chord consists of 3 or 5 notes together and this pianism undoubtedly makes the multi-source separation challenging work.

4.1.2 Results under Sound-prism Separation System

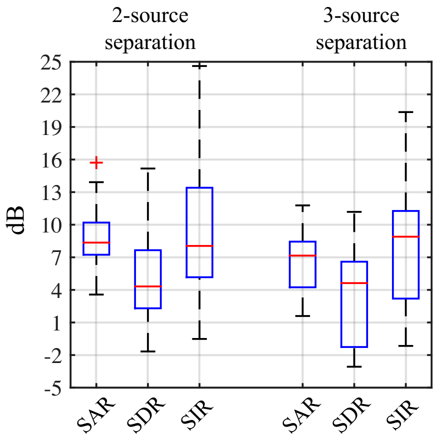
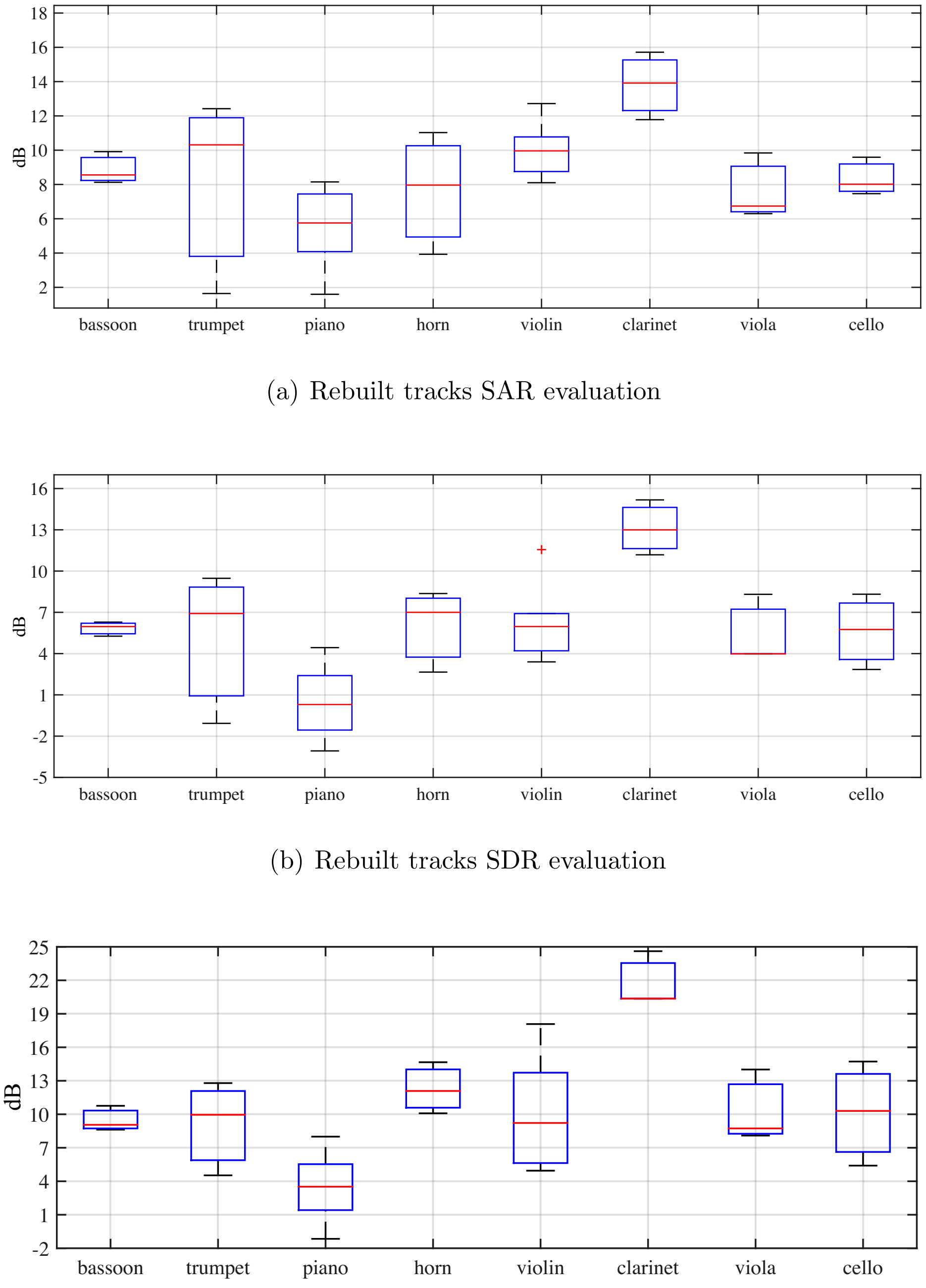


Figure 4.5: Source separation evaluation on TRIOS dataset under the sound-prism system



(a) Rebuilt tracks SAR evaluation

(b) Rebuilt tracks SDR evaluation

(c) Rebuilt tracks SIR evaluation

Figure 4.6: Rebuilt track evaluation on TRIOS dataset under the sound-prism system

In this dissertation, the sound-prism on-line music separation system [2] is adopted as one comparison against the proposed source template NMF algorithm. Generally, its separation work relies on the support of MIDI files as the capability to provide score alignment to calculate onset-offset like features by a Hidden Markov model (HMM). With the prior information of synchronized score events from the dataset of TRIOS, the sound-prism algorithm grabs the accurate score onset-offset features, then its separation work reaches high values of SAR, SDR and SIR parameters in Fig. 4.5. However in an aspect of 2-source separation, sound-prism performances of SAR, SDR and SIR are about 3.5 dB, 1.5 dB and 2 dB less than SAR, SDR and SIR in Fig. 4.2. For 3-source separation, the sound-prism performances of SAR, SDR and SIR keep their performances as well as in 2-source separation, with the aid from the score alignment information of the actual recordings from synchronized MIDI. One reason of its lower 2-source separation than the proposed source template NMF is the multi-pitch estimation algorithm used in sound-prism system. Though its multi-pitch estimation model has been trained by thousands of isolated musical chords from 16 kinds of instruments [2], the complex chords generated from the piano may go beyond its collections and deteriorate the reconstructions of source signals.

The rebuilt channel audio evaluation in the form of SAR, SDR and SIR are presented in Fig. 4.6. Compared with the separation performances in Fig. 4.3, the rebuilt channel audio performances are higher because of better performances in 3-source separation. But the rebuilt piano channel performance is close to the work in Fig. 4.6(b) and still gets the worst SAR, SDR and SIR values among the other sources. This suggests that the synchronized MIDI informed score alignment information improves the efficiency of multi-polyphony separation but the trained multi-pitch estimation algorithm has a limited chord recognition with actual samples beyond its collections.

4.1.3 Results under Oracle Separation System

The near-optimal time-frequency masks for single-channel source separation [144] from the Oracle-tool box is employed to show the best theoretical performance in our thesis, and its performance estimation results is described in Fig. 4.7. Whatever the performances of SAR, SDR and SIR in 2-source or 3-source separation, they imply

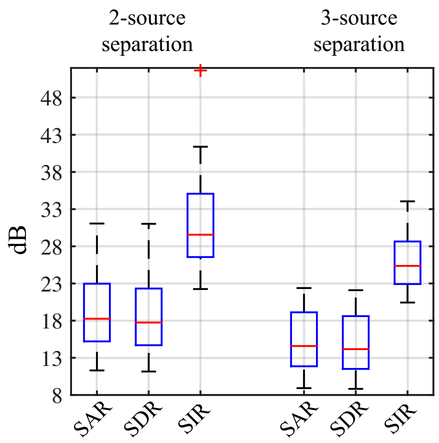
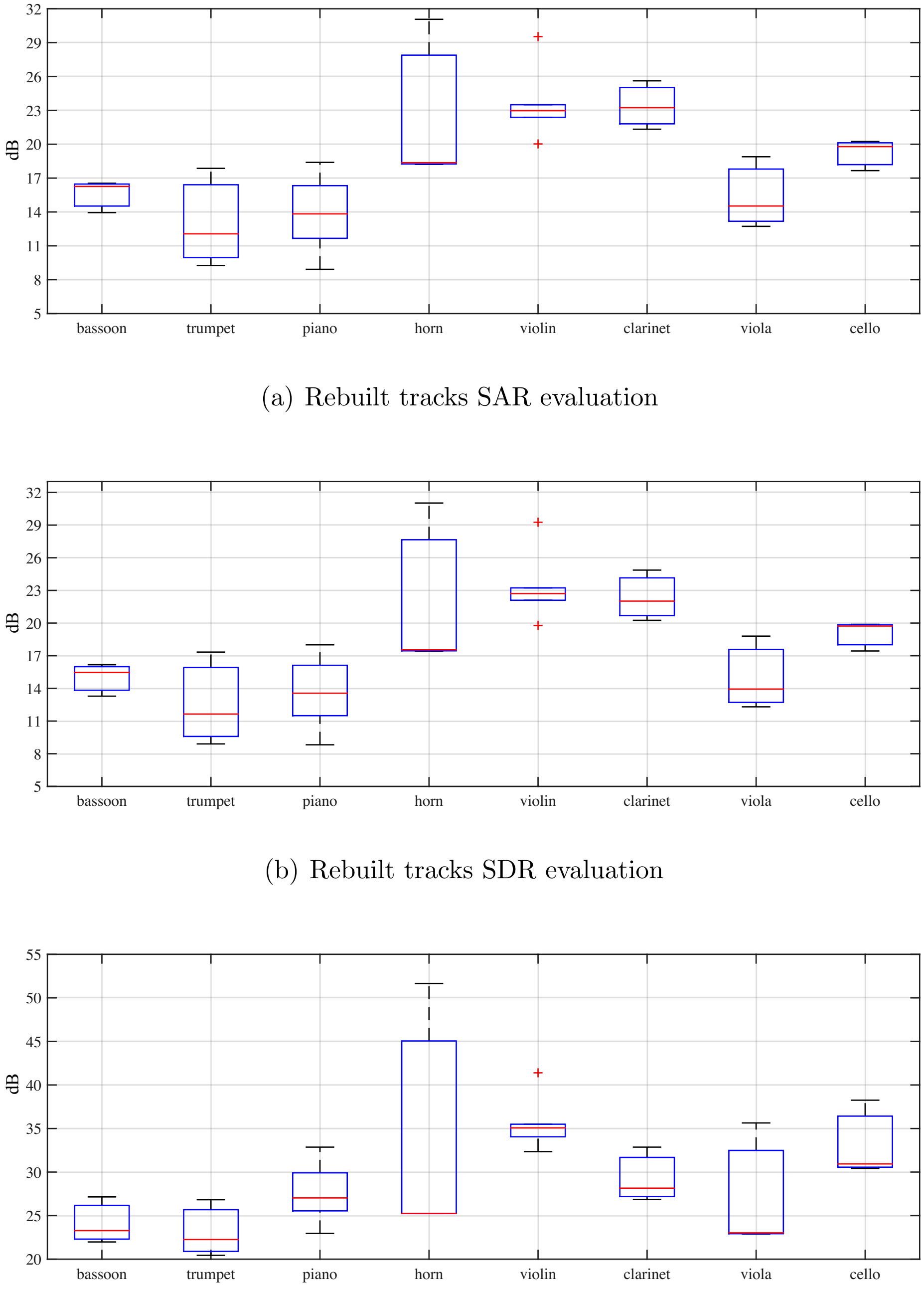


Figure 4.7: Source separation on TRIOS dataset under the near-optimal time-frequency masks for single-channel source separation algorithm from Oracle-tool box

that big potential improvement exists in the proposed source template NMF and sound-prism algorithms. The theoretical best performances of rebuilt channel audio are given in Fig. 4.8. Unlike the experiment results of the proposed source template NMF algorithm in Fig. 4.3 and sound-prism algorithms in Fig. 4.6, the rebuilt channel audio of piano does not have the worst performances among the others. It means that more side information concerned with the piano chords will improve the performances of rebuilt piano track audio in the proposed source template NMF and sound-prism algorithms.



(a) Rebuilt tracks SAR evaluation

(b) Rebuilt tracks SDR evaluation

(c) Rebuilt tracks SIR evaluation

Figure 4.8: Rebuilt track evaluation on TRIOS dataset under the Near-optimal time-frequency masks for single-channel source separation algorithm from Oracle tool box by instrument

4.2 Separation Performance on Bach10 Dataset

Table 4.4: Bach10 dataset

Melody	**Solo**	**Mixture audio**											**NO.**
		2-source						3-source				4 source	
Each Piece	bas	✓	✓	✓				✓	✓	✓		✓	11
	cla	✓			✓	✓		✓	✓		✓	✓	
	vln		✓		✓		✓	✓		✓	✓	✓	
	sax			✓		✓	✓		✓	✓	✓	✓	

Table 4.5: Time length (seconds) per melody piece in Bach 10 dataset

Melody name	AchGottundHerr	AchLiebenChristen	Christederdubist	ChristeDuBeistand	DieNacht	DieSonne	HerrGott	FuerDeinenThron	Jesus	NunBitten
Time length (second)	26	42	27.5	41.6	36.8	34.3	33.1	33.6	30	38.2
Mean value (second)	34.31									

The Bach 10 database [31] is composed of ten polyphony melodies, and each of them contains 4 isolated solos. These isolated solos in one melody piece are performed by players under the circumstance of listening to the mixture of all previous recorded audio instead of the metronome based synchronization. Because the players of solos are not involved in the group synchronized recording, the Bach 10 belongs to a less-than-ideal synchronization database. Because of the variation in the tempo, it is not difficult to find notes in inarticulate places. Actually, the fermata signs in pieces prolonged the notes beyond their normal duration. The structure and the mixed audio samples are listed in Table 4.4. Each piece of melody contains four isolated audio of different instruments. They are bassoon, clarinet, violin and saxophone. There are six different polyphony-2 audio mixtures per melody, if we choose from four different instruments of the two instruments of the combination number. With the same theory, there are four different polyphony-3 audio mixtures per melody. So for each melody piece, it can provide 11 audio mixtures. For 10 melodies in the Bach10 database, the total amount of available audio mixtures is 110 which includes 60 two-source audio mixtures, 40 three-source audio mixtures and 10 four-source audio mixtures. Their MIDI files are obtained from online web resources. The time length of each melody is listed in Table 4.5. Therefore, there are many changes of natural tempo, while the MIDI file has a constant tempo. We use this database to test our proposed source template NMF, and repeat the experiments sound-prism and Oracle toolbox algorithms. Typically, we calculate the possible number of rebuilt track audio per melody and give the final results in Table 4.6. The multi-source separation performances are plotted in Fig. 4.9. By observing the SAR, SDR and SIR results in Fig. 4.9(a), it shows the separation performance in two to four polyphony separation. Being consistent with the experiment results in TRIOS dataset. In Fig. 4.9(a), the performances decrease when the number of polyphony is increased. Its best results are SAR at 10.2 dB, SDR at 6 dB and SIR at 10.5 dB in two-source separation. The sound-prism based separation results given in Fig .4.9(b) shows its stability performances of SAR around 7.8 dB, SDR around 5.2 dB and SIR around 11 dB whenever the sound-prism algorithm is working at two, three or four-source separation conditions. Finally, the separation performances based on Oracle-toolbox are presented in Fig. 4.9 to indicate the theoretical limit of the state-of-the-art algorithms.

Table 4.6: Numbers of rebuilt audio channels per melody in Bach10

	Rec solo	**From audio mixture**			**Total NO.**
		2-source	3-source	4-source	
Per melody	bassoon	6	4	1	11
	clarinet	6	4	1	11
	violin	6	4	1	11
	saxophone	6	4	1	11

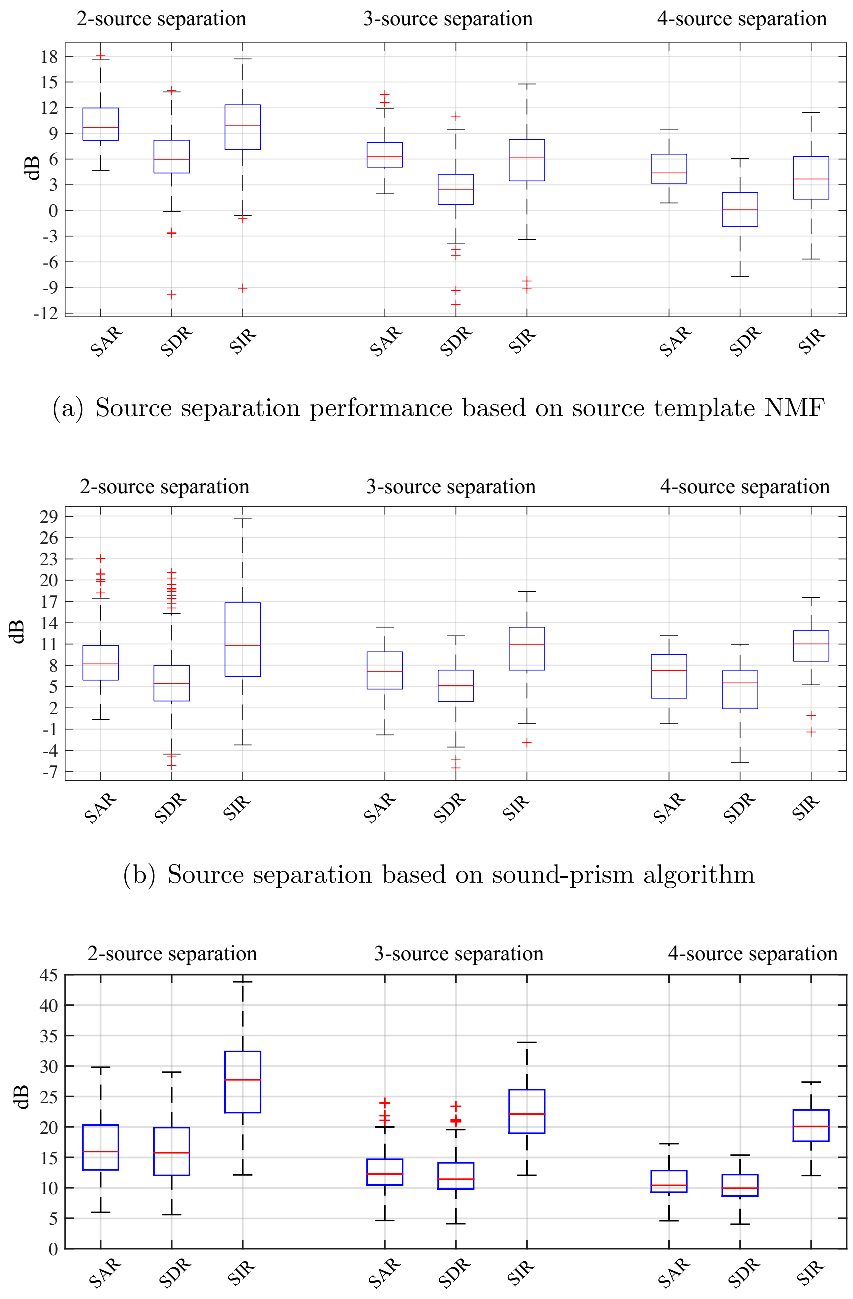
It is also interesting to see the performances of rebuilt tracks in Fig. 4.10, 4.11 and 4.12 which were based on the proposed source template NMF, the sound-prism and the Oracle-toolbox algorithm respectively. In polyphony-2 separation condition, the rebuilt clarinet track audio has the best performance in SAR, SDR and SIR values based on all proposed source template NMF, sound-prism and Oracle-toolbox algorithms. The performances of the proposed source template NMF, the rebuilt violin track audio has violent in changes performances on multi-polyphony separation experiments.

4.3 Separation Performance Test by Simulating Actual Environment

As the experiments mentioned above, these algorithms are all tested using audio at laboratory level. But in real circumstances, the multi-polyphony audio mixture is generated and recorded in noisy situations because of outside noise such as the bi-acoustics sounds from crowds, air conditioning, power supplies, reverberation and so on. But the real noise background is hard to simulate. If we only consider a simplified noise background with the similar characteristics of white Gaussian noise, and have a flat spectrum over the range of frequencies, the simulated audio mixture is generated by adding Gaussian type white noise to the audio mixture from the Bach 10 dataset. It is also interesting to see the variations of separation performances with the different levels of background noise. We use the signal-to-noise ratio (SNR) [145, 146] parameter to measure the noise energy and observe the behaviours of these three algorithms with various SNR. The SNR here indicates an average signal-to-noise ratio and its definition is

$$SNR_{dB} = E\left[10log_{10}\left[\left(\frac{V_{signal}}{V_{noise}}\right)^2\right]\right]. \tag{4.1}$$

In experiments, we plan to test the polyphony-2 separation performances of these algorithms on simulated audio mixtures with different levels of noise. In order to control the noise level, the SNR is adjusted. Fig. 4.13 shows the original audio mixture from the Bach 10 dataset and lists its simulated audio in the min and max SNR value of 4 dB and 16 dB. From the plot, it is easy to observe that the curves of



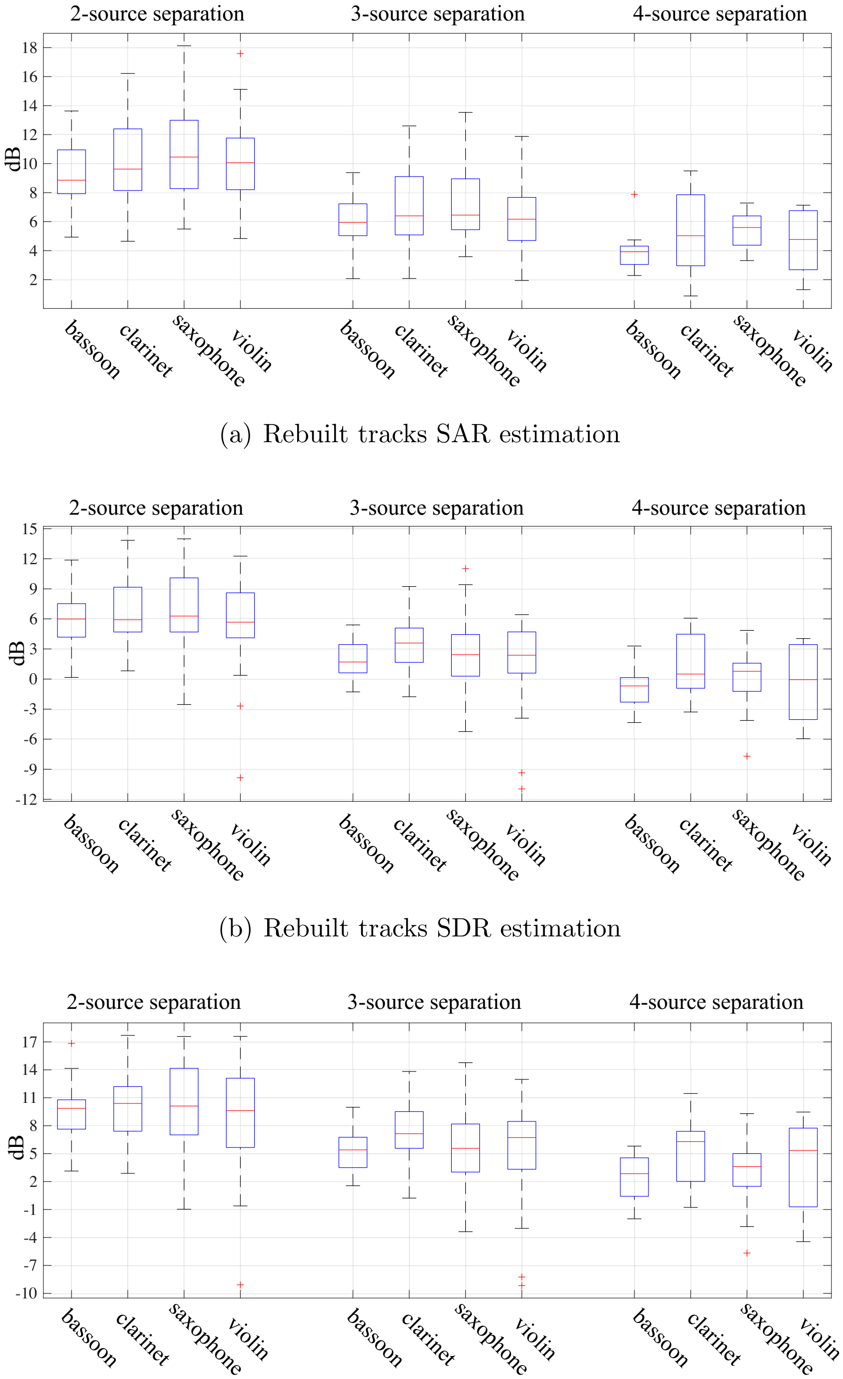
(a) Source separation performance based on source template NMF

(b) Source separation based on sound-prism algorithm

(c) Source separation based on oracle-toolbox algorithm

Figure 4.9: Source separation on Bach 10 dataset among these three algorithms

the simulated audio mixture at SNR 16 dB are smoother than the simulated audio mixture at SNR 4 dB. Therefore the expected separation performances at SNR 16 dB should be higher than the results at SNR 4 dB.

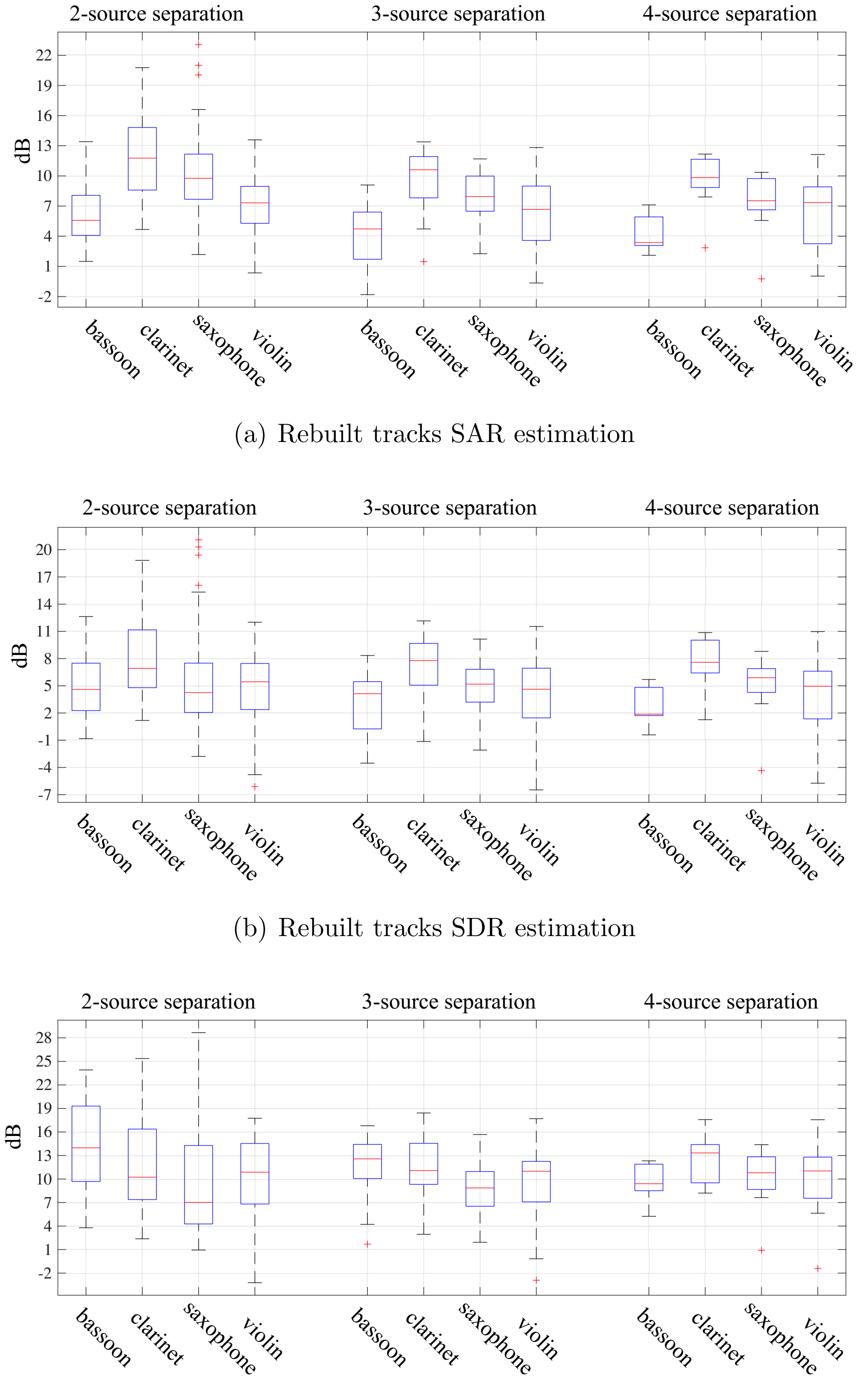


(a) Rebuilt tracks SAR estimation

(b) Rebuilt tracks SDR estimation

(c) Rebuilt tracks SIR estimation

Figure 4.10: Rebuilt track performance evaluation on Bach 10 dataset under the proposed source template NMF

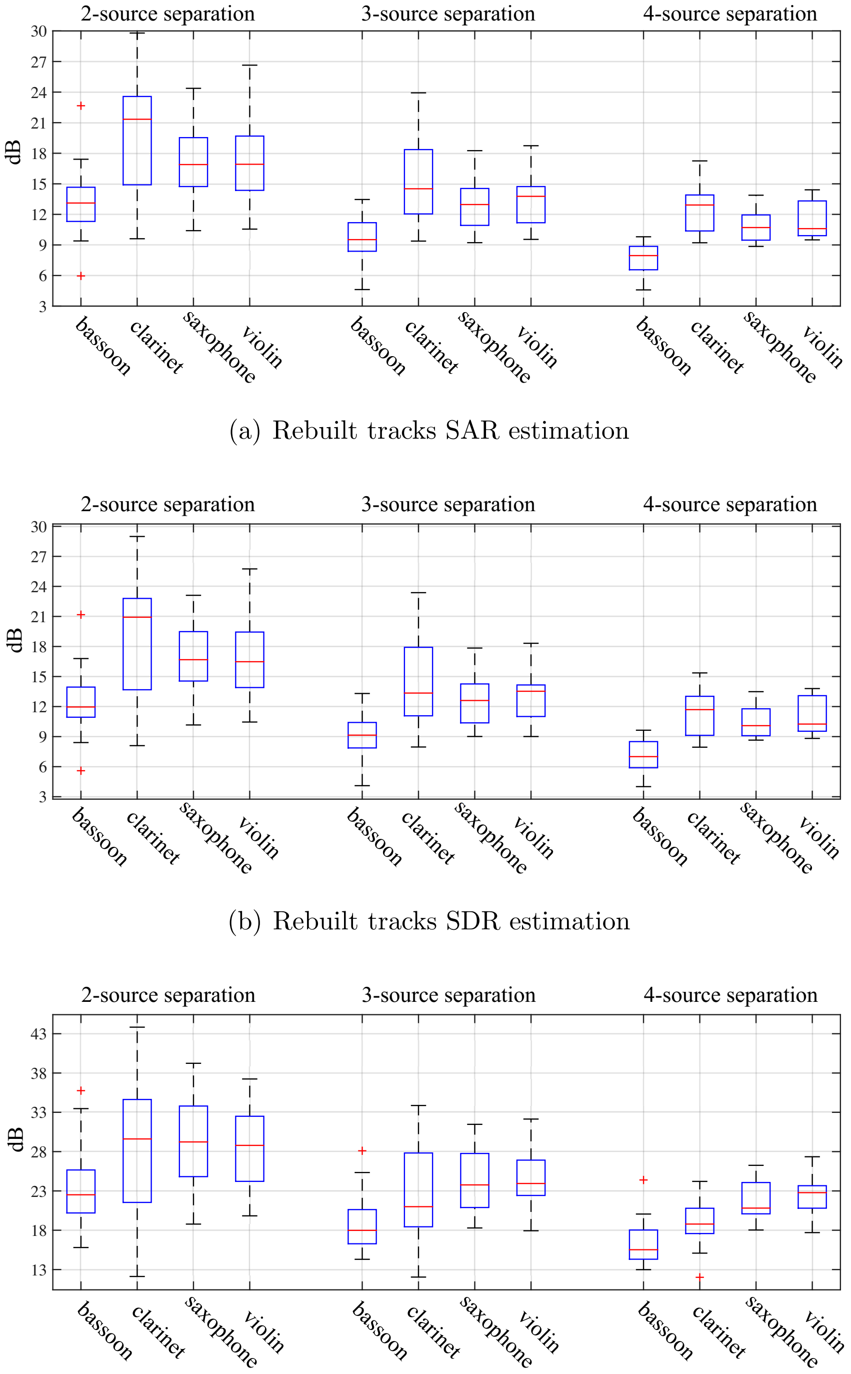


(a) Rebuilt tracks SAR estimation

(b) Rebuilt tracks SDR estimation

(c) Rebuilt tracks SIR estimation

Figure 4.11: Rebuilt track performance evaluation on Bach 10 dataset under the SoundPrism



(a) Rebuilt tracks SAR estimation

(b) Rebuilt tracks SDR estimation

(c) Rebuilt tracks SIR estimation

Figure 4.12: Rebuilt track performance evaluation on Bach 10 dataset under the Oracle-toolbox

The audio mixture separation is only referring to the polyphony-2 separation work. The polyphony-2 audio mixtures in Bach 10 dataset are 120 as well as we can derive that the simulated polyphony two audio mixtures are equal to the same number. The first procedure of the test is to choose one of the source separation algorithms from the three and calculate the SAR, SDR and SIR parameters in Fig. 4.14, 4.16 and 4.18. It is obvious to see that except for the Oracle-toolbox algorithm, the two-source separation performance on the simulated audio mixtures under the proposed source template NMF and sound-prism algorithms are both sensitive to the effects of Gaussian white noise. Particularly, the sound-prism algorithm has the lowest performances. For an example, the mean SDR of the Oracle algorithm is

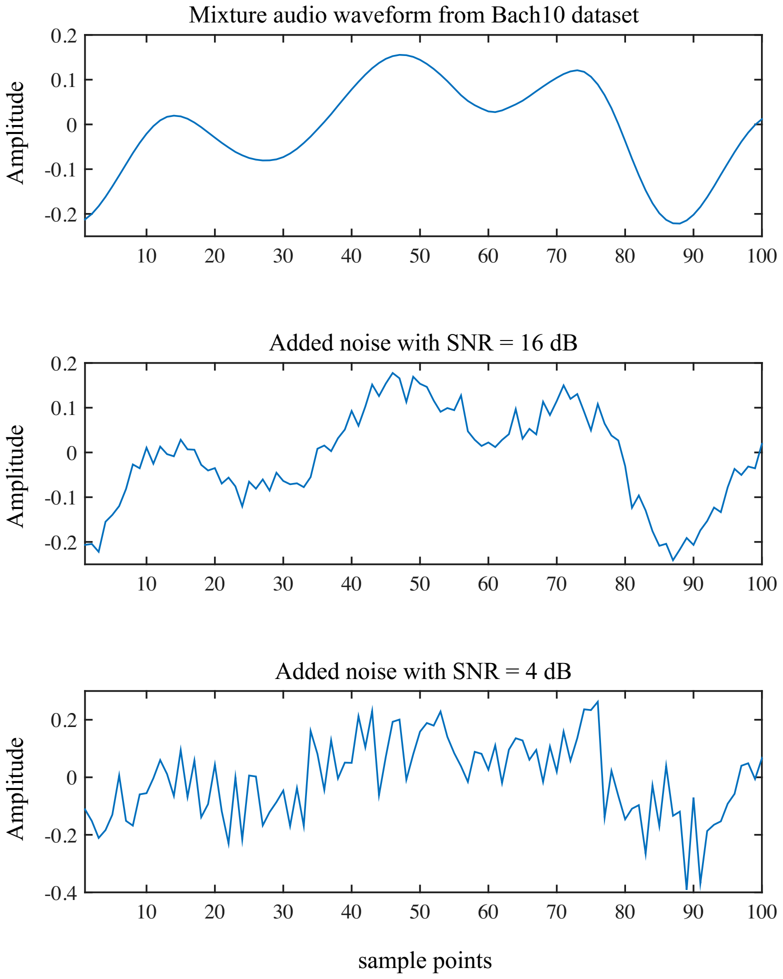
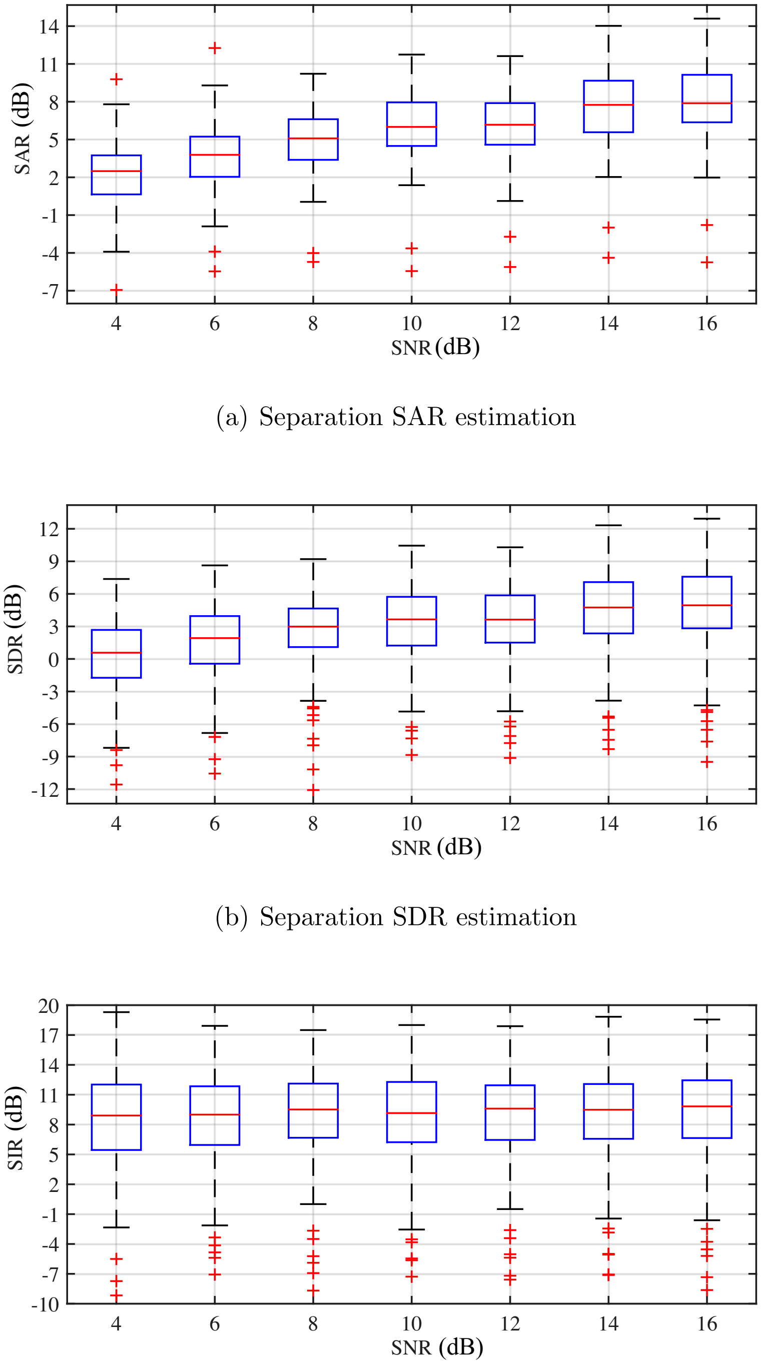


Figure 4.13: Audio mixture and its simulations with noise

around 15 dB at Fig. 4.9(c), and its mean SDR is still located around 15 dB in Fig. 4.18(b) when it works on the stimulated mixture audio with mean SNR at 16 dB; the mean SDR of the proposed source template NMF is around 6 dB at Fig. 4.9(a) and its mean SDR keeps the value at 5 dB in Fig. 4.14(b) with SNR at 16 dB; but the value of average SDR goes down when the SoundPrism algorithm is utilized to process the simulated audio mixtures, and the value of average SDR is reduced from 5 dB in Fig. 4.9(b) to 1.8 dB in Fig. 4.16(b). Furthermore, this reduction goes into the negative quadrant with the noise signal energy increasing to SNR 4 dB.

The polyphony-2 separation performances of SAR, SDR and SIR show their values of rebuilt audio track performances by the instrument in Fig. 4.15, 4.17 and 4.19. The number of rebuilt audio samples of each instrument is 30 s and they are all reconstructed from polyphony-2 separation results. The instruments include bassoon, clarinet, saxophone and violin, and they are grouped together under the different SNR values. For the rebuilt audio track performances, the proposed source template NMF and the sound-prism algorithm both have the problem of getting lower average values of SAR, SDR and SIR with different values of SNR. In particular, the rebuilt violin track always gets a lower mark than others on the rebuilt track. However the proposed source template NMF keeps its competitive edge on values of average SAR, SDR and SIR compared with the sound-prism algorithm.

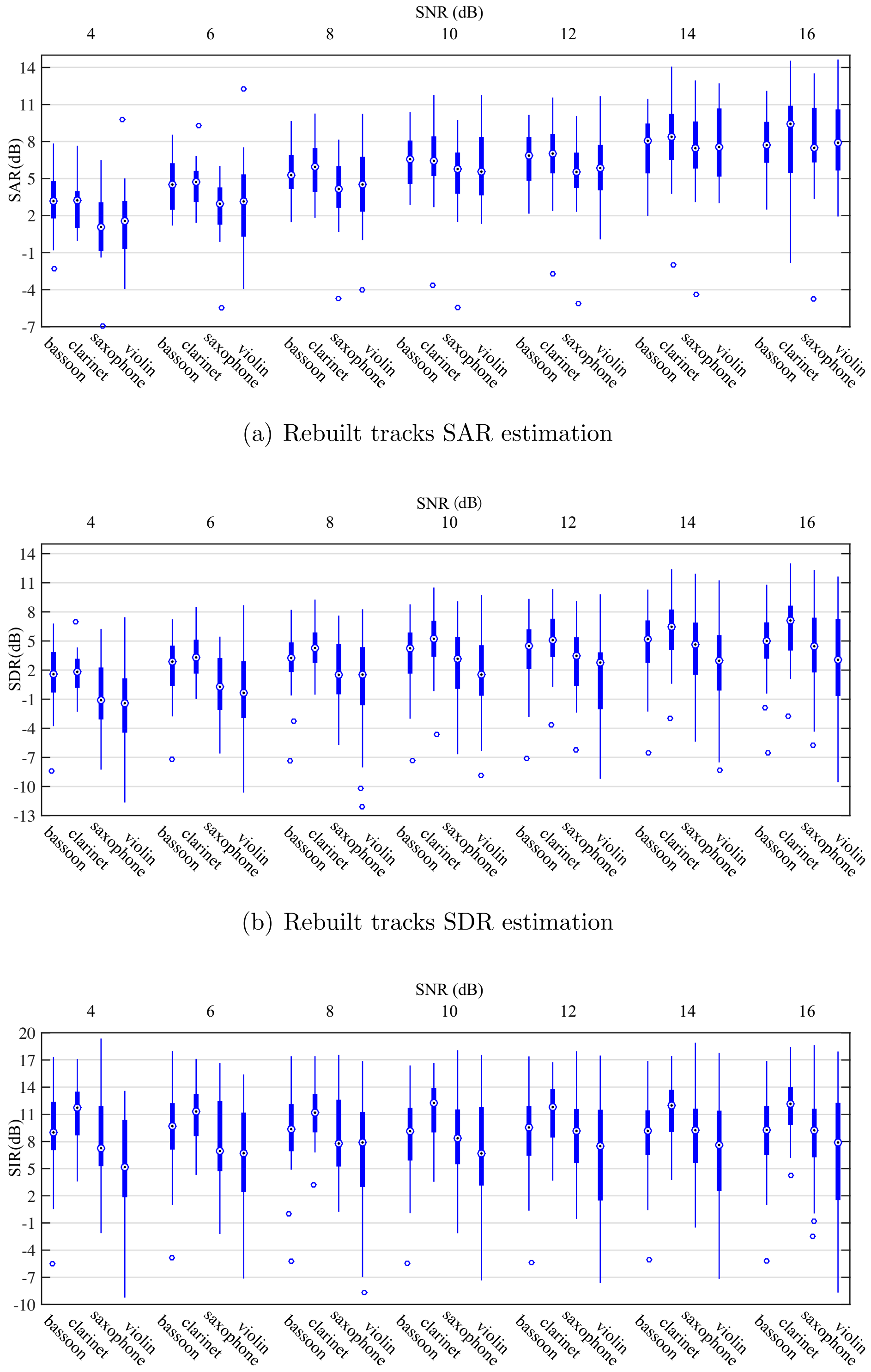


(a) Separation SAR estimation

(b) Separation SDR estimation

(c) Separation SIR estimation

Figure 4.14: Polyphony two separation performance evaluation on noise added stimulation audio mixture under the source template NMF

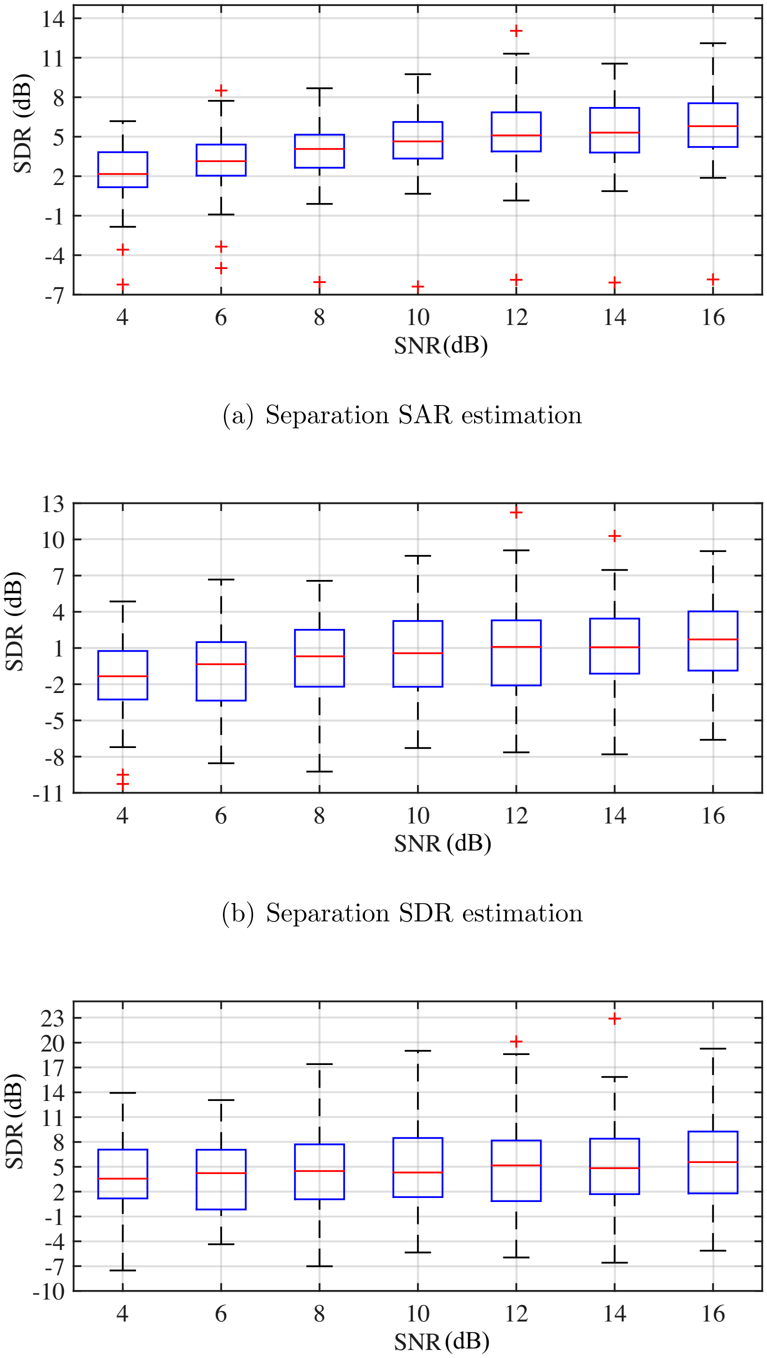


(a) Rebuilt tracks SAR estimation

(b) Rebuilt tracks SDR estimation

(c) Rebuilt tracks SIR estimation

Figure 4.15: Rebuilt track performance evaluation on noise added stimulation audio mixture under the source template NMF

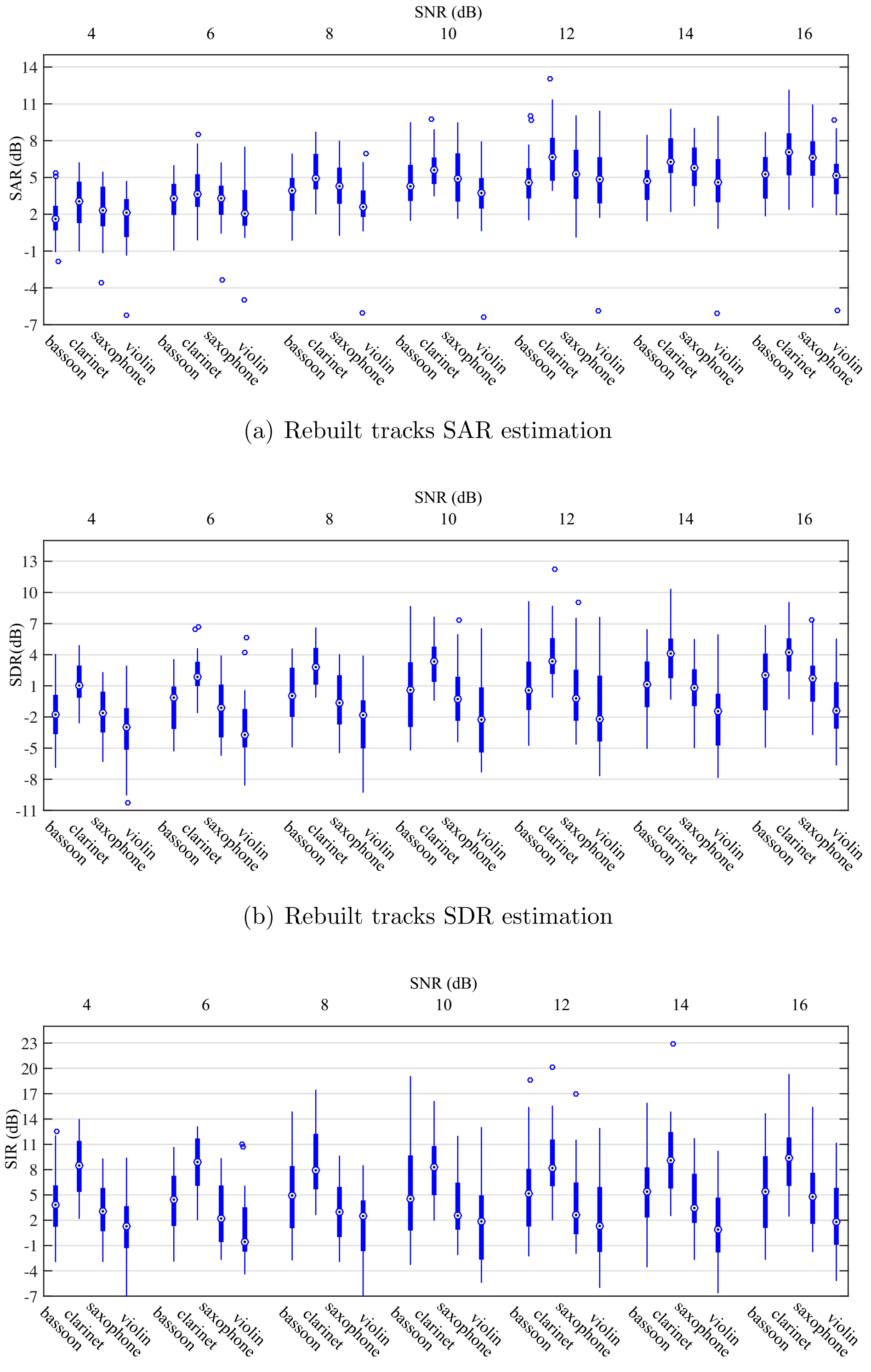


(a) Separation SAR estimation

(b) Separation SDR estimation

(c) Separation SIR estimation

Figure 4.16: Polyphony two separation performance evaluation on noise added stimulation audio mixture under the sound-prism algorithm

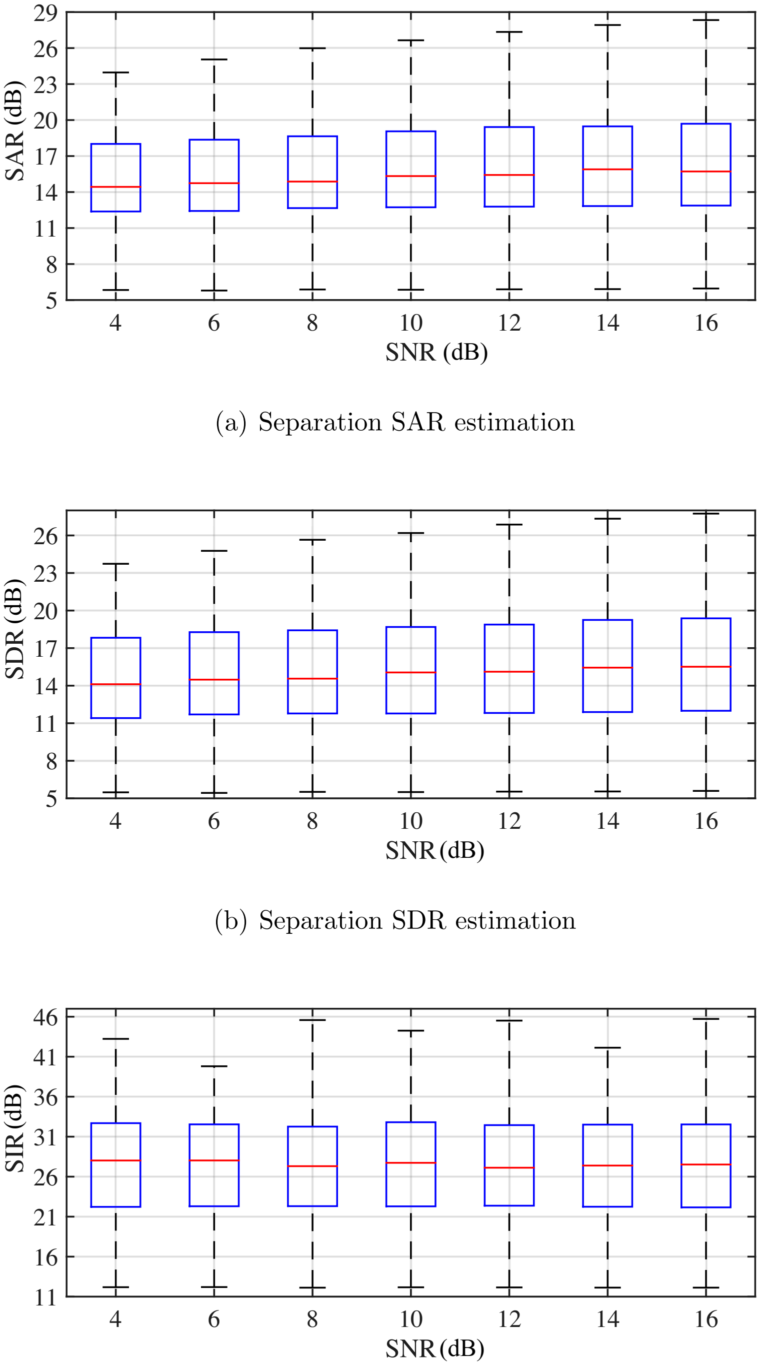


(a) Rebuilt tracks SAR estimation

(b) Rebuilt tracks SDR estimation

(c) Rebuilt tracks SIR estimation

Figure 4.17: Rebuilt track performance evaluation on noise added stimulation audio mixture under the sound-prism algorithm

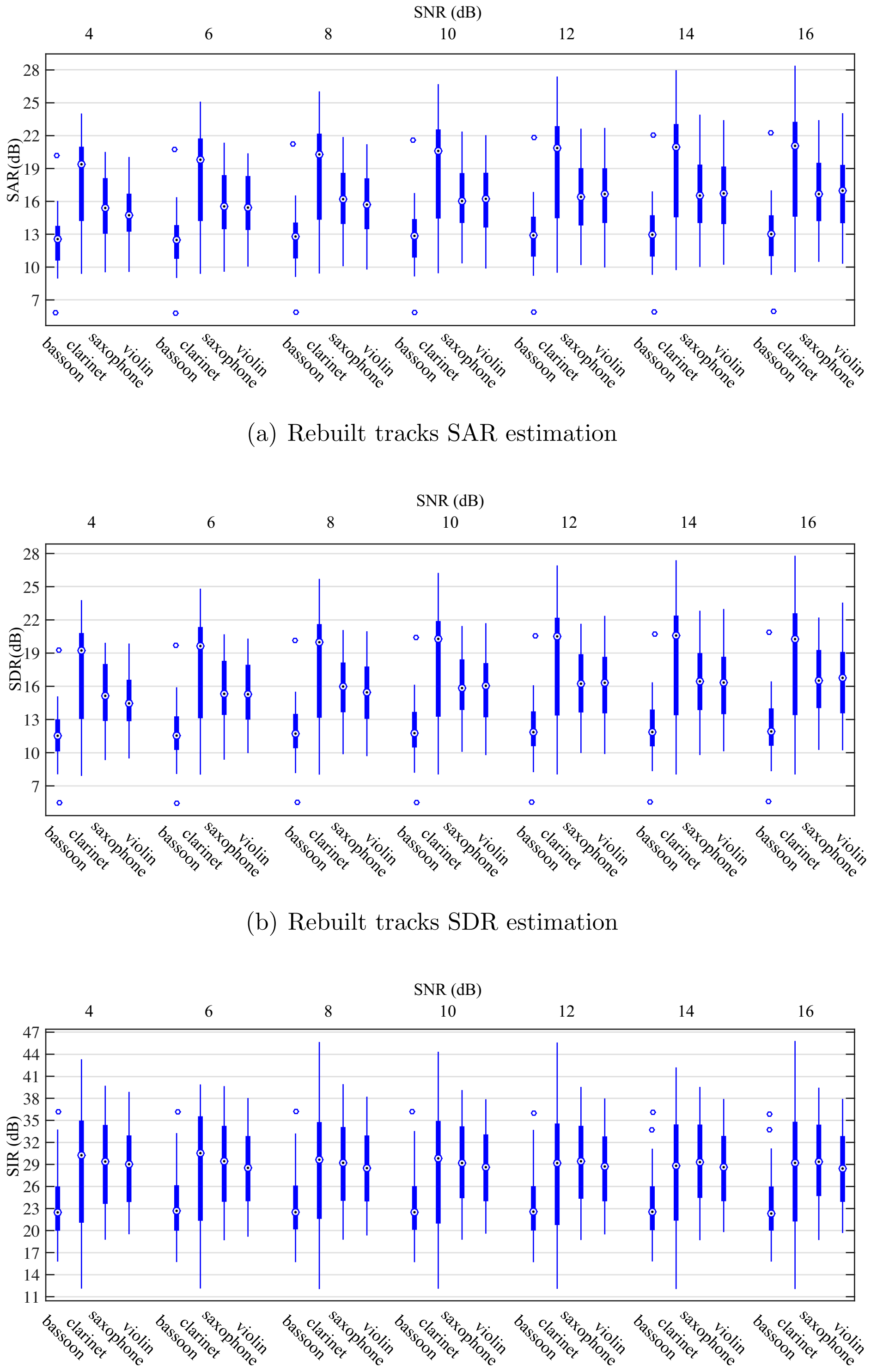


(a) Separation SAR estimation

(b) Separation SDR estimation

(c) Separation SIR estimation

Figure 4.18: Polyphony two separation performance evaluation on noise added stimulation audio mixture under the Oracle-toolbox algorithm



(a) Rebuilt tracks SAR estimation

(b) Rebuilt tracks SDR estimation

(c) Rebuilt tracks SIR estimation

Figure 4.19: Rebuilt track performance evaluation on noise added stimulation audio mixture under the Oracle-boolbox algorithm

Chapter 5

Conclusion and Discussion

In this dissertation, we developed the proposed source template NMF method for multi-instrument separation for monaural musical audio and designed the experimental tests based on TRIOS and Bach10 datasets compared with the SoundPrism and Oracle-toolbox algorithms. The conclusion of these experiments is made here from the performance evaluations in Chapter 4 to summarize the value of the research. Meanwhile the discussion is also made to discover the advantages and disadvantages theoretically in our proposed algorithm and to look forward to potential and realistic improvements in the future.

5.1 Conclusion

In this research, we assumed that each instrument notes' templates in the frequency domain are generated by the production of a basis impulse excitation with the shift-variant resonance functions of different fundamental frequency f_0s. Based on this assumption, we developed the proposed source template NMF model which uses linear combinations of the source labelled score spectra templates to approximate the spectrogram of the observed audio mixture. The prior knowledge of the note spectra templates are obtained from a recorded dataset full of isolated instrument notes. The proposed source template NMF model evolves the prior informed resonance functions approximately to the actual recorded instrument and estimates the combination coefficients of each template to build the separation mask for instrument audio track reconstruction. We test the proposed source template model with the TRIOS and Bach10 datasets, and compared the multi-source separation performances with

SoundPrism and Oracle-toolbox algorithms. On one side, from the performance comparisons, we found that the proposed source template NMF model had equivalent or higher qualities in polyphony-2 separation depending on the datasets. In polyphony-3 or -4 separation, the performances of our proposed source template NMF method is behind the SoundPrism algorithm. But the necessary special requirements of the SoundPrism algorithm includes a score-alignment defined MIDI file and multi-pitch detection algorithm trained by thousands of isolated chords, which are far beyond the necessary, isolated notes' recordings and an actual recording scores list used in the proposed source template NMF. On the other side, from the SNR tests, the performance of the proposed source template NMF exceeded the SoundPrism algorithm in polyphony-2 separation. The overall experimental test results tell us that the proposed source template NMF has advantages in two-source separation and has better noise performance.

5.2 Discussion

The proposed source template NMF is designed in this dissertation to solve the source separation problem for monaural musical audio. Compared with the SoundPrism algorithm, the proposed source template NMF achieved a higher accuracy of two-source separation on the TRIOS dataset and almost the same performance of two-source separation on the Bach-10 dataset with fewer supports from side information. Furthermore, the proposed source template NMF has a better noise performance on two-source separation on simulated audio. There were many reasons to explain the performance on two or multi-source separation.

In the proposed model, the isolated notes' audio informed instrument templates have the potential to promote the efficiency of two-source separation on ensemble and polyphony audio. The isolated notes' audio contains prior information of certain frequencies' oscillation under strings or airflow vibrations which lead to the excitation impulse and driving force for an instrument. With the help of this prior information, the proposed source template NMF model estimates the maximum likelihood of the note's onset-offset features corresponding to the prior template. The notes' maximum likelihood estimations are supervised by the prior information of notes list and the pitch-checking algorithm to improve the accuracy. Then the supervised estimations of notes' onset-offset features are used alternatively to evolve the actual recorded instruments' templates from prior ones. The proposed algorithm proved its efficiency with results in Fig. 4.2 and Fig. 4.9(a).

But the proposed source template NMF algorithm is limited to some samples in two, three or four-source separation because of two main reasons. The first reason is called the octave error problem which is still a challenge in a multi-pitch detection algorithm. A scenario of the octave error problem in our experiments is described and explained in Fig. 5.1 which happened in the sample of Brahms's rhythm from the TRIOS dataset. In Ⓐ of Fig. 5.1, it shows the score onset-offset-like features of the rebuilt violin track. The feature A in the row of MIDI number 87 and the feature B in the row of MIDI number 75 are both onset-offset-like features which have one octave gap. But according to the prior information of notes in actual recordings in Ⓑ, the feature A is a misleading estimation. We found the corresponding time of events A and B in the spectrogram of the audio mixtures of Ⓒ and marked it with a dotted vertical line. Then we extracted a slice of the mixture audio spectrogram at the time of events A and B and presented the time slice in the item Ⓓ to observe the frequency profile's history at that time. Combined with the prior knowledge of the notes template in Ⓔ and Ⓕ, the goal of our proposed algorithm is to find out that which gives the contributions to Ⓓ. However, because of the frequency components of Ⓕ are covered by Ⓔ it's hard to decide whether Ⓓ is a combination of Ⓔ and Ⓕ, or just evolving from only one of them. The misleading part cannot be recognized until some necessary side information is given.

The second reason of misleading estimations under our proposed source template NMF is the harmonic component compression phenomenon in some instrument playing. The example of harmonic components compression of a wind pipe type instrument, alto saxophone, is given in Fig. 5.2. The note templates show the frequency profiles of the notes from the row of MIDI number 49 to MIDI number 80 which are provided by the dataset [109]. The harmonic component compression phenomenon is more obvious in the note spectrum template at MIDI number 80 than at MIDI number 66. In an actual recording scenario shown by item Ⓕ of Fig. 5.3, the fundamental frequency component is more outstanding than other harmonics and causes a new type of misleading in note onset-offset-like features estimations. In Fig. 5.3, we prescribed this kind of misleading estimation in source separation of the sample called "ChristeDuBeistand" from the Bach-10 dataset. In Ⓐ and Ⓒ, we give the estimated note's onset-offset-like features of clarinet track and violin track compared with their prior score alignment information in Ⓑ and Ⓓ. From this comparison, we found the estimated A in Ⓐ was an effective estimate but the estimated B in Ⓒ was misleading. Then we located the time of the effective A and the misleading B in the spectrogram of mixture audio in Ⓔ and spread out the frequency profiles history in Ⓕ. It clearly shows the phenomenon of harmonic components compression in which it is observed as the primary outstanding fundamental component and withered harmonics. Under this circumstance, though the proposed source template NMF model is informed by the prior note templates in Ⓖ and Ⓗ, it is still hard to say which one of them, Ⓖ or Ⓗ, has the main contributions to the actual mixture recording at the specific time when it is indicated by a dotted vertical line. Because of the absence of the necessary information from harmonic components, this type of misleading feature cannot be solved without the help from the further information of notes playing. We should note that these kinds of misleading feature will accumulate with increasing number of sources and prove to be the explanation of the low performance in three or four-source separation behaviours.

However, these limitations do not affect the effectiveness of our proposed source template NMF algorithm in conditions with simulated noise background. In order to explain this advantageous characteristic, we illustrate the frequency profiles at a specific time in the spectrograms of the original and simulated audio mixture at different levels of SNR. Item Ⓐ in Fig. 5.4 is a time slice from spectrograms of original audio. The lowest peak is located around 91 Hz and can be recognized clearly from other peaks. But its shape becomes more ambiguous to recognize when noise energy is higher in Ⓒ and Ⓓ. This deterioration of the peak value leads to a challenge to the multi-pitch detection algorithm in the SoundPrism algorithm, which is very sensitive to noise and even goes on strike if a silence period exists in the audio mixture. But in our proposed source template NMF algorithm, the score onset-offset-like feature estimation is based on the prior note templates, which is hard to be affected by the trivial changes in the frequency profiles of the audio mixture.

5.3 Future Work

According to the discussion, future work on the proposed approach can start from the work of solving the three problems of octave error, harmonic component compression and missing fundamental frequency. In order to solve the octave error problem, the estimation of note activation has to be detected to find out whether the notes have an octave relationship and are active simultaneously. If they are, we could use the note timbre of the source as side information to evaluate the spectrum of the observed audio mixtures and determine whether the spectrum is a frequency profile of only one note or the sum of notes. In this way, the note timbre provides clues to recognize the situation of multi-pitch estimation. Following a similar approach, we may find solutions of harmonic component compression and missing fundamental frequency phenomena. The research of note timbre characteristics by source would become a key issue in the future research.

Additionally, the proposed algorithm can also be developed in exploring the phase side information from the observed audio mixture STFT. In source separation work, the STFT of observed audio mixtures give us two aspects of information: the first is the magnitude of the spectrogram, the second is its phase. Because it is difficult to establish a model with constraints of both the note frequency magnitudes and the phase, in the proposed algorithm, we only develop an updatable constraint for notes by source and focus on the magnitudes. So the phase of the rebuilt signal of the target channel is inherited from the original STFT and is not denoted by source. For the future development, we should pay attention to the research of the side information of the phase and establish a statistic constraint to rebuild the source defined phase characteristics.

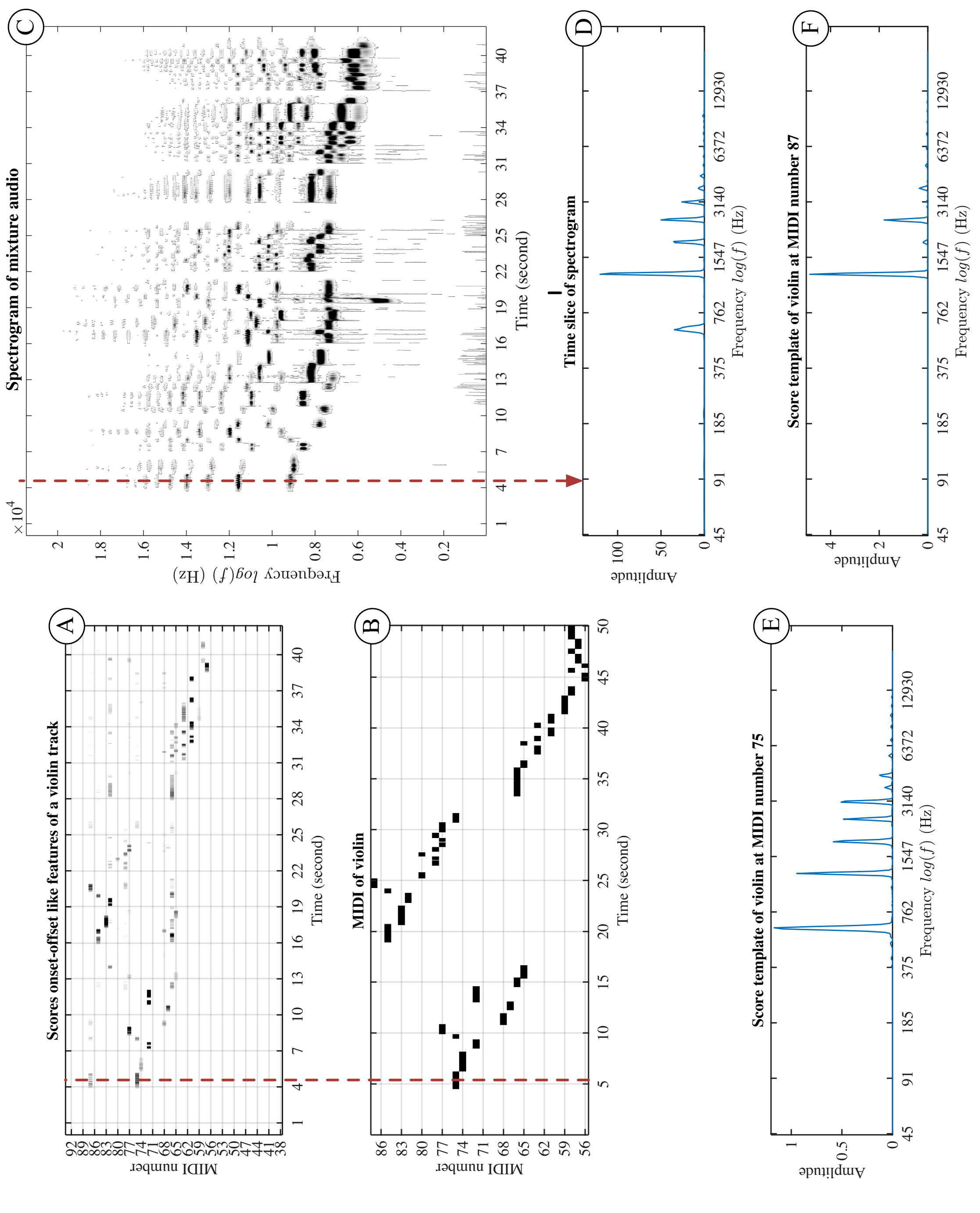


Figure 5.1: Scenario of octave error problem

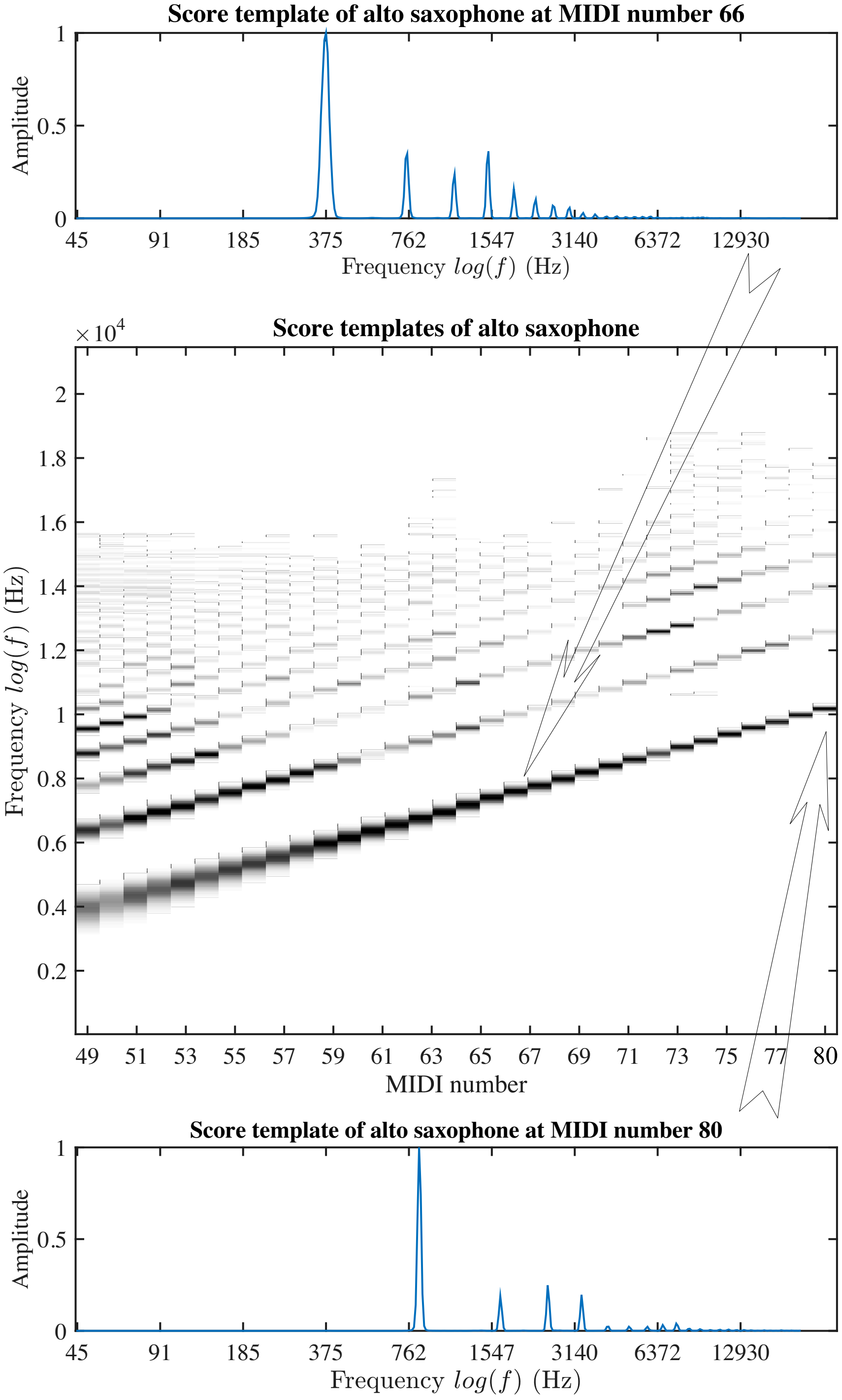


Figure 5.2: Harmonic components compression of an alto saxophone

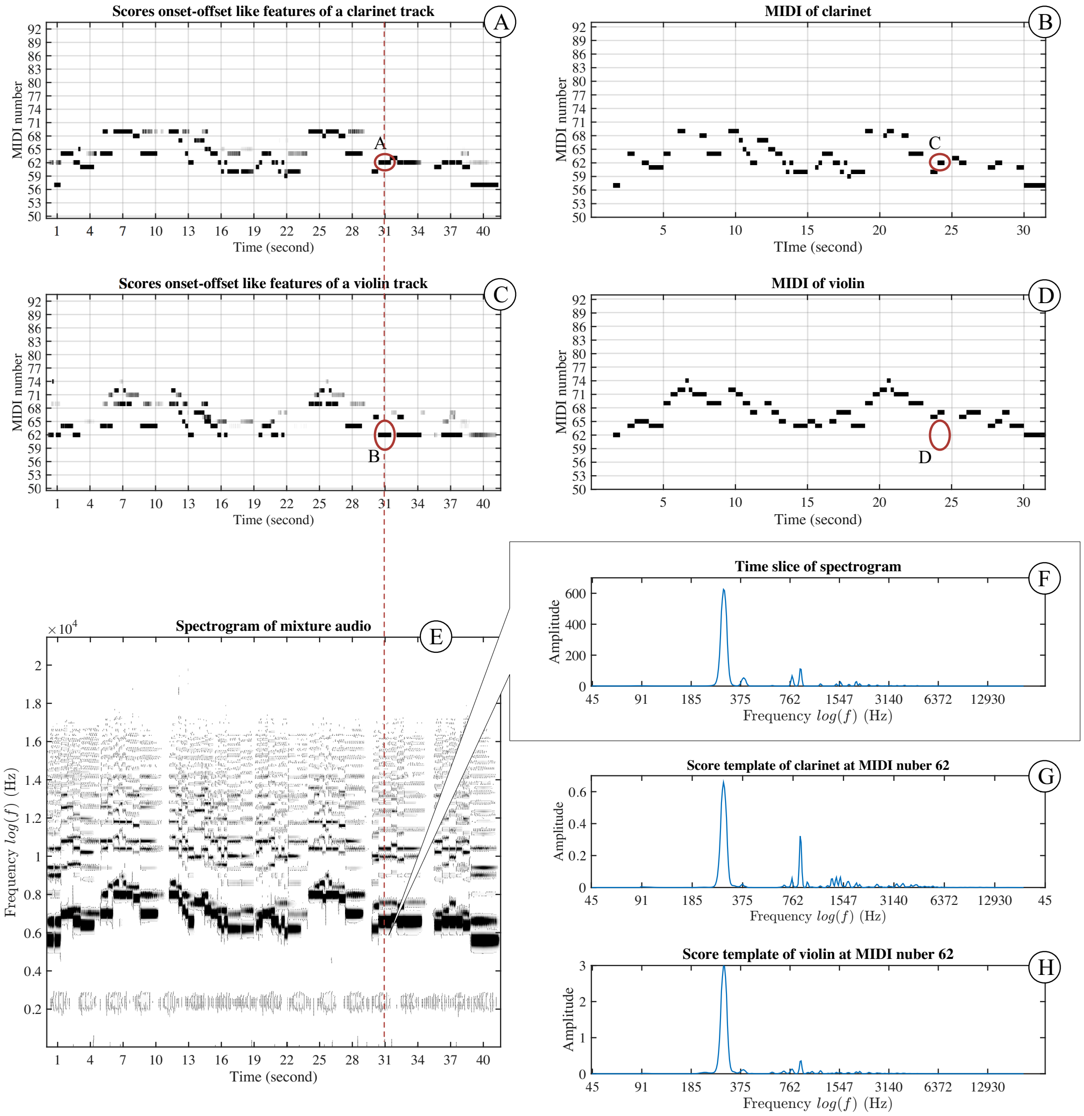
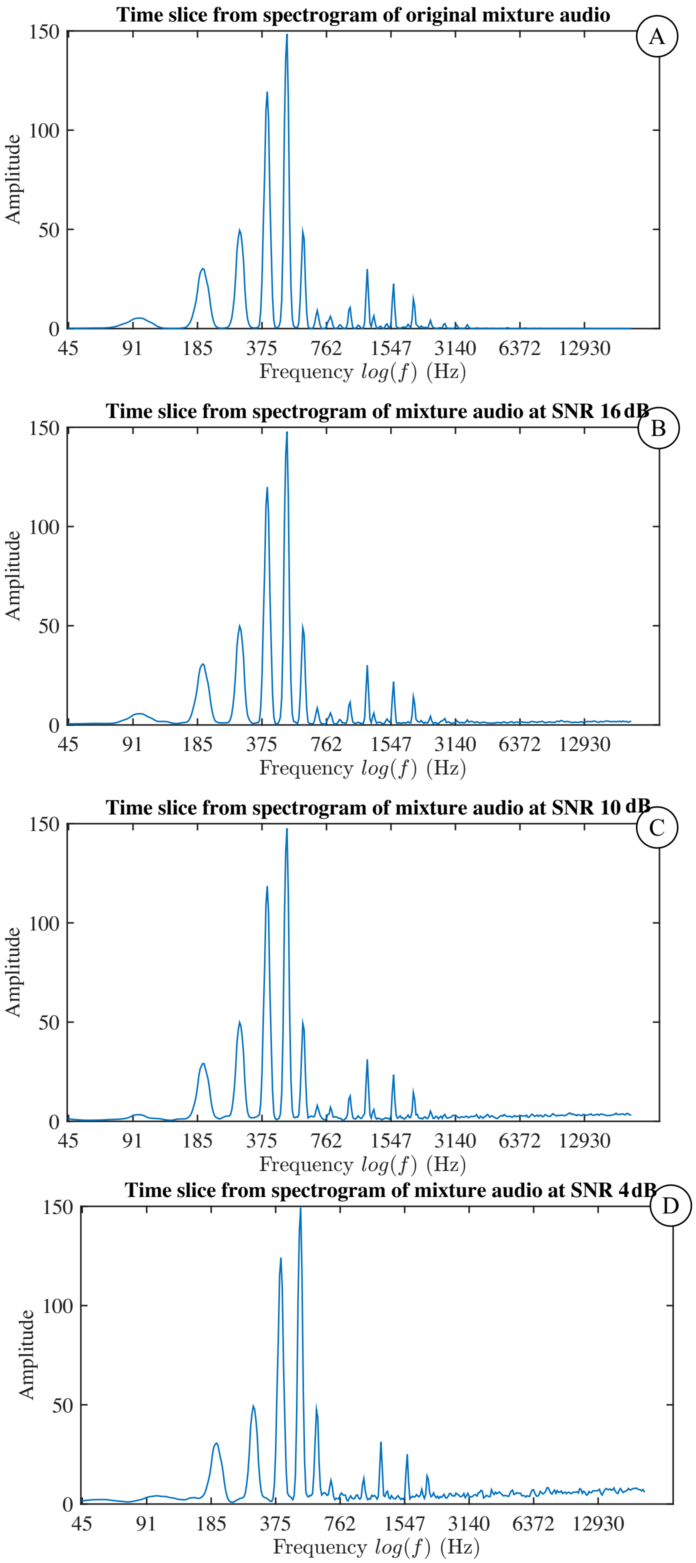


Figure 5.3: Misleading score onset-offset estimation because of harmonic components absence


Time slice from spectrogram of original mixture audio

Time slice from spectrogram of mixture audio at SNR 16 dB

Time slice from spectrogram of mixture audio at SNR 10 dB

Time slice from spectrogram of mixture audio at SNR 4 dB

Figure 5.4: Frequency profiles at specific time in spectrograms of original and simulated mixture audio at different level of SNR.

List of References

[1] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[2] 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.

[3] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *ISMIR*, vol. 2003, pp. 229–230, 2003.

[4] M. Goto, "Development of the RWC music database," in *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, vol. 1, pp. 553–556, 2004.

[5] M. Mørup and M. N. Schmidt, "Sparse non-negative matrix factor 2-d deconvolution," tech. rep., Technical University of Denmark, 2006.

[6] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-d deconvolution for blind single channel source separation," in *Independent Component Analysis and Blind Signal Separation*, pp. 700–707, Springer, 2006.

[7] M. Morup, K. H. Madsen, and L. K. Hansen, "Shifted non-negative matrix factorization," in *2007 IEEE Workshop on Machine Learning for Signal Processing*, pp. 139–144, IEEE, 2007.

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

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

[10] R. Jaiswal, D. Fitzgerald, E. Coyle, and S. Rickard, "Shifted NMF using an efficient constant-Q transform for monaural sound source separation," in *22nd IET Irish Signals and Systems Conference*, pp. 23–24, 2011.

[11] H. Kirchhoff, S. Dixon, and A. Klapuri, "Shift-variant non-negative matrix deconvolution for music transcription," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 125–128, IEEE, 2012.

[12] WIKIPEDIA, "Boxplotwiki." https://en.wikipedia.org/wiki/Box_plot. Accessed: 2017-03-20.

[13] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[14] D. Wu, N. Fang, C. Sun, X. Zhang, W. J. Padilla, D. N. Basov, D. R. Smith, and S. Schultz, "Terahertz plasmonic high pass filter," *Applied Physics Letters*, vol. 83, no. 1, pp. 201–203, 2003.

[15] K. Furutani, M. Kato, and T. Tsuru, "Low-pass filter," Sept. 16 1997. US Patent 5,668,511.

[16] F. Kobayashi and I. Umino, "Band pass filter," June 4 1991. US Patent 5,021,757.

[17] R. H. Bamberger and M. J. Smith, "A filter bank for the directional decomposition of images: Theory and design," *IEEE transactions on signal processing*, vol. 40, no. 4, pp. 882–893, 1992.

[18] S. Haykin, "Adaptive filter theory," *2nd. ed., Prentice-Hall, Englewood Cliffs, NJ*, 1991.

[19] S. Haykin, "Adaptive filter theory, 1996," *Telecommunication Systems, Radio Resource Management, Sensor Network and Particularly Their Applicable Issues to 4G Mobile Communication Systems and Cognitive Radio Systems*, pp. 12–13, 2000.

[20] 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.

[21] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

[22] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle, and S. Rickard, "Clustering NMF basis functions using shifted NMF for monaural sound source separation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, pp. 245–248, IEEE, 2011.

[23] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," tech. rep., Georgia Institute of Technology, 2008.

[24] S.-i. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Advances in neural information processing systems*, pp. 757–763, 1996.

[25] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean, “Boosting algorithms as gradient descent,” in *Advances in neural information processing systems*, pp. 512–518, 2000.

[26] C.-J. Lin, “On the convergence of multiplicative update algorithms for nonnegative matrix factorization,” *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1589–1596, 2007.

[27] 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.

[28] R. Badeau, N. Bertin, and E. Vincent, “Stability analysis of multiplicative update algorithms for non-negative matrix factorization,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, pp. 2148–2151, IEEE, 2011.

[29] S. Ewert and M. Müller, “Score-informed source separation for music signals,” *Multimodal music processing*, vol. 3, pp. 73–94, 2012.

[30] E. Vincent, R. Gribonval, and M. D. Plumbley, “BSS Oracle Toolbox Version 2.1.” `https://bass-db.gforge.inria.fr/bassoracle/`.

[31] Z. Duan, B. Pardo, and C. Zhang, “Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, 2010.

[32] Y.-G. Zhang and C.-S. Zhang, “Separation of music signals by harmonic structure modeling,” in *Advances in Neural Information Processing Systems*, pp. 1617–1624, 2006.

[33] T. Virtanen, “Unsupervised learning methods for source separation in monaural music signals,” *Signal Processing Methods for Music Transcription*, pp. 267–296, 2006.

[34] M. Viswanathan and M. Viswanathan, “Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale,” *Computer Speech & Language*, vol. 19, no. 1, pp. 55–83, 2005.

[35] R. C. Streijl, S. Winkler, and D. S. Hands, “Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives,” *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[36] I. REC, “Mean opinion score (MOS) terminology,” *International Telecommunication Union, Geneva*, 2006.

[37] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, “PEAQ-The ITU standard for objective measurement of perceived audio quality,” *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[38] C. Colomes, C. Schmidmer, T. Thiede, and W. C. Treurniet, "Perceptual quality assessment for digital audio: PEAQ-The New ITU standard for objective measurement of the perceived audio quality," in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*, Audio Engineering Society, 1999.

[39] M. Salovarda, I. Bolkovac, and H. Domitrovic, "Estimating perceptual audio system quality using PEAQ algorithm," in *Applied Electromagnetics and Communications, 2005. ICECom 2005. 18th International Conference on*, pp. 1–4, IEEE, 2005.

[40] P. M. Lavrador, N. B. de Carvalho, and J. C. Pedro, "Evaluation of signal-to-noise and distortion ratio degradation in nonlinear systems," *IEEE transactions on microwave theory and techniques*, vol. 52, no. 3, pp. 813–822, 2004.

[41] C. Févotte, R. Gribonval, and E. Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.

[42] 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.

[43] H.-T. Gao, T.-H. Li, K. Chen, W.-G. Li, and X. Bi, "Overlapping spectra resolution using non-negative matrix factorization," *Talanta*, vol. 66, no. 1, pp. 65–73, 2005.

[44] A. Tan, Y. Zhao, and X. Guo, "NIR spectroscopy and NMF algorithm for identification of oil pollutants in water," in *2014 International Conference on Mechatronics, Electronic, Industrial and Control Engineering (MEIC-14)*, Atlantis Press, 2014.

[45] R. Hennequin, R. Badeau, and B. David, "NMF with time–frequency activations to model nonstationary audio events," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744–753, 2011.

[46] M. D. Hoffman, "Poisson-uniform nonnegative matrix factorization," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5361–5364, IEEE, 2012.

[47] J. Zhang, L. Wei, Q. Miao, and Y. Wang, "Image fusion based on nonnegative matrix factorization," in *2004 International Conference on Image Processing, 2004 (ICIP'04)*, vol. 2, pp. 973–976, IEEE, 2004.

[48] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–273, ACM, 2003.

[49] C. C. Aggarwal and C. K. Reddy, *Data clustering: algorithms and applications.* CRC press, 2013.

[50] Y. Chen, M. Rege, M. Dong, and J. Hua, "Non-negative matrix factorization for semi-supervised data clustering," *Knowledge and Information Systems*, vol. 17, no. 3, pp. 355–379, 2008.

[51] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 792–799, ACM, 2005.

[52] P. O. Hoyer, "Non-negative sparse coding," in *Proceedings of the 2002 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557–565, IEEE, 2002.

[53] C. Lazar, A. Doncescu, and N. Kabbaj, "Non negative matrix factorisation clustering capabilities; application on multivariate image segmentation," *International Journal of Business Intelligence and Data Mining*, vol. 5, no. 3, pp. 285–296, 2010.

[54] T. Li and C. H. Ding, *Nonnegative Matrix Factorizations for Clustering: A Survey.* In Data Clustering Algorithms and Applications, Chapman and Hall/CRC, 2013.

[55] A. Mirzal, "NMF versus ICA for blind source separation," *Advances in Data Analysis and Classification*, pp. 1–24, 2014.

[56] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics & Amp; Data Analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[57] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints.," in *ICASSP*, pp. 5365–5368, 2012.

[58] J. Eggert and E. Korner, "Sparse coding and NMF," in *2004 IEEE International Joint Conference on Neural Networks, 2004. Proceedings*, vol. 4, pp. 2529–2533, IEEE, 2004.

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

[60] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[61] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 713–730, 2008.

[62] L. Rosasco, "Sparsity based regularization," tech. rep., MIT, 2010.

[63] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

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

[65] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010)*, pp. 4206–4209, IEEE, 2010.

[66] B. J. King, *New methods of complex matrix factorization for single-channel source separation and analysis*. PhD thesis, University of Washington, 2013.

[67] H. Li, T. Adali, W. Wang, and D. Emge, "Non-negative matrix factorization with orthogonality constraints for chemical agent detection in raman spectra," in *2005 IEEE Workshop on Machine Learning for Signal Processing*, pp. 253–258, IEEE, 2005.

[68] H. Li, T. Adal, W. Wang, D. Emge, and A. Cichocki, "Non-negative matrix factorization with orthogonality constraints and its application to raman spectroscopy," *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 48, no. 1-2, pp. 83–97, 2007.

[69] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *2008 IEEE International Joint Conference on Neural Networks (IJCNN 2008)*, pp. 1828–1832, IEEE, 2008.

[70] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons, and S.-I. Amari, "Novel multi-layer non-negative tensor factorization with sparsity constraints," in *Adaptive and Natural Computing Algorithms*, pp. 271–280, Springer, 2007.

[71] R. Jaiswal, *Non-Negative Matrix Factorization Based Algorithms to Cluster Frequency Basis Functions for Monaural Sound Source Separation*. PhD thesis, Dublin Institute of Technology, 2013.

[72] S. Kırbız and B. Günsel, "Perceptually enhanced blind single-channel music source separation by non-negative matrix factorization," *Digital Signal Processing*, vol. 23, no. 2, pp. 646–658, 2013.

[73] P. Kabal, "An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality," tech. rep., TSP Lab, Dept. Electrical & Computer Engineering, McGill University, 2002.

[74] T. Nakamura and H. Kameoka, "Shifted and convolutive source-filter non-negative matrix factorization for monaural audio source separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 489–493, IEEE, 2016.

[75] J. Fritsch, "High quality musical audio source separation," Master's thesis, Centre for Digital Music, Universitäts-und Landesbibliothek Bonn, 2012.

[76] S. Ewert, *Signal Processing Methods for Music Synchronization, Audio Matching, and Source Separation.* PhD thesis, Universitäts-und Landesbibliothek Bonn, 2012.

[77] C. Cannam, C. Landone, and M. Sandler, "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1467–1468, ACM, 2010.

[78] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, pp. 185–188, 2003.

[79] P. Cano, A. Loscos, and J. Bonada, "Score-performance matching using HMMs," in *Proceedings of the International Computer Music Conference*, pp. 441–444, 1999.

[80] N. Orio and D. Schwarz, "Alignment of monophonic and polyphonic music to a score," in *International Computer Music Conference (ICMC)*, pp. 1–1, 2001.

[81] S. Ewert, M. Muller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1869–1872, IEEE, 2009.

[82] C. Raphael, "Automatic segmentation of acoustic musical signals using hidden markov models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 4, pp. 360–370, 1999.

[83] Y.-x. Yuan, "Step-sizes for the gradient method," *AMS IP Studies in Advanced Mathematics*, vol. 42, no. 2, p. 785, 2008.

[84] C.-b. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[85] C. L. Byrne, "Auxiliary-function methods in iterative optimization," in *Lecture Notes on Iterative Optimization Algorithms*, Lowell: Department of Mathematical Sciences, University of Massachusetts Lowell, 2015.

[86] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[87] A. Cichocki, S. Cruces, and S.-i. Amari, "Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134–170, 2011.

[88] S. Kullback, "Letter to the editor: the Kullback−Leibler distance," *The American Statistician*, vol. 41, pp. 340–341, 1987.

[89] S. Kullback, *Information theory and statistics.* Courier Corporation, 1997.

[90] D. Commenges, "Information theory and statistics: An overview," tech. rep., Epidemiology and Biostatistics Research Center, INSERM, Bordeaux University, 2015.

[91] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," tech. rep., Tokyo Institute of Statistical Mathematics, Tokyo, Japan, 2001.

[92] Z. Chen, A. Cichocki, and T. M. Rutkowski, "Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer disease," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 5, pp. V–V, IEEE, 2006.

[93] A. Cichocki, S.-i. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He, "Extended smart algorithms for non-negative matrix factorization," in *Artificial Intelligence and Soft Computing–ICAISC 2006*, pp. 548–562, Springer, 2006.

[94] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," *Choice*, vol. 10, p. 1, 2010.

[95] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons, and S.-I. Amari, "Non-negative tensor factorization using alpha and beta divergences," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. III–1393, IEEE, 2007.

[96] A. Cichocki and S.-i. Amari, "Families of alpha-beta and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[97] P. Smaragdis and B. Raj, "Shift-invariant probabilistic latent component analysis," *Journal of Machine Learning Research*, 2007.

[98] M. Shashanka, *Latent variable framework for modeling and separating single-channel acoustic sources*. PhD thesis, Boston University, Boston, 2007.

[99] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence and Neuroscience*, vol. 2008, 2008.

[100] M. Shashanka, "Probabilistic latent variable model for sparse decompositions of non-negative data," *Computational Intelligence and Neuroscience*, 2008.

[101] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[102] R. Swinburne, *Bayes' theorem*. Oxford University Press, 2005.

[103] E. S. Yudkowsky, "An intuitive explanation of Bayes′ theorem." http://www.yudkowsky.net/rational/bayes. Accessed: 2016-10-19.

[104] J. Schnupp, I. Nelken, and A. King, *Auditory neuroscience: Making sense of sound*. MIT press, 2011.

[105] M. Davenport, M. Davenport, and S. Hannahs, *Introducing phonetics and phonology*. Routledge, 2010.

[106] C. J. Plack, A. J. Oxenham, and R. R. Fay, *Pitch: neural coding and perception*, vol. 24. Springer Science & Business Media, 2006.

[107] 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.

[108] C. Févotte, “Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1980–1983, IEEE, 2011.

[109] L. Fritts, “The University of Iowa electronic music studios musical instrument samples.” http://theremin.music.uiowa.edu/MISflute.html, 1997. Accessed: 2016-10-19.

[110] J. Snyman, *Practical mathematical optimization: An introduction to basic optimization theory and classical and new gradient-based algorithms*, vol. 97. Springer Science & Business Media, 2005.

[111] H. F. Pollard and E. V. Jansson, “A tristimulus method for the specification of musical timbre,” *Acta Acustica United with Acustica*, vol. 51, no. 3, pp. 162–171, 1982.

[112] M. Basu, “Gaussian-based edge-detection methods-a survey,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 32, no. 3, pp. 252–260, 2002.

[113] A. C. Berg and J. Malik, “Geometric blur for template matching,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–I, IEEE, 2001.

[114] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 2366–2369, IEEE, 2010.

[115] J. J. Koenderink, “The structure of images,” *Biological cybernetics*, vol. 50, no. 5, pp. 363–370, 1984.

[116] F. Durand and J. Dorsey, “Fast bilateral filtering for the display of high-dynamic-range images,” in *ACM transactions on graphics (TOG)*, vol. 21, pp. 257–266, ACM, 2002.

[117] C. Lopez-Molina, B. De Baets, H. Bustince, J. Sanz, and E. Barrenechea, “Multiscale edge detection based on Gaussian smoothing and edge tracking,” *Knowledge-Based Systems*, vol. 44, pp. 101–111, 2013.

[118] G. Taubin, “Curve and surface smoothing without shrinkage,” in *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pp. 852–857, IEEE, 1995.

[119] A. M. Wink and J. B. Roerdink, “Denoising functional MR images: a comparison of wavelet denoising and Gaussian smoothing,” *IEEE transactions on medical imaging*, vol. 23, no. 3, pp. 374–387, 2004.

[120] W. Wu, "Paralleled Laplacian of Gaussian (LoG) edge detection algorithm by using GPU," in *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, pp. 1003309–1003309, International Society for Optics and Photonics, 2016.

[121] C. Deng, X.-B. Gao, D.-C. Tao, and X.-L. Li, "Digital watermarking in image affine co-variant regions," in *Machine Learning and Cybernetics, 2007 International Conference on*, vol. 4, pp. 2125–2130, IEEE, 2007.

[122] T.-W. Chu, S.-F. Su, M.-C. Chen, S. S.-D. Xu, and K.-S. Hwang, "Edge enhanced sift for moving object detection," in *Informative and Cybernetics for Computational Social Systems (ICCSS), 2016 3rd International Conference on*, pp. 11–14, IEEE, 2016.

[123] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[124] MATLAB, *version 8.4.0 (R2014b)*. Natick, Massachusetts: The MathWorks Inc., 2014.

[125] A. Hugill, "The orchestra: a user's manual," 2004.

[126] B. David, "Composition," *http://www.compositiontoday.com/*.

[127] K. Peason, "On lines and planes of closest fit to systems of point in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[128] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[129] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[130] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[131] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[132] H. Kameoka, "Non-negative matrix factorization and its variants for audio signal processing," in *Applied Matrix and Tensor Variate Data Analysis*, pp. 23–50, Springer, 2016.

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

[134] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research," in *ISMIR*, vol. 14, pp. 155–160, 2014.

[135] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *ISMIR*, vol. 2, pp. 287–288, 2002.

[136] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 318–326, 2008.

[137] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking systems.," in *ISMIR*, pp. 315–320, 2009.

[138] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[139] G. E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 154–154, 2007.

[140] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Source separation by score synthesis.," in *ICMC*, 2010.

[141] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Evaluation of a score-informed source separation system.," in *ISMIR*, pp. 219–224, 2010.

[142] Z. Duan, J. Han, and B. Pardo, "Song-level multi-pitch tracking by heavily constrained clustering," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 57–60, IEEE, 2010.

[143] M. Cousins and R. Hepworth-Sawyer, *Logic Pro X: Audio and Music Production.* CRC Press, 2014.

[144] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

[145] D. H. Johnson, "Signal-to-noise ratio," *Scholarpedia*, vol. 1, no. 12, p. 2088, 2006.

[146] P. R. Griffiths and J. A. de Haseth, "Signal-to-noise ratio," *Fourier Transform Infrared Spectrometry, Second Edition*, pp. 161–175, 2007.