공학박사 학위논문

일반화된 디리클레 사전확률을 이용한 비지도적 음원 분리 방법

Unsupervised Approach to Music Source Separation using Generalized Dirichlet Prior

2018 년 2 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

박 정 수

공학박사 학위논문

일반화된 디리클레 사전확률을 이용한 비지도적 음원 분리 방법

Unsupervised Approach to Music Source Separation using Generalized Dirichlet Prior

2018 년 2 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

박 정 수

일반화된 디리클레 사전확률을 이용한 비지도적 음원 분리 방법

Unsupervised Approach to Music Source Separation using Generalized Dirichlet Prior

지도교수 이 교 구

이 논문을 공학박사 학위논문으로 제출함

2018 년 1 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

박 정 수

박정수의 공학박사 학위논문을 인준함

2018 년 1 월

위 원 장	이 원 종	(인)
부위원장	이 교 구	(인)
위 원	곽 노 준	(인)
위 원	남 주 한	(인)
위 원	이 석 진	(인)

Abstract

Music source separation aims to extract and reconstruct individual instrument sounds that constitute a mixture sound. It has received a great deal of attention recently due to its importance in the audio signal processing. In addition to its stand-alone applications such as noise reduction and instrument-wise equalization, the source separation can directly affect the performance of the various music information retrieval algorithms when used as a pre-processing. However, conventional source separation algorithms have failed to show satisfactory performance especially without the aid of spatial or musical information about the target source. To deal with this problem, we have focused on the spectral and temporal characteristics of sounds that can be observed in the spectrogram. Spectrogram decomposition is a commonly used technique to exploit such characteristics; however, only a few simple characteristics such as sparsity were utilizable so far because most of the characteristics were difficult to be expressed in the form of algorithms. The main goal of this thesis is to investigate the possibility of using generalized Dirichlet prior to constrain spectral/temporal bases of the spectrogram decomposition algorithms. As the generalized Dirichlet prior is not only simple but also flexible in its usage, it enables us to utilize more characteristics in the spectrogram decomposition frameworks. From harmonic-percussive sound separation to harmonic instrument sound separation, we apply the generalized Dirichlet prior to various tasks and verify its flexible usage as well as fine performance.

Keywords: Source separation, musical instrument, matrix decomposition, probabilistic latent component analysis, non-negative matrix factorization, Dirichlet prior

Student Number: 2012-31246

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Source separation in a digital signal processing aims to recover original signals of interest from given signal mixtures. It has been attracted considerable attention as a research topic in the past few decades and applied to many research fields [1]. The applications of source separation include music and audio analysis such as instrument-wise equalizing, stereo-to-surround up-mixing, karaoke systems, and crosstalk cancellation, biomedical signal analysis such as electroencephalographic (EEG) and electromyographic (EMG) [2, 3, 4], and chemical signal analysis [5]. Nevertheless, the mainstream of the recent source separation research focuses on the audio signal due to the easily overlapping nature of sound and its diverse applications.

In almost every situation we hear a variety of sounds that occur simultaneously, and humans are able to find meaningful information in the sounds.

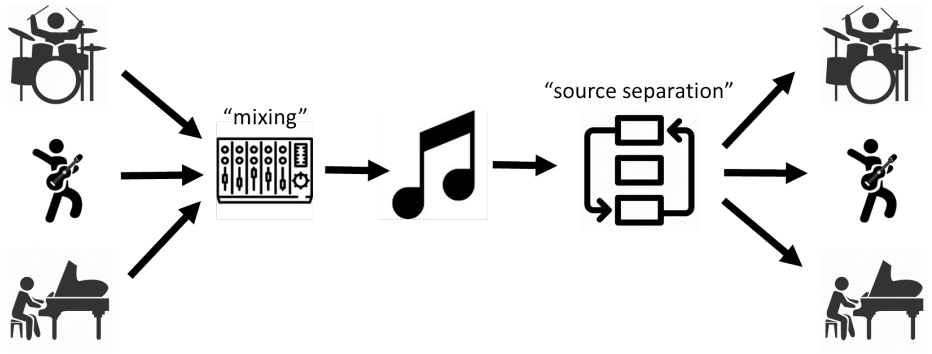


Fig. 1.1 Concept of music source separation.

"Cocktail party problem" demonstrates an interesting phenomenon concerning the human's capability of listening. Even in a noisy environment like the cocktail party, people are able to concentrate on a sound that they want to attend such as the voice of a person they are in conversation. This selective attention enables humans to catch crucial auditory information, with being insensitive to the magnitude of the sound. As this process happens in a human brain unconsciously, machines are not capable of imitating their magnitude-robust operation. Hence, in order to make the machines work correctly, pre-processing to amplify, or separate, the sounds of interest is necessary. This leads to the necessity of audio source separation, and this is why the source separation algorithms have an enormous impact on the audio signal processing and machine learning research.

In particular, the source separation has been extensively used in the field of music signal analysis. Music source separation can be comprehended as the opposite process of audio mixing that combines multitrack recordings as shown in Fig. 1.1. As a pre-processing method, it has contributed to the improvement of the various music information retrieval (MIR) algorithms enabling the ex-

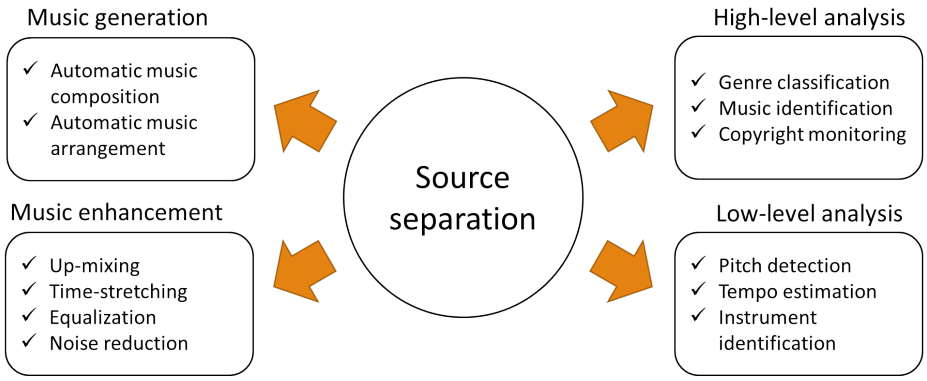


Fig. 1.2 Applications of source separation in music and audio research.

traction of musical information to work with high accuracy as shown in Fig. 1.2. Such algorithms include low-level music analyses such as pitch detection, tempo estimation, and instrument identification, and high-level analyses such as genre classification, music identification, and copyright monitoring [6]. Besides, it has also assisted music generation algorithms such as automatic music composition as well as music enhancement algorithms such as up-mixing, time-stretching, instrument-wise equalization, and noise reduction [7, 8]. In addition, the spectral and temporal characteristics of music signals are often considered stationary. Hence the algorithms to analyze the music signals can be extended to other applications to investigate noisier signals such as EEG signal.

According to the aforementioned necessity of music source separation, we have investigated its essential problems and presented the works in this thesis. In the rest of this chapter, we introduce problems in music source separation and describe the scope of our research. Then the motivation of our approach is presented. Finally, we summarize our major focuses and outline the following chapters.

1.2 Task of interest

In this section, we introduce the focus of our research in the music source separation field. To this end, we first categorize conventional approaches according to the two criteria: the number of channels and the use of side-information. Then our main task of interest is introduced as single-channel blind source separation.

1.2.1 Number of channels

When the input mixture signal is composed of multiple channels, the estimation of original sources can be achievable via spatial filtering. Conventional studies that use spatial filtering assume that the mixture signal is the linear combination [9] or convolutional mixing of [10, 11] of the individual sounds. In this case, the separation process is identical to obtaining the inverse of the actual mixing matrix. When the number of channels is equal to the number of original sources, the task is categorized as the *determined* case since the perfect reconstruction of the original sounds is theoretically available. This is also applied to the *overdetermined* case where the number of channels exceeds the number of sources. However, when the number of channels is smaller than the number of sources, which is referred to as *underdetermined* case, the perfect reconstruction is not possible via spatial filtering; hence, the assistance of spectro-temporal characteristics is necessary.

Since most of the music signals are comprised of a single-channel (*mono*) or two channels (*stereo*), the music source separation task is often presumed as underdetermined. Accordingly, intensive study about the single channel-based

music source separation methods is essential, and it is commonly utilized as an essential background of the multi-channel music source separation [12]. In this thesis, we concentrate on the single-channel scenario. We aim to show that proper utilization of the spectro-temporal information can greatly improve the separation performance even without the aid of spatial information.

1.2.2 Utilization of side-information

Meanwhile, the amount of spectro-temporal information we use is considered as an important criterion to categorize single-channel source separation studies. In the early stage of the music source separation, the blind approach was intensively investigated [13], where no additional information about the target source exists. These blind source separation (BSS) studies often assume that the target sound has certain statistical features such as non-Gaussianity and independence [14] or sparsity [15, 16, 13, 17]. BSS techniques are useful in some cases; however, such statistical assumptions cannot be guaranteed in many practical situations, which eventually causes performance degradation.

To overcome their low performance, informed source separation (ISS) was widely studied [18]. These studies assume situations where side-information about the target sources is available. Such information includes spectro-temporal characteristics such as music score [6, 19, 20], partial information such as onset [21], and direct information such as manually provided annotations [22, 23, 24] and user-guided audio signal [25]. Some studies even assumed that the information about the sources can be embedded inside the music signal by employing watermarking or encoding step [26, 27, 28].

Especially, recent approaches have examined the effect of artificial neural

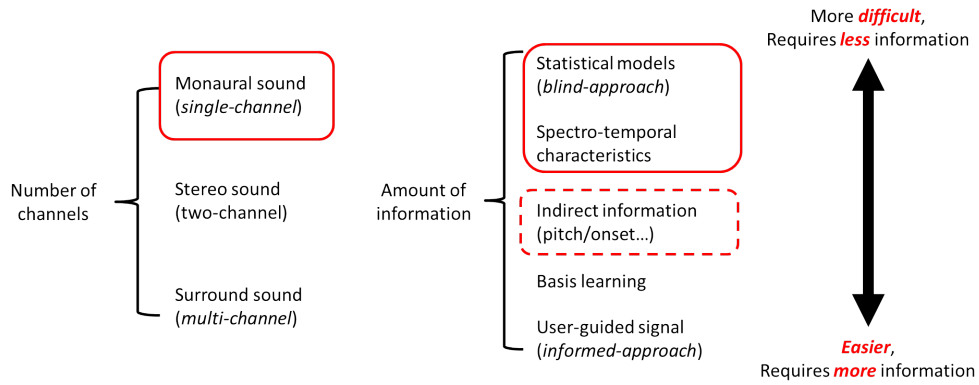


Fig. 1.3 Focus of our research.

network-based methods on the source separation tasks. While some of the research efforts directly applied the deep learning-based techniques like autoencoder as Lim and Lee's work [29], Osako *et al.*'s work [30], and Grais and Plumbley's work [31], others attempted to enhance conventional approaches like matrix decomposition [32] and time-frequency mask [33].

However, it is still a big issue to reduce the amount of information needed for successful separation, as the situation where side-information can be sufficiently provided is highly limited. Especially in deep learning-based approaches, a significant amount of training data is required for each sound source. In this thesis, we focus on using a least amount of side-information while maintaining satisfactory performance. Fig. 1.3 summarizes the focus of our research.

1.3 Approach

In this section, our approach to the unsupervised single-channel music source separation is described. This section consists of three subsections that intro-

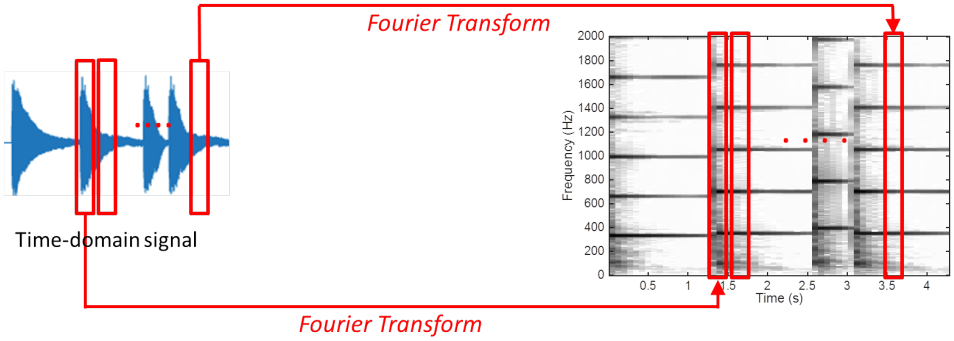


Fig. 1.4 Generation of a spectrogram with a time-domain signal.

duce constrained spectrogram decomposition techniques, Dirichlet prior, and our contribution.

1.3.1 Spectrogram decomposition with constraints

As most of the conventional approaches concerning the single-channel music source separation problem, we focus on the spectro-temporal characteristics of the music signals. However, it is unable to examine the spectral characteristics with the time-domain signals. Hence, we often convert the input audio signal to *spectrogram* in the first step. Fig. 1.4 describes how a spectrogram is generated from a time-domain signal. The input time-domain signal is windowed and transformed to the frequency domain via fast Fourier transform (FFT). As the transform is carried out for all segments of short duration, the spectrogram can represent temporal transitions as well as spectral characteristics.

In other MIR-related studies, other two-dimensional representations such as mel-spectrogram [34], constant-Q spectrogram [35], and mel-frequency cepstral coefficients (MFCC) [36] were used for the same purpose. However, feasibility of

inverse operation is also important for the source separation, since the output spectrograms have to be converted to time-domain signals in the end. As the spectrogram is invertible unlike other two-dimensional representations, it is generally used for the source separation research.

As Fourier transform is one of the linear transformations, linearity of the sound signals is preserved. The spectrogram components are complex numbers, however, the phase information is generally removed to simplify the analysis. By doing so, the linearity is no longer preserved, but conventional studies have often assumed that the linearity between the magnitude spectrograms is approximately preserved. In this case, a magnitude spectrogram is mathematically interpreted as a *non-negative matrix*. In the rest of this thesis, the term "spectrogram" is used to indicate magnitude spectrogram.

In order to separately investigate and utilize the spectral and temporal characteristics of the spectrogram, *matrix decomposition* is widely used. The matrix decomposition aims to obtain matrices that approximate the original spectrogram when multiplied. From a practical point of view, it is equivalent to learning the spectral and temporal bases that represent characteristics of each side from the given spectrogram. Probabilistic latent component analysis (PLCA), probabilistic latent semantic analysis (PLSA), and non-negative matrix factorization (NMF) are the representative matrix decomposition algorithms. The algorithms use iterative update to estimate matrices that can accurately approximate the spectrogram.

Fig. 1.5 shows the illustration of a spectrogram and the results of the matrix decomposition. As shown in the figure, a spectrogram can be approximated as a multiplication of spectral bases ($\mathbf{W}$) and corresponding activations ($\mathbf{H}$). Then

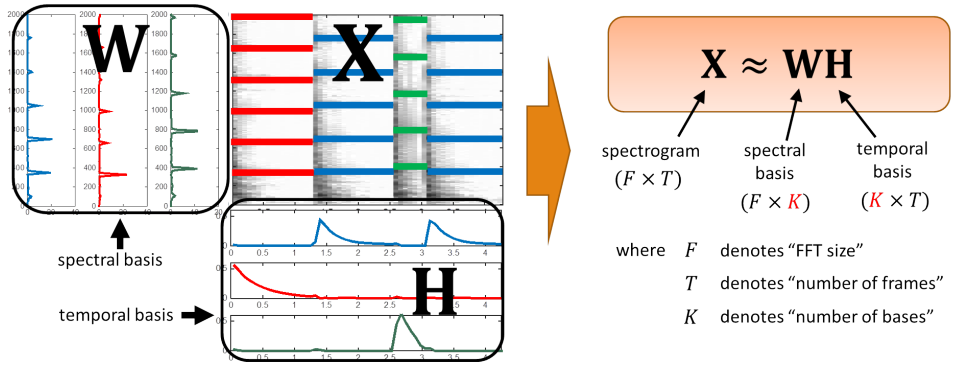


Fig. 1.5 Illustration of spectrogram decomposition.

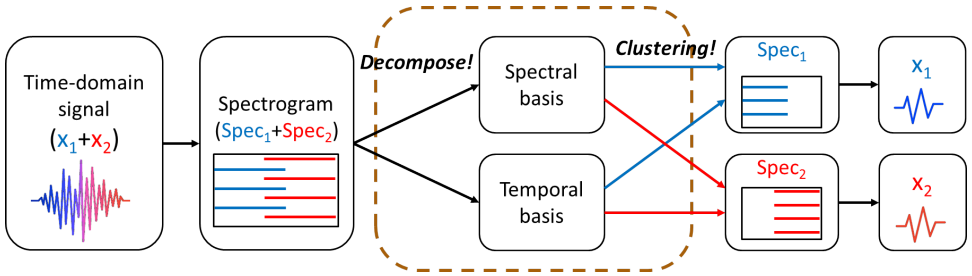


Fig. 1.6 Overall procedure of source separation based on spectrogram decomposition.

the estimated bases are clustered to reconstruct the spectrogram of each source as in [37, 38]. The overall procedure of the spectrogram decomposition-based source separation is shown in Fig. 1.6.

Nevertheless, this *decompose and cluster* strategy does not guarantee stable performance. In general, the matrix decomposition has infinite number of solutions because sufficient number of bases are often given. This is due to the fact that the optimal number of bases cannot be predicted in advance. Therefore, the random initialization of the bases causes the spectrogram decomposition to make different outputs each time. Fig. 1.7 illustrates the unstableness of the

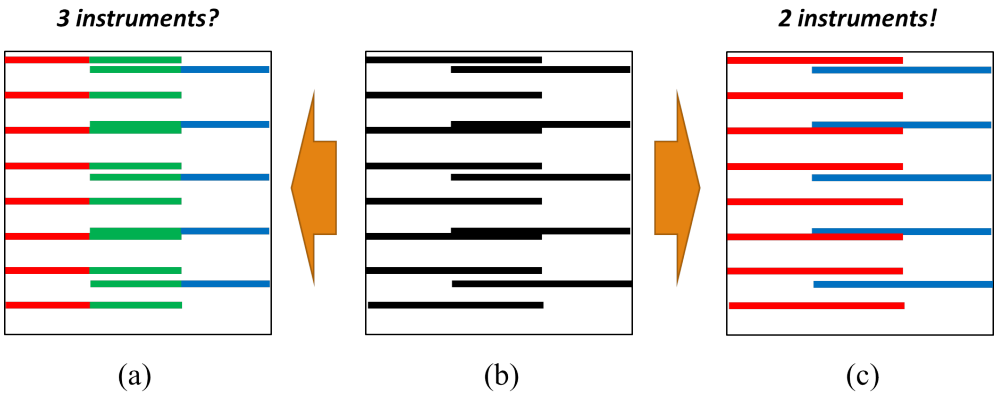


Fig. 1.7 Illustration of (b) original spectrogram, (a) undesired decomposition, and (c) intended decomposition.

spectrogram decomposition. In this illustration, sounds from two instruments are assumed to be mixed with partial overlap as Fig. 1.7 (c). As can be confirmed in the figure, the decomposition can result in learning musically meaningless bases as Fig. 1.7 (a). These bases can interrupt the clustering process and result in the degradation of the separation performance.

To alleviate this problem, constraints are often imposed to the spectral and temporal bases during the iteration. Fig. 1.8 illustrates the effect of constraint imposition. The constraints prevent bases from converging to the meaningless points that are close to the initial location. Instead, they enforce the bases to have a certain structure, which may reflect the characteristics of the target source. In order to make the constraint imposition work properly, the bases are partitioned into several groups in advance and enforced with different constraints. Through the constrained iteration, we can estimate and cluster the bases simultaneously.

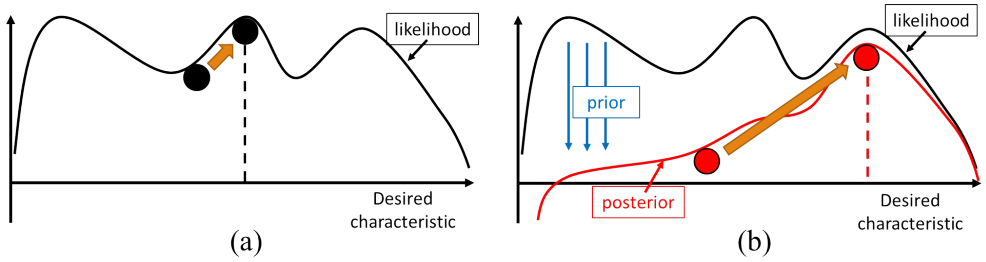


Fig. 1.8 Effect of imposing prior as a constraint.

1.3.2 Dirichlet prior

Our approach uses *Dirichlet prior* as it is the simple but effective constraint imposition. The term "prior" is originated from the probability theory and refers to the predetermined probability of an event to occur. In the PLCA or PLSA framework, constraints can be imposed in a form of prior probability. Since their iterative update equations can be derived by applying the principle of maximum likelihood (ML), including prior probability makes the estimation follow maximum a posteriori (MAP) estimation criterion.

Dirichlet prior has been adopted in the source separation as in Smaragdis and Mysore's work [25] and Kim and Smaragdis' work [39]. However, its applications were limited to shaping the bases like the trained ones [25] and making the bases to have sparse structure [39]. According to our observation, the Dirichlet prior has much more potential to be used for various purposes. This can be clearly addressed with the comparison with other constraint imposition methods. Table 1.1 shows the comparison between the Dirichlet prior and the entropic prior which is used to enforce the basis sparseness. It can be seen that not only the Dirichlet prior is more flexible, it also requires less computations. Motivated by these advantages of using the Dirichlet prior, we aim to apply it to

Table 1.1 Comparison of Dirichlet prior and entropic prior.

	Entropic prior	Dirichlet prior
Flexibility	Low (Only for sparsity imposition)	High (Can be used for many purposes)
Computation complexity	High (Includes iterative operation)	Low (Simple weighted summation)

various tasks from harmonic-percussive sound separation (HPSS) to harmonic instrument sound separation (HISS).

1.3.3 Contribution

The major contributions of the studies presented in this thesis can be summarized as follows:

1. **Unsupervised approach**: Unlike conventional studies that focus on the informed source separation task and the neural network-based source separation methods, our approach concentrates on the blind case where least information is required. Through the efficient utilization of the suitable spectro-temporal information, we can achieve fine performance with the reduced amount of information. To this end, we investigate the spectro-temporal characteristics and use it for the separation. This approach also differs from the conventional statistical characteristics-based blind source separation methods [40] such as independent component analysis (ICA) in that the features used are predefined by humans.

2. **Utilization of novel spectro-temporal characteristics**: The spectro-temporal characteristics of the musical sources used in the studies – harmonicity, sparsity, unsparsity, continuity, discontinuity, and spectral envelope – are not only musically meaningful but also novel; especially, harmonicity, continuity, and discontinuity were never considered in the conventional studies. With the aid of these characteristics, we can achieve significant performance improvement.

3. **Simple implementation**: The implementation of the algorithms is simplified using the concept of the Dirichlet prior. Conventional studies have often reinduced the iterative update formula because they often attempted to give constraints by modifying the cost function and the changes in the cost function require the entire formula to be reinduced. In this thesis, we use the Dirichlet prior to simplify the implementation of the constraint imposition. By adopting the Dirichlet prior, we do not have to reinduce the entire equations to impose constraints; we can achieve it by setting the exemplar hyperparameters to the values suitable to the task. Furthermore, we extend the concept of the Dirichlet prior to the NMF framework, which enables further reduction of the computational complexity.

1.4 Outline of the thesis

Chapter 2 provides mathematical descriptions about the core concepts – PLCA, NMF, and Dirichlet prior – that are frequently used in the following chapters. Derivations of the matrix decomposition algorithms are presented in the first place. Next, how the Dirichlet prior changes the iterative update equa-

tions of PLCA is described. Then the application of Dirichlet prior is extended to the NMF according to the equivalence of PLCA and NMF. This extension is meaningful in that it dramatically reduces the computational cost.

Chapter 3 examines the HPSS problem focusing on the spectral-side difference of the harmonic and percussive sounds. As the harmonic sounds appear to have periodical energy distribution in the spectral domain whereas the percussive sounds have flat spectrum, these characteristics are imposed through the Dirichlet prior. Conventional HPSS methods have focused on temporal continuity of the harmonic components and spectral continuity of the percussive components. However, it may not be appropriate to use them to separate time-varying harmonic signals such as vocals, vibratos, and glissandos, as they lack in temporal continuity. With the proposed algorithm, we successfully separate the rapidly time-varying harmonic signals from the percussive signals by imposing different constraints on the two disjoint groups of the spectral bases. Experiments with real recordings as well as synthesized sounds show that the proposed method outperforms the conventional methods.

Chapter 4 also discusses HPSS and presents a novel method that exploits continuity/discontinuity properties in the matrix decomposition framework. It is widely accepted in the HPSS research that the harmonic and percussive components have anisotropic characteristics; the spectra of the harmonic sounds and the time activations of the percussive sounds have uneven energy distribution, whereas the spectra of the percussive sounds and the time activations of the harmonic sounds are smooth in their shapes. However, conventional methods fail to fully utilize the characteristics leading to the suboptimal performance.

Based on the observations that not the degree of sparseness but the degree of fluctuation is an accurate measure for distinguishing the harmonic and percussive components, we propose a novel HPSS algorithm by incorporating the continuity control in the iterative update formula of the matrix decomposition algorithm. The comparative evaluation results show that the proposed method outperforms conventional methods in terms of both objective and subjective evaluation.

Chapter 5 presents an informed approach to HISS problem. As the HISS is a more complicated task compared to the HPSS, we use the assistance of side-information. Since *spectral envelope* is one of the features that best represent an instrument according to the source-filter model, it is imposed on the spectral bases in the NMF algorithm in order for them to have the spectral envelope of the target instrument. In so doing, the spectral envelopes are estimated via linear predictive coding (LPC). As the iteration proceeds, the spectral bases are shaped to have the extracted envelope of the target instrument. The proposed approach is evaluated using the real recordings and shows outperforming results over the conventional methods.

Chapter 6 extends the HISS method presented in the previous chapter to the blind approach. Unlike the informed approach, the spectral envelope of the target instrument is not given. Instead, spectral bases of the NMF algorithm are grouped in advance, and the bases belong to a group are forced to have the same spectral envelope. In this way, the spectral envelopes and the details of the bases simultaneously converge to the target instruments' envelopes and details, respectively. In addition, the proposed method does not require pre-training

or post-processing because the estimation and the clustering of the spectral bases can be performed simultaneously in a single spectrogram decomposition framework. The comparative evaluation results with real recordings show that the proposed method outperforms the conventional methods.

Chapter 2

Theoretical background

In this chapter, we present descriptions about the important definitions and theories related to the matrix decomposition techniques. This chapter is composed of two categories: matrix decomposition techniques and their variations with the Dirichlet prior. As aforementioned in the introduction, spectrogram decomposition-based method is one of the major approaches in the source separation. Hence, we first describe two spectrogram decomposition techniques – PLCA and NMF – with derivations. In the next section, how the update equations of the matrix decomposition methods change with the application of the Dirichlet prior is described.

This chapter is organized as follows. At first PLCA and NMF algorithms are described in detail with mathematical derivations. Then we present the concept of Dirichlet prior followed by its application to the PLCA framework. Finally, we generalize the concept of the Dirichlet prior to the NMF framework on the basis of the duality between PLCA and NMF.

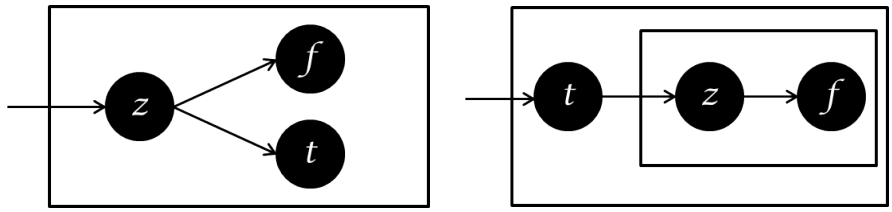
2.1 Probabilistic latent component analysis

The central concept of the PLCA algorithm is to estimate the latent distributions that constitute the parameter model θ. This method interprets a magnitude spectrogram as a multivariate distribution representing a histogram that is generated from the one-dimensional marginal probability distributions [41].

The generative model of the PLCA can be easily explained with the concept shown in Fig. 2.1 (a). According to it, the generation of the magnitude spectrogram is modeled with the following steps: 1. Draw z with the probability $p(z)$ 2. Draw f and t with the marginal probability distributions $p(t|z)$ and $p(f|z)$ 3. Iterate step 1 and step 2. Mathematically, the latent variable z is determined at first, and then the frame index t and the frequency bin index f is simultaneously determined. This is discriminated from the PLSA of which generative model is presented in Fig. 2.1 (b). Each z is considered to have a two-dimensional probability distribution, and the magnitude spectrogram can be generated via weighted sum of them. According to this model, the magnitude spectrogram can be mathematically interpreted as

$$\begin{aligned} \mathbf{X}_{t,f} &= p(f,t) \\ &= \sum_{z=1}^{K} p(z)\, p(f,t|z) \\ &= \sum_{z=1}^{K} p(z)\, p(t|z)\, p(f|z) \end{aligned} \tag{2.1}$$

where $\mathbf{X}_{t,f}$ denotes the (t,f)-th element of the magnitude spectrogram $\mathbf{X}$, $p(z)$ denotes the marginal distribution of the latent variable z, and $p(t|z)$ and $p(f|z)$ denote the marginal distributions of t and f (respectively) for the given value of z. This is due to our assumption that t and f are independently determined.



(a) (b)

Fig. 2.1 Generative model of (a) PLCA and (b) PLSA.

Unlike PLSA that assumes t, z, and f are determined sequentially, PLCA assumes the two-dimensional normalization of the spectrogram because t and f are simultaneously determined.

According to this analysis, X can be represented as the multiplication of three matrices as

$$\mathbf{X} = \mathbf{VSG} \tag{2.2}$$

where $\mathbf{V}$ is a $(F \times K)$ matrix of which k-th column is $p\left(f|z = z_k\right)$, and $\mathbf{G}$ is a $(K \times T)$ matrix of which k-th row is $p\left(t|z = z_k\right)$, and $\mathbf{S}$ is a diagonal matrix of which (k,k)-th element is $p\left(z = z_k\right)$. Note that this is the linear algebraic expression and will be compared to the NMF later. From this formula, it can be observed that it can have multiple solutions including the extraordinary and meaningless solutions as $\mathbf{V} = \mathbf{X}$, and $\mathbf{S} = \mathbf{G} = \mathbf{I}_{T\times T}$ when $K = T$.

To derive the iterative update equations, we maximize the log-likelihood that can be represented as

$$\begin{aligned} L &= \kappa \log p\left(\mathbf{X}|\theta\right) \\ &= \kappa \sum_{z=1}^{K} \sum_{f,t} \mathbf{X}_{f,t} p\left(z|f,t\right) \log\left\{p\left(z\right) p\left(f,t|z\right)\right\} \end{aligned} \tag{2.3}$$

where κ is the normalizing constant of $\mathbf{X}$. We define the re-scaled log-likelihood

as $L' = L/\kappa$. Now, we aim to get the variables maximizing this log-likelihood by means of an expectation-maximization (EM) algorithm. For the expectation step, the *a posteriori* probability is given as

$$p\left(z|f,t\right) \leftarrow \frac{p\left(z\right)p\left(f|z\right)p\left(t|z\right)}{\sum\limits_{z} p\left(z\right)p\left(f|z\right)p\left(t|z\right)}. \tag{2.4}$$

To derive the maximization step equations, we adopt a Lagrange multiplier method with normalization constraints. The Lagrange function can be represented as

$$\begin{aligned} & f_{Lagrange}\left(L', p\left(f|z\right), p\left(z\right), p\left(t|z\right), \lambda_f, \lambda_z, \lambda_t\right) \\ & = L' + \lambda_f\left(1 - \sum_{f} p\left(f|z\right)\right) + \lambda_z\left(1 - \sum_{z} p\left(z\right)\right) \\ & + \lambda_t\left(1 - \sum_{t} p\left(t|z\right)\right). \end{aligned} \tag{2.5}$$

To maximize the Lagrange function, the partial derivatives of $f_{Lagrange}$ must be zero, which can be calculated as

$$\frac{\partial f_{Lagrange}}{\partial p\left(f|z\right)} = \sum_{t} \frac{\mathbf{X}_{f,t} p\left(z|f,t\right)}{p\left(f|z\right)} - \lambda_f \tag{2.6}$$

$$\frac{\partial f_{Lagrange}}{\partial p\left(z\right)} = \sum_{f,t} \left\{ \frac{\mathbf{X}_{f,t} p\left(z|f,t\right)}{p\left(z\right)} \right\} - \lambda_z \tag{2.7}$$

$$\frac{\partial f_{Lagrange}}{\partial p\left(t|z\right)} = \sum_{f} \left\{ \frac{\mathbf{X}_{f,t} p\left(z|f,t\right)}{p\left(t|z\right)} \right\} - \lambda_t. \tag{2.8}$$

According to the normalization condition, the Lagrange multipliers are obtained as

$$\lambda_f = \sum_{f} \left\{ \sum_{t} \mathbf{X}_{f,t} p\left(z|f,t\right) \right\} \tag{2.9}$$

$$\lambda_z = \sum_{z} \left\{ \sum_{f} \sum_{t} \mathbf{X}_{f,t} p\left(z|f,t\right) \right\} \tag{2.10}$$

$$\lambda_t = \sum_t \left\{ \sum_f \mathbf{X}_{f,t} p\left(z|f,t\right) \right\}. \tag{2.11}$$

Thus, the iterative update equations of the maximization step of PLCA are derived as

$$p\left(f|z\right) \leftarrow \frac{\sum\limits_t \mathbf{X}_{f,t} p\left(z|f,t\right)}{\sum\limits_f \left\{ \sum\limits_t \mathbf{X}_{f,t} p\left(z|f,t\right) \right\}} \tag{2.12}$$

$$p\left(z\right) \leftarrow \frac{\sum\limits_f \sum\limits_t \mathbf{X}_{f,t} p\left(z|f,t\right)}{\sum\limits_z \left\{ \sum\limits_f \sum\limits_t \mathbf{X}_{f,t} p\left(z|f,t\right) \right\}} \tag{2.13}$$

$$p\left(t|z\right) \leftarrow \frac{\sum\limits_f \mathbf{X}_{f,t} p\left(z|f,t\right)}{\sum\limits_t \left\{ \sum\limits_f \mathbf{X}_{f,t} p\left(z|f,t\right) \right\}}. \tag{2.14}$$

Alteration of the expectation step and the maximization step makes the variables converge.

Note that the iterative update equations require the parameters to be initialized in advance of the iteration. Also, the decomposition results vary according to the initial condition, since it can have multiple valid solutions. Among the possible solutions, it is important to distinguish meaningful information.

2.2 Non-negative matrix factorization

The NMF algorithm aims to decompose a matrix that contains non-negative elements as multiplication of two non-negative matrices. It is commonly used for the audio source separation by interpreting magnitude or power spectrogram as a matrix to be decomposed [42], [43] similar to PLCA algorithm. In such cases, the outputs of the NMF would be the matrix of spectral bases whose columns

may denote the frequency domain representation of the source sounds, and the matrix of temporal bases whose rows denote the time activations or mixture weights of the corresponding frequency bases. In mathematical formula, it can be represented as

$$\begin{aligned} \mathbf{X} &\approx \tilde{\mathbf{X}} \\ &= \mathbf{WH} \end{aligned} \tag{2.15}$$

where $\mathbf{X}$ denotes the estimated magnitude spectrogram, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K]$ denotes an $(F \times K)$ non-negative matrix where $\mathbf{w}_k$ is its k-th column, and $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_K]^H$ denotes an $(K \times T)$ non-negative matrix where $\mathbf{h}_k$ is its k-th row that denotes the temporal activation of $\mathbf{w}_k$. The number of bases K is often considered to be much smaller than the number of frequency bins F and the number of frames T.

The error between the original matrix $\mathbf{X}$ and the reconstructed one $\tilde{\mathbf{X}}$, which is measured by cost function, converges to the minimum value as iteration proceeds. The multiplicative update rule for NMF algorithm was presented by Lee and Seung [44] for the case where Euclidean distance or Kullback-Leibler (KL) divergence was used. Here, we consider KL divergence for the cost function, which is widely used in conventional source separation methods [45] since it is better to be used to approximate the spectrogram. The KL divergence of arbitrary matrices $\mathbf{A}$ and $\mathbf{B}$ is represented as

$$D_{KL}\left(\mathbf{A}||\mathbf{B}\right) = \sum_{m,n} \left\{ \mathbf{A}_{m,n} \log \frac{\mathbf{A}_{m,n}}{\mathbf{B}_{m,n}} - \mathbf{A}_{m,n} + \mathbf{B}_{m,n} \right\} \tag{2.16}$$

where m and n are the row and column index, respectively.

The multiplicative update rule for the minimization of KL divergence is

represented as

$$\mathbf{H}_{k,t}^{(l+1)} \leftarrow \frac{\mathbf{H}_{k,t}^{(l)} \sum_{f} \left\{ \mathbf{W}_{f,k}^{(l)} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{f'} \mathbf{W}_{f',k}^{(l)}} \tag{2.17}$$

$$\mathbf{W}_{f,k}^{(l+1)} \leftarrow \frac{\mathbf{W}_{f,k}^{(l)} \sum_{t} \left\{ \mathbf{H}_{k,t}^{(l+1)} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{t'} \mathbf{H}_{k,t'}^{(l+1)}}. \tag{2.18}$$

where l denotes the iteration index. Note that each update equation is computed for all elements before we move on to the next equation. These two equations are iteratively calculated until the matrices converge. We can either set the number of iterations as in [37] or let the iteration stop when the distance between $\mathbf{X}$ and $\tilde{\mathbf{X}}$ falls below the certain threshold as in [46].

From 2.2 and 2.15, the similarity of PLCA and NMF is intuitively observed. As $\mathbf{S}$ is a diagonal matrix, its function is limited to scaling up or down the bases. Hence, if we define $\mathbf{W} = \mathbf{VS}$ and $\mathbf{H} = \mathbf{G}$, PLCA becomes equivalent to NMF. On the other hand, If we normalize each of the columns of $\mathbf{W}$ and each of the rows of $\mathbf{H}$, they can be interpreted as $\mathbf{V}$ and $\mathbf{G}$, respectively. From these facts, we can say that there is a duality between PLCA and NMF. Further analysis about this equivalence is described in Ding *et al.*'s work [47].

2.3 Dirichlet prior

In probabilistic analysis framework, imposition of Dirichlet prior is a proper and convenient method for shaping spectral/temporal bases of the matrix decomposition techniques. By giving the prior information about the bases' shapes, we can roughly determine the bases where to converge. This is due to the fact that the probabilistic analysis-based matrix decomposition techniques such as PLCA

or PLSA interpret the magnitude spectrogram as a histogram of the multinomial distribution, and Dirichlet prior is a conjugate prior of the multinomial distribution.

Without the use of Dirichlet prior, several methods have been frequently used to shape the bases. As in the [48], we can reinduce the update equations to make bases sparse/smooth. Or, as in the Shashanka [49] and Smaragdis's method [50], we can impose entropic prior to obtain sparse bases. However, they are limited to the sparsity imposition, so it is not appropriate to use them to impose continuity and discontinuity characteristics. For these reasons, we have chosen to use the Dirichlet prior in our research.

2.3.1 PLCA framework

In many applications, it is desirable to shape the marginal distributions representing the spectral or temporal bases to assign their characteristics. In such cases, additional information about the marginal distributions must be provided as a form of *prior distribution*. When the log-prior R is given, we can maximize the log-posterior P, which is represented as

$$\begin{aligned} P &= L' + R \\ &= \log p\left(\mathbf{X}|\theta\right) + \log p\left(\theta\right) \end{aligned} \tag{2.19}$$

where θ denotes the set of parameters. Now, we aim to get the variables maximizing this log-posterior by means of EM algorithm. Here, we assume that the prior distribution is determined in the form of a Dirichlet distribution as

$$p\left(\theta\right) = \prod_{z} \left\{ \prod_{f} p(f|z)^{c_f \xi(f|z)} \times p(z)^{c_z \psi(z)} \times \prod_{t} p(t|z)^{c_t \zeta(t|z)} \right\} \tag{2.20}$$

where c_f, c_z, and c_t are the constant coefficients, and $\xi(f|z)$, $\psi(z)$, and $\zeta(t|z)$ are the *exemplar hyperparameters*.

In this case the Lagrange function in the former formation changes as

$$\begin{aligned} & f_{Lagrange}\left(P, p\left(f|z\right), p\left(z\right), p\left(t|z\right), \lambda_f, \lambda_z, \lambda_t\right) \\ & = P + \lambda_f \left(1 - \sum_f p\left(f|z\right)\right) + \lambda_z \left(1 - \sum_z p\left(z\right)\right) \\ & + \lambda_t \left(1 - \sum_t p\left(t|z\right)\right). \end{aligned} \tag{2.21}$$

To maximize this Lagrange function, the partial derivatives of it must be zero, which is represented as

$$\frac{\partial f_{Lagrange}}{\partial p\left(f|z\right)} = \frac{c_f \xi\left(f|z\right)}{p\left(f|z\right)} + \sum_t \frac{\mathbf{X}_{f,t} p\left(z|f,t\right)}{p\left(f|z\right)} - \lambda_f \tag{2.22}$$

$$\frac{\partial f_{Lagrange}}{\partial p\left(z\right)} = \frac{c_z \psi\left(z\right)}{p\left(z\right)} + \sum_{f,t} \left\{ \frac{\mathbf{X}_{f,t} p\left(z|f,t\right)}{p\left(z\right)} \right\} - \lambda_z \tag{2.23}$$

$$\frac{\partial f_{Lagrange}}{\partial p\left(t|z\right)} = \frac{c_t \zeta\left(t|z\right)}{p\left(t|z\right)} + \sum_f \left\{ \frac{\mathbf{X}_{f,t} p\left(z|f,t\right)}{p\left(t|z\right)} \right\} - \lambda_t. \tag{2.24}$$

According to the normalization condition, the Lagrange multipliers are obtained as

$$\lambda_f = \sum_f \left\{ c_f \xi\left(f|z\right) + \sum_t \mathbf{X}_{f,t} p\left(z|f,t\right) \right\} \tag{2.25}$$

$$\lambda_z = \sum_z \left\{ c_z \psi\left(z\right) + \sum_f \sum_t \mathbf{X}_{f,t} p\left(z|f,t\right) \right\} \tag{2.26}$$

$$\lambda_t = \sum_t \left\{ c_t \zeta\left(t|z\right) + \sum_f \mathbf{X}_{f,t} p\left(z|f,t\right) \right\}. \tag{2.27}$$

Thus, the iterative update equations of the maximization step of PLCA are derived as

$$p(f|z) \leftarrow \frac{c_f \xi(f|z) + \sum_t \mathbf{X}_{f,t} p(z|f,t)}{\sum_f \left\{ c_f \xi(f|z) + \sum_t \mathbf{X}_{f,t} p(z|f,t) \right\}} \tag{2.28}$$

$$p(z) \leftarrow \frac{c_z \psi(z) + \sum_f \sum_t \mathbf{X}_{f,t} p(z|f,t)}{\sum_z \left\{ c_z \psi(z) + \sum_f \sum_t \mathbf{X}_{f,t} p(z|f,t) \right\}} \tag{2.29}$$

$$p(t|z) \leftarrow \frac{c_t \zeta(t|z) + \sum_f \mathbf{X}_{f,t} p(z|f,t)}{\sum_t \left\{ c_t \zeta(t|z) + \sum_f \mathbf{X}_{f,t} p(z|f,t) \right\}}. \tag{2.30}$$

Alteration of the expectation step and the maximization step makes the variables converge. In these equations, the prior information is reflected in the form of a weighted sum in the iterative update formula.

2.3.2 NMF framework

The term *Dirichlet prior* is used in the probabilistic analysis as it is a kind of probability distribution. Hence, it is impossible to apply it to the linear algebraic approach as NMF. However, when we focus on the duality between PLCA and NMF, it can be observed that it is possible to make the NMF function like Dirichlet prior-applied PLCA. In this subsection, the application of the Dirichlet prior is investigated in detail, and how it can be extendedly applied to the NMF is described. We define this extended utilization of the concept of Dirichlet prior as *generalized Dirichlet prior*

In the PLCA framework, Dirichlet prior is a kind of "prior distribution" that contains prior knowledge about the distributions to be estimated – $(f|z)$, (z),

$(t|z)$. As we can confirm in the Eq. 2.19, if any kind of prior distribution is not used, maximization of the posterior probability will become same as maximizing the likelihood. For this reason, the iterative update formula will remain same as the original PLCA update formula. Meanwhile, the use of Dirichlet prior changes the update formula of PLCA to be the weighted sum of the original formula and the hyperparameters. For example, Eq. 2.28 can be disassembled into the following three equations.

$$p\left(f|z\right) \leftarrow \sum_{t} \mathbf{X}_{f,t} p\left(z|f,t\right) \tag{2.31}$$

$$p\left(f|z\right) \leftarrow c_f \xi\left(f|z\right) + p\left(f|z\right) \tag{2.32}$$

$$p\left(f|z\right) \leftarrow \frac{p\left(f|z\right)}{\sum_{f'} p\left(f'|z\right)} \tag{2.33}$$

Here, only Eq. 2.32 is the newly added equation by imposing Dirichlet prior. As it has a form of weighted sum, we have directly applied it to the NMF's update equations. Note that Eq. 2.31 and 2.33 correspond to the NMF's spectral basis update equation presented in Eq. 2.18. To adopt the Dirichlet prior imposition as its post-processing, we assume that switching Eq. 2.32 and 2.33 does not interfere the variables' long-term convergence.

When the Dirichlet prior is extensively applied, the spectral basis update equation of the NMF is converted into the following equations as

$$\tilde{\mathbf{W}}_{f,k} \leftarrow \frac{\mathbf{W}_{f,k}^{(l)} \sum_{t} \left\{ \mathbf{H}_{k,t}^{(l+1)} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{t'} \mathbf{H}_{k,t'}^{(l+1)}} \tag{2.34}$$

$$\mathbf{W}_{f,k}^{(l+1)} \leftarrow w^{(freq)} \tilde{\mathbf{W}}_{f,k} + \left(1 - w^{(freq)}\right) \mathbf{\Xi}_{f,k} \tag{2.35}$$

where Ξ and $\tilde{\mathbf{W}}$ are the temporarily adopted variables each has the same size as $\mathbf{W}$, and $w^{(freq)}$ is the mixing weight. These two equations replace the Eq.

2.18. Similarly, the update equation of the time activation is changed as

$$\tilde{\mathbf{H}}_{k,t} \leftarrow \frac{\mathbf{H}^{(l)}_{k,t} \sum_{f} \left\{ \mathbf{W}^{(l)}_{f,k} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{f'} \mathbf{W}^{(l)}_{f',k}} \tag{2.36}$$

$$\mathbf{H}^{(l+1)}_{k,t} \leftarrow w^{(time)} \tilde{\mathbf{H}}_{k,t} + \left(1 - w^{(time)}\right) \mathbf{Z}_{k,t} \tag{2.37}$$

where $\tilde{\mathbf{H}}$ and $\mathbf{Z}$ are the temporarily adopted variables each has the same size as $\mathbf{H}$, and $w^{(time)}$ is the mixing weight. These two equations replace the Eq. 2.17. Note that these equations are equivalent to the original update equations when $\mathbf{\Xi} = \mathbf{W}$ and $\mathbf{Z} = \mathbf{H}$.

The proposed reformulation is partially heuristic, however, this can reduce the computational complexity dramatically. As PLCA requires massive computation, reducing it is one of the main focuses to have them used in the realistic environment. By using this NMF with the generalized Dirichlet prior, we can both get the similar results to the Dirichlet prior-applied PLCA.

2.4 Summary

In this chapter, we have given an overview of the theoretical backgrounds that are crucial to describe our following studies. To this end, we have reviewed the representative matrix decomposition algorithms – PLCA and NMF – and investigated their equivalence. Then the Dirichlet prior imposition is studied in the probabilistic framework followed by examining the transition of PLCA's iterative update equations. By the fact that the Dirichlet prior imposition is equivalent to adding the weighted summation formula with the exemplar hyperparameter in the post-processing step, we could generalize it to be applied

to the NMF framework. It is meaningful in that it significantly reduces the computational complexity.

The four main topics will be discussed in detail in the following chapters. In the first half of the chapters, we deal with the harmonic-percussive sound separation problem with two different approaches. In Chapter 3, the first approach is presented that focuses on the spectral aspects of the harmonic and percussive sounds. On the other hand, in Chapter 4, we focus on both spectral and temporal characteristics of the harmonic and percussive sounds. In the second half of the chapters, we examine the harmonic instrument sound separation problem. As it is a comparatively difficult problem, we first present informed approach in Chapter 5 that uses pre-trained envelope information. Then it is extended to the blind approach in Chapter 6. All of these studies are based on the basis shaping using the generalized Dirichlet prior.

Chapter 3

Harmonic-Percussive Source Separation Using Harmonicity and Sparsity Constraints

3.1 Introduction

In this chapter, based on the original work by Park et al. [51], we describe the first HPSS method that uses spectral characteristics of the harmonic and percussive sounds. Recently, musical signal processing has received a great deal of attention especially with the rapid growth of digital music sales. Automatic musical feature extraction and analysis for a large amount of digital music data has been enabled with the support of computational power. The major purposes of such tasks include extracting musical information such as melody extraction, chord estimation, onset detection, and tempo estimation.

Because most music signals often consist of both harmonic and percussive signals, the extraction of tonal attributes is often severely degraded by the pres-

ence of percussive interference. On the other hand, when we analyze rhythmic attributes such as tempo estimation, the harmonic signals act as interference that may prevent accurate analysis. Consequently, the separation of harmonic and percussive components in music signals will function as an important pre-processing step that allows efficient and precise analysis.

For these reasons, many researchers have focused on investigating HPSS using various approaches. Uhle *et al.* performed singular value decomposition (SVD) followed by independent component analysis (ICA) to separate drum sounds from the mixture [52]. Gillet *et al.* presented a drum-transcription algorithm based on band-wise decomposition using sub-band analysis [53].

Other researchers have employed matrix factorization techniques such as NMF. Helen *et al.* proposed a two-stage process composed of a matrix-factorization step and a basis-classification step [54]. Kim *et al.* employed the matrix co-factorization technique, where spectrograms of the mixture sound and drum-only sound are jointly decomposed [55]. NMF with smoothness and sparseness constraints was utilized by Canadas-Quesada *et al.* [48]. The algorithm was developed based on assumptions regarding the anisotropic characteristics of the harmonic and percussive components; harmonic components have temporal continuity and spectral sparsity, whereas percussive components have spectral continuity and temporal sparsity.

Most HPSS algorithms have employed the same assumption. Ono *et al.* presented a simple technique to represent a mixture sound spectrogram as a sum of harmonic and percussive spectrograms based on the Euclidean distance [56]. Their technique aims to minimize the temporal dynamics of harmonic components and the spectral dynamics of percussive components. They fur-

ther extended their work to use an alternative cost function based on the KL divergence [57]. More recently, FitzGerald presented a median filtering-based algorithm [58], where a median filter is applied to the spectrogram in a row-wise and column-wise manner for the extraction of harmonic and percussive sounds, respectively. Gkiokas *et al.* also proposed a non-linear filter-based HPSS algorithm [59].

However, the assumption regarding the temporal continuity, which is considered to be crucial for conventional harmonic-percussive studies, does not account for the rapidly time-varying harmonic signals often present in vocal sounds and musical expressions such as slides, vibratos, or glissandos. This is because their spectrograms often fluctuate over short periods of time. Thus, it may degrade the performance of the algorithms, particularly when loud vocal components or such musical expressions are mixed.

In this chapter, we propose a HPSS algorithm that is classified as a spectrogram decomposition-based method. We consider the spectrum of harmonic components to have a harmonic and sparse structure in the frequency domain, whereas the spectrum of percussive components to have an unsparse structure. To realize the successful separation of harmonic/percussive sounds, we apply constraints that impose a particular structure of the spectral bases. The novelty of the proposed method resides in the harmonicity constraint, which is an extension of the sparsity constraint presented in previous works [39]. The constraint is closely related to the Dirichlet prior, which is frequently used in probabilistic analysis. Because the proposed algorithm does not assume temporal continuity for the separation of harmonic signals, we can successfully separate harmonic signals from the mixture sound, even when there are significant fluctuations

over time.

The rest of this chapter is organized as follows. Section 3.2 explains in detail how the proposed method works. In Section 3.3, we present experimental results, and in Section 3.4, we draw our conclusion.

3.2 Proposed method

In this section, we present a detailed explanation of the proposed HPSS method. The proposed algorithm uses the spectrogram-decomposition technique, NMF, with the harmonicity and sparsity constraints based on the Dirichlet prior. For the efficient description of the proposed method, we first introduce the algorithm description for the proposed method. Then, the theoretical relations of the proposed method to the Dirichlet prior are described.

3.2.1 Formulation of Harmonic-Percussive Separation

We present a modified NMF algorithm to impose the characteristics of harmonic/percussive sounds. The update rule is separately represented for the harmonic source basis and percussive source basis as follows:

$$\mathbf{H}_{k,t}^{(l+1)} \leftarrow \frac{\mathbf{H}_{k,t}^{(l)} \sum_{f} \left\{ \mathbf{W}_{f,k}^{(l)} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{f'} \mathbf{W}_{f',k}^{(l)}} \tag{3.1}$$

$$\tilde{\mathbf{W}}_{f,k} \leftarrow \frac{\mathbf{W}_{f,k}^{(l)} \sum_{t} \left\{ \mathbf{H}_{k,t}^{(l+1)} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{t'} \mathbf{H}_{k,t'}^{(l+1)}} \tag{3.2}$$

$$\hat{\mathbf{w}}_k \leftarrow \left(1 - \gamma_H^H\right) \tilde{\mathbf{w}}_k + \gamma_H^H ifft\left(\left\{fft\left(\tilde{\mathbf{w}}_k\right)\right\}^p\right), k \in \Phi_H \tag{3.3}$$

$$\hat{\mathbf{w}}_k \leftarrow \max\left(\hat{\mathbf{w}}_k, 0\right), k \in \Phi_H \tag{3.4}$$

$$\begin{cases} \mathbf{w}_k^{(l+1)} \leftarrow \left(1-\gamma_S^H\right)\hat{\mathbf{w}}_k + \gamma_S^H(\hat{\mathbf{w}}_k)^q, k \in \Phi_H \\ \mathbf{w}_k^{(l+1)} \leftarrow \left(1-\gamma_S^P\right)\tilde{\mathbf{w}}_k + \gamma_S^P(\tilde{\mathbf{w}}_k)^r, k \in \Phi_P \end{cases} \tag{3.5}$$

where $\tilde{\mathbf{w}}_k$, $\hat{\mathbf{w}}_k$, and $\mathbf{w}_k^{(l+1)}$ denote the k-th column of $\tilde{\mathbf{W}}$, $\hat{\mathbf{W}}$, and $\mathbf{W}^{(l+1)}$, respectively, Φ_H and Φ_P denote a set of harmonic bases and percussive bases, respectively, $fft(\cdot)$ and $ifft(\cdot)$ denote the functions of the fast Fourier transform (FFT) and the inverse FFT (IFFT), respectively, $\mathbf{w}_k$ denotes the kth column of $\mathbf{W}$, γ_H^H denotes the harmonicity weight parameter for the harmonic signal, and γ_S^H and γ_S^P denote the sparsity weight parameters for harmonic and percussive signals, respectively. Note that Eq. 3.1 and 3.2 are identical to the original NMF update equations. Eq. 3.3–3.5 contribute to shaping the spectral bases as desired as the iteration proceeds.

Mixing weights that have values between 0 and 1 represent the importance of each constraint imposition, and indicate the degree to which we need to impose the characteristic. To enable the harmonic bases to have a harmonic and sparse structure while preserving the original figures of spectral bases, γ_H^H and γ_S^H are set to have small positive numbers, as the effect of the constraint is accumulated over the iteration.

The exponents p, q, and r have to be determined considering the range of each parameter, $0 \leq r \leq 1 \leq p, q$. Here, p and q respectively reflect the degree of harmonicity and sparsity of the destination, and they have to be controlled considering the spectral characteristics of the original harmonic sources. Likewise, r reflects the degree of "unsparsity" of the percussive sources.

Among the update equations shown above, the function of the conventional NMF update equations is to minimize the error between $\mathbf{X}$ and its estimation $\tilde{\mathbf{X}}$. On the other hand, the remainders of the equations aim to shape the spectral bases. The sparsity constraint in Eq. 3.5 has been similarly adopted for the matrix decomposition [39], and it is based on the fact that the square operation increases the differences among the vector components. If the square root operation is used instead, as in the percussive case of Eq. 3.5, unsparsity can be imposed to the basis. Similarly, we can extend this concept to the harmonicity. The second term in Eq. 3.5 denotes the harmonics-emphasized basis, which is due to the fact that the *spectrum of the spectrum* is sparse. To prevent elements from being negative, the $\max(\cdot,\cdot)$ operation in Eq. 3.4 has to be jointly involved.

The harmonic and percussive sounds are reconstructed using the corresponding bases as follows:

$$\mathbf{X}^{(Harmonic)} = \sum_{k \in \Phi_H} \mathbf{w}_k \mathbf{h}_k \tag{3.6}$$

$$\mathbf{X}^{(Percussive)} = \sum_{k \in \Phi_P} \mathbf{w}_k \mathbf{h}_k \tag{3.7}$$

where $\mathbf{h}_k$ denotes the kth row of $\mathbf{H}$.

3.2.2 Relation to Dirichlet Prior

The proposed update equations can be intuitively comprehended. However, the equations are based on a firm theoretical background presented in section 2, not heuristically induced. In this subsection, we investigate the relations between the generalized Dirichlet prior and the proposed method.

As shown in the previous chapter, we can generalize the Dirichlet prior of the PLCA by applying it to the NMF algorithm as follows:

$$\mathbf{H}_{k,t} \leftarrow (1-\gamma_1)\frac{\mathbf{H}_{k,t}\sum_f \left\{\mathbf{W}_{f,k}\mathbf{X}_{f,t}\Big/\tilde{\mathbf{X}}_{f,t}\right\}}{\sum_{f'}\mathbf{W}_{f',k}} + \gamma_1\mathbf{A}_{k,t} \tag{3.8}$$

$$\mathbf{W}_{f,k} \leftarrow (1-\gamma_2)\frac{\mathbf{W}_{f,k}\sum_t \left\{\mathbf{H}_{k,t}\mathbf{X}_{f,t}\Big/\tilde{\mathbf{X}}_{f,t}\right\}}{\sum_{t'}\mathbf{H}_{k,t'}} + \gamma_2\mathbf{B}_{f,k} \tag{3.9}$$

where $\mathbf{A}$ and $\mathbf{B}$ denote the matrices of hyper parameters with respect to $\mathbf{H}$ and $\mathbf{W}$, respectively, and γ_1 and γ_2 denote the mixing weights. In our research, we focus only on the spectral bases, and thus Eq. 3.8 is discarded. As can be observed, the proposed update equations, Eq. 3.1–3.5, have the same form as Eq. 3.9, and the way in which we shape the spectral bases depends on the form of $\mathbf{B}$ matrix.

Frequency-domain sparsity imposition can be easily achieved by setting the hyper parameter $\mathbf{B}$ as [39]

$$\mathbf{b}_k = (\mathbf{w}_k)^u \tag{3.10}$$

where $\mathbf{b}_k$ denotes the kth column of $\mathbf{B}$, and u denotes an exponent that controls the degree of sparsity of $\mathbf{b}_k$.

On the other hand, harmonicity imposition can be achieved when the hyper parameter is represented as

$$\mathbf{b}_k = ifft\left(\{fft\left(\mathbf{w}_k\right)\}^v\right) \tag{3.11}$$

where v denotes the exponent that controls the degree of harmonicity of $\mathbf{b}_k$. This is because a periodic signal can be represented as a sum of sinusoids, and the spectrum of the periodic signal is sparse. Conversely, if a spectrum is sparse, we

can assume that the original signal has a strongly periodic characteristic. Thus, we aim to make the *spectrum of the spectrum* to be sparse in order to shape a signal such that it has a harmonic structure. Note that in order to prevent destructive interference caused by phase distortion, we have to manipulate only the magnitudes within the IFFT function, preserving the original phases of $fft(\mathbf{w}_k)$.

3.3 Performance evaluation

3.3.1 Sample Problem

In this section, we apply the proposed method and the conventional methods to simple sample examples, which is suitable for showing the novelty and validity of the proposed method. Spectrograms of synthesized sounds that consist of horizontal and vertical lines are presented in Fig. 3.1 (a) and Fig. 3.2 (a). Fig. 3.1 (a) models the case where a pitched harmonic sound is sustained for a certain period. The sounds of harmonic instruments such as guitars, pianos, flutes, and violins fall within this scenario. On the other hand, Fig. 3.2 (a) illustrates the case where a harmonic signal alters its frequency over time. In this case, vibratos, glissandos, and vocal signals correspond to the harmonic components. We compare the performance of the proposed method to the separation results obtained using three conventional methods: Ono *et al.*'s Euclidean distance-based method [56], Ono *et al.*'s KL divergence-based method [57], and FitzGerald's method [58].

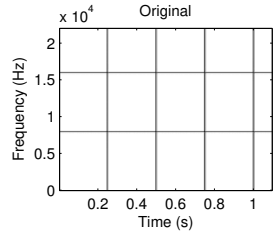
As shown in Fig. 3.1 (b), both the conventional methods and the proposed method are able to successfully separate the sounds. This is because the hori-

zontal lines in this example have horizontally continuous characteristics, which are assumed by the conventional methods to be present. However, when the harmonic sound vibrates and the horizontal lines fluctuate, as shown in Fig. 3.2 (a), conventional methods cannot distinguish the horizontal lines from vertical lines. As we can see in Fig. 3.2 (b), the estimated percussive components of conventional methods contain harmonic partials, and only the proposed method can successfully separate them. Thus, we can claim that the proposed method is not affected by variations in the pitch because it relies on the harmonic structure of the vertical axis, and not the degree of horizontal transition.

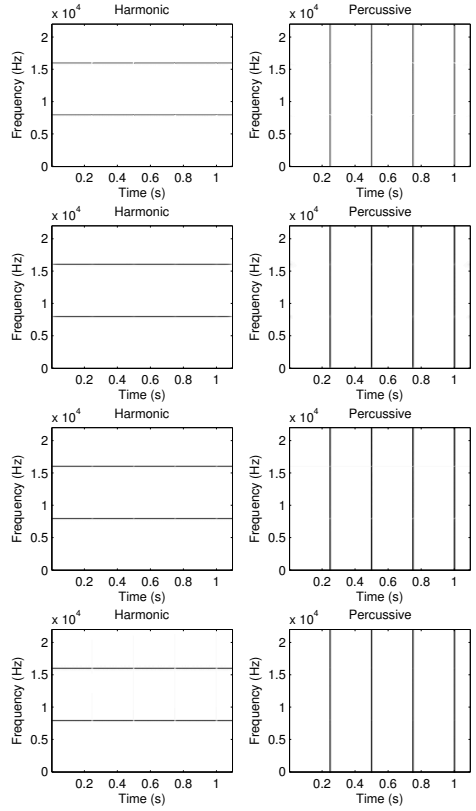
3.3.2 Qualitative Analysis

We evaluated the performance of the proposed method using a real recording example. Fig. 3.3 shows a log-scale plot of the spectrogram of an excerpt from "*Billie Jean*," by *Michael Jackson*. The signal was sampled at 22,050 Hz, and the frame size and overlap size were set to 1,024 and 512, respectively. We can observe from the spectrogram that the excerpt contains both harmonic and percussive components. The harmonic components can be seen as horizontally connected lines, whereas the percussive components are seen as vertical lines as in the sample examples.

Fig. 3.4 (a) and (b) show the separation results of the harmonic sound (up) and percussive sound (down), which were obtained using Ono *et al.*'s Euclidean distance-based method and KL divergence-based method, respectively. Here, we set the parameters to the values recommended in the references. We observe that the estimated percussive components still contain harmonic components that may correspond to the vocal components. This is because Ono *et al.*'s algo-

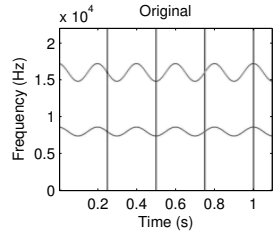


(a) Original figure

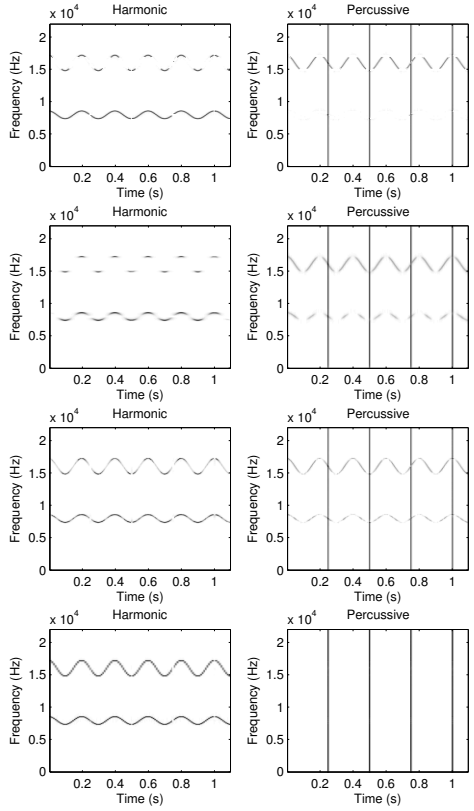


(b) Separation results of Ono's Euclidean distance-based method, Ono's KL divergence-based method, FitzGerald's method, and the proposed method (from top to bottom)

Fig. 3.1 Sample example of separating straight horizontal lines and vertical lines.



(a) Original figure



(b) Separation results of Ono's Euclidean distance-based method, Ono's KL divergence-based method, FitzGerald's method, and the proposed method (from top to bottom)

Fig. 3.2 Sample example of separating fluctuating horizontal lines and vertical lines.

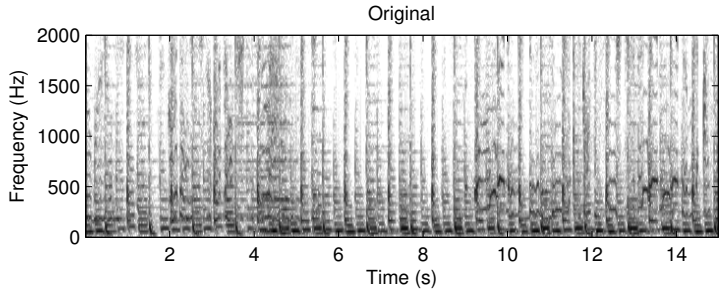
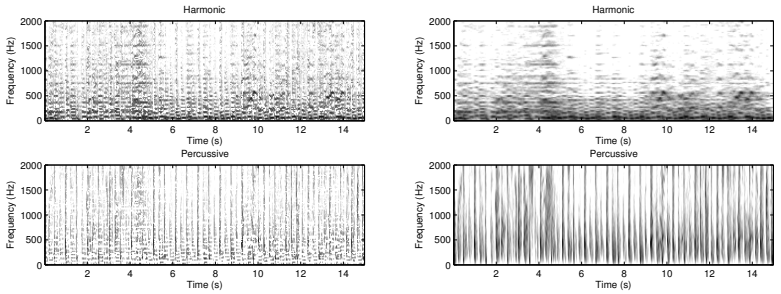


Fig. 3.3 Spectrogram of a real audio recording example ("Billie Jean" by Michael Jackson).

rithms aim to minimize the temporal transition of the harmonic spectrogram. However, vocal components in the original spectrogram do not match well with the underlying assumption.

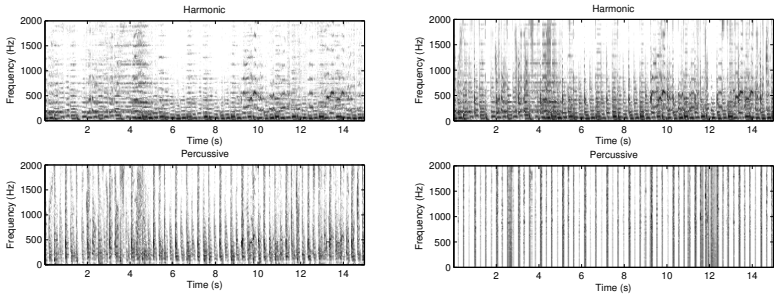
Fig. 3.4 (c) shows the result of FitzGerald's method with a median filter length of 17 and when the exponent for the Wiener filter-based soft mask is two, as recommended by FitzGerald [58]. We also observe that the separated percussive components still contain harmonic components, as in the previous case. This is because of the use of a one-dimensional median filter, which assumes that the harmonic components are sustained for several periods.

Fig. 3.4 (d) shows the performance of the proposed method. We observe that the harmonic and percussive components are clearly separated, and the percussive components do not have any vocal components in these results. This is because unlike conventional methods, the proposed algorithm does not rely on the horizontal continuity principle. Rather, the proposed algorithm tries to account for the harmonic components using the harmonic and sparse spectral bases.



(a) Ono's Euclidean distance-based method

(b) Ono's KL divergence-based method



(c) FitzGerald's method

(d) Proposed method

Fig. 3.4 Qualitative performance comparison of conventional and proposed methods.

3.3.3 Quantitative Analysis

We performed a quantitative analysis to verify the validity of the proposed algorithm. First, we compiled a dataset that consists of 10 audio samples, which is a subset of the MASS database [60], but two sets of data, namely *tamy-que_pena_tanto_faz_6-19* and *tamy-que_pena_tanto_faz_46-57*, were excluded in this experiment because they lack percussive signals. Then, we obtained a spectrogram for each audio sample with the frame size and hop size set to 2,048 samples and 1,024 samples, respectively. Note that the sampling rate of the songs in

the MASS dataset is 44,100 Hz. Finally, we measured the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) using the BSS_EVAL toolbox (http://bass-db.gforce.inria.fr /bss_eval/) supported by [61]. They are mathematically defined as

$$SDR = 20\log_{10}\left(\frac{||s_{target}||}{||s_{interf} + s_{artif}||}\right) \tag{3.12}$$

$$SIR = 20\log_{10}\left(\frac{||s_{target}||}{||s_{interf}||}\right) \tag{3.13}$$

$$SAR = 20\log_{10}\left(\frac{||s_{target} + s_{interf}||}{||s_{artif}||}\right) \tag{3.14}$$

where s_{target}, s_{interf}, and s_{artif} denote the target sound, interference, and artifact, respectively. SIR and SAR have a performance trade-off relationship with each other; thus, we consider SDR as the representative performance value. Table 1 shows the parameter values of the proposed method used in this experiment. The parameters of the conventional methods are set to the recommended values, as in the previous experiment.

The evaluation results are summarized in Fig. 3.5. We can see that the proposed method guarantees a better average SDR result compared to conventional methods, even though the proposed method has a lower SIR performance than Ono *et al.*'s Euclidean distance-based method. This is because the proposed method far outperforms other methods with respect to the SAR, which has a trade-off relation with the SIR [62].

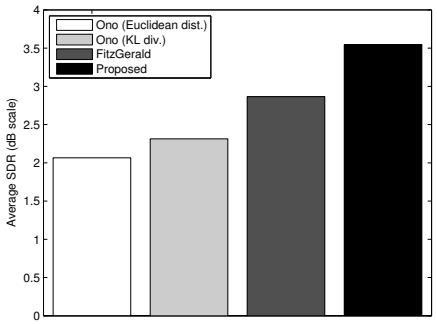
3.4 Summary

In this chapter, we proposed a novel HPSS algorithm based on NMF with harmonicity and sparsity constraints. Conventional methods assumed that the

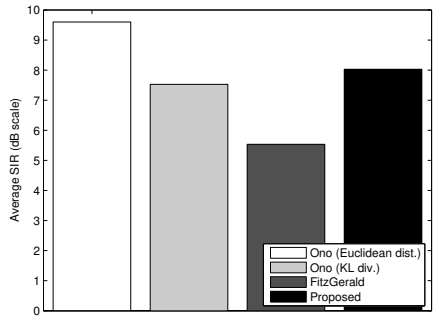
Parameter	Value
p	1.1
q	1.1
r	0.5
γ_H^H	0.001
γ_S^H	0.001
γ_S^P	0.1
Number of bases (H,P)	(300,200)

Table 3.1 Experimental parameters.

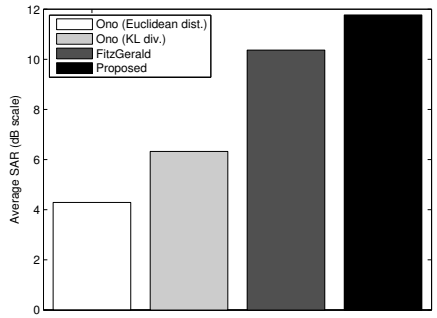
harmonic components were represented as horizontal lines with temporal continuity. However, such an assumption could not be applied to the vocal components or various musical expressions of harmonic instruments. To overcome this problem, we presented a harmonicity constraint, which is a generalized Dirichlet prior. By letting the spectrum of the spectrum be harmonic and sparse, we could refine the harmonic components and eliminate inharmonic components. The experimental results showed the validity of the proposed method by comparing it with conventional methods.



(a) SDR



(b) SIR



(c) SAR

Fig. 3.5 Quantitative performance comparison of conventional and proposed methods.

Chapter 4

Exploiting Continuity/Discontinuity of Basis Vectors in Spectrogram Decomposition for Harmonic-Percussive Sound Separation

4.1 Introduction

In this chapter, the former HPSS algorithm is extended to employ temporal characteristics. To this end, spectral/temporal continuity is controlled by means of Dirichlet prior. This chapter is based on the research published in the IEEE/ACM Transactions on Audio, Speech, and Language Processing [63].

Recently, the digital music sales market has grown rapidly because of the use of smart devices and high-speed wireless internet connectivity. Consequently, it has become an important task to automatically extract musical information for

massive databases. This information includes tonal attributes such as pitches, chords, and keys and rhythm-related features such as beat and tempo.

In the general case, musical signals are composed of harmonic sounds produced by pitched instruments, and percussive sounds such as those produced by drums. The difference between harmonic and percussive sounds may cause severe performance degradation of the musical information extraction algorithms, because percussive signals act as interference when analyzing tonal features, and vice versa. For this reason, HPSS functions as an essential pre-processing step in a number of music information retrieval (MIR) tasks, such as tempo estimation [64], music structural segmentation [65], chord estimation [66], vocal separation [67], and melody line estimation [68].

Other research efforts related to digital audio effects have also used HPSS algorithms. Tachibana *et al.* applied the HPSS algorithm to singing voice enhancement [69]. Driedger *et al.* used HPSS for time-scale modification to reduce artifacts [70]. Buyens *et al.* also used HPSS for music pre-processing for cochlear implant users [71].

In the early stage of HPSS research, drum sound transcription and separation was studied. Uhle *et al.* applied an independent component analysis for the drum track extraction [52]. Helen and Virtanen used NMF followed by a SVM [54]. Gillet and Richard presented subband analysis and a Wiener-filtering-based drum separation and transcription method [72].

Recent studies interpret the magnitude spectrogram as a non-negative matrix and often utilize matrix decomposition algorithms. Kim *et al.* proposed an NMF-based partial co-factorization algorithm for drum separation [55]. To fully utilize this algorithm, drum-only sounds must be jointly provided. Other studies

focus on the characteristics of the harmonic and percussive components present in the spectrogram. Ono *et al.* presented a matrix division method based on row-wise and column-wise continuity [56]; they extended their work to the case in which Kullback-Leibler divergence was used as a cost function [57]. They also presented a real-time equalizer for harmonic and percussive sounds in the paper. Ono *et al.*'s algorithms were further analyzed and evaluated in a review paper [73].

Fitzgerald applied a one-dimensional median filter to the magnitude spectrogram to exploit the anisotropy of the harmonic and percussive components [58]; a post-processing method to refine the separated sounds was presented by Thoshkahna and Ramakrishnan [74] and Driedger *et al.* [75]. Fitzgerald *et al.* also applied kernel additive modeling (KAM) to the HPSS problem, which can be interpreted as a generalization of the median filtering [76]. Gkiokas and Papavassiliou used non-linear image filtering that includes morphological operation [59]. Canadas-Quesada *et al.* re-derived the update equations of the NMF algorithm by inserting sparseness and smoothness constraints into the cost function [48]. Park and Lee focused on the spectral aspects of harmonic and percussive components targeting the successful separation of the vocal components that may lack the feature of smooth time activation [51]. Duong *et al.*'s method focused on the multichannel HPSS environment [77].

Most of the conventional research efforts considered the anisotropic characteristics of the harmonic and percussive components presented in the spectrogram. Especially, Ono *et al.*'s methods, Fitzgerald's median filter-based method, and KAM-based method rely on the assumption that the harmonic components are continuous in the temporal domain, whereas the percussive components are

continuous in the spectral domain. However, these methods are not able to separate vocal components, as the bases of the vocal components do not remain for a sufficient time due to its rapidly time-varying nature like vibrato or slur.

Park and Lee's method focused on the spectral aspects of the harmonic and percussive components, attempting to clearly separate the vocal components. They assumed that only a few harmonically distributed frequency bins of the harmonic components possess a large portion of the frame energy, where the spectrum of the percussive components was flat and non-sparse. However, their assumption cannot be applied to the case in which the kick drum is mixed, since most of the energy of the kick drum's spectrum is concentrated in the low frequency band.

Canadas-Quesada *et al.*'s method presents a different approach to vocal separation: it fully utilizes the sparseness and smoothness characteristics for both harmonic and percussive components, which had been only partially adopted in previous studies. Their method aims to shape the bases such that they have sparse or smooth structures that satisfy the anisotropic characteristics of the harmonic and percussive components. Here, the smoothness was measured using the sum of squared differences with the adjacent components, whereas the sparseness was measured using L1-norm.

Canadas-Quesada *et al.*'s approach partially solves the problem of vocal component separation, because the sparseness constraint applied to the harmonic spectrum might compensate for the mismatch of on temporal side. However, whether it is accurate to use the sparsity measures to distinguish the harmonic and percussive components is unclear. Harmonic sounds are naturally "harmonic", which means they contain energy not only at the fundamental fre-

quency but also at integer multiples of the fundamental. Clearly, the sparseness would be increased if the energy spread through the harmonic frequency bins were concentrated in a single bin. However, both the pure sinusoidal wave and the kick drum signal have this spectral structure, which may cause difficulty in distinguishing harmonic and percussive spectra when using only the sparseness measure.

In this chapter, we exploit continuity control to separately estimate the harmonic and percussive bases when performing the matrix decomposition. This method is based on our observation that not the sparsity but rather the continuity is a representative indicator to better differentiate the harmonic and percussive spectra. The proposed algorithm can be derived using PLCA enforced with Dirichlet prior. The reason is that in probabilistic analysis framework, the imposition of Dirichlet prior is a proper and convenient method for shaping spectral/temporal bases of the matrix decomposition techniques [25]. By giving the prior information about the bases' shapes, we can roughly determine the bases where to converge. This is due to the fact that the probabilistic analysis-based matrix decomposition techniques such as PLCA interpret the magnitude spectrogram as a histogram of the multinomial distribution, and that Dirichlet prior is a conjugate prior of the multinomial distribution. In so doing, we design the algorithm to control the basis convergence point such that the continuity is minimized or maximized. However, the PLCA algorithm requires far more computations than the NMF, which causes slower iteration and convergence. To solve this problem, we reformulate the PLCA with Dirichlet prior in a NMF framework. Because the update formulas use the weighted sum with the hyperparameters, they can be easily extended to the NMF algorithm which is

mathematically identical.

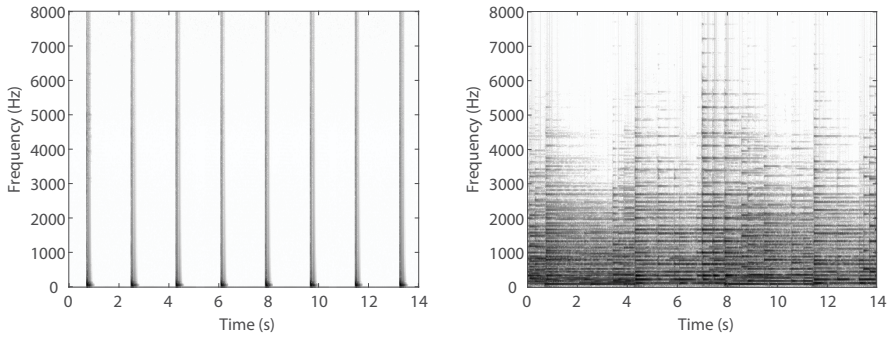
The remainder of this chapter is organized as follows. Section 2 gives a detailed description of the proposed algorithm. In Section 3, performance evaluation results are presented. Conclusions follow in Section 4 with directions for future work.

4.2 Proposed Method

4.2.1 Characteristics of harmonic and percussive components

Fig. 4.1 (a) and (b) show the spectrograms of the kick drum sound and the piano sound, respectively. Both sound sources are from "bearlin-roads_85-99" in the music audio signal separation (MASS) database [60]. We can confirm the anisotropic characteristics of the harmonic and percussive components from the figures: the kick drum's spectrogram is seen as a group of vertical lines and the piano's spectrogram is seen as a group of horizontal lines. The conventional assumptions about the harmonic and percussive components seem to be confirmed.

However, closer observation reveals that most of the kick drum energy is concentrated in the low frequency band. This property may cause the sparsity of the kick drum's spectrum to be higher than that of the piano spectrum. The more obvious measure is the energy concentration, because the activated frequency bins are harmonically distributed in the piano's spectrum. As the dominant peaks concentrate in a narrow region, the average difference between the adjacent values (or the degree of fluctuation) in the spectrum decreases as kick drum's spectrum.



(a) (b)

Fig. 4.1 Spectrograms of representative (a) percussive sound (kick drum) and (b) harmonic sound (piano).

To find out relevant measures that appropriately distinguish the harmonic and percussive spectra, we have inspected the spectral bases of a few sound samples: a kick drum, snare drum, hi-hat, piano, violin and pure tone sound samples. Before the analysis, a single basis was trained using the NMF algorithm. The trained bases are shown in Fig. 4.2. For the piano, violin, and pure tone samples, the E4 note was used. As shown in Fig. 4.2 (a), (b), and (c), the kick drum is an extreme case of percussive instruments, because most of the other percussive instruments usually have a flatter spectrum.

The considered measures are the L1-norm normalized with the L2-norm, the Gini index, entropy, and the degree of fluctuation (DF) normalized with the L1-norm. Note that the L1-norm measure is used in Canadas-Quesada *et al.*'s method, by modifying the cost function of the NMF algorithm by adding the L1-norm term. The Gini index originated from economics, but it also acts as a representative sparsity measure, according to Hurley and Rickard [78]. In the paper, the inverse term of the L1-norm measure was also investigated.

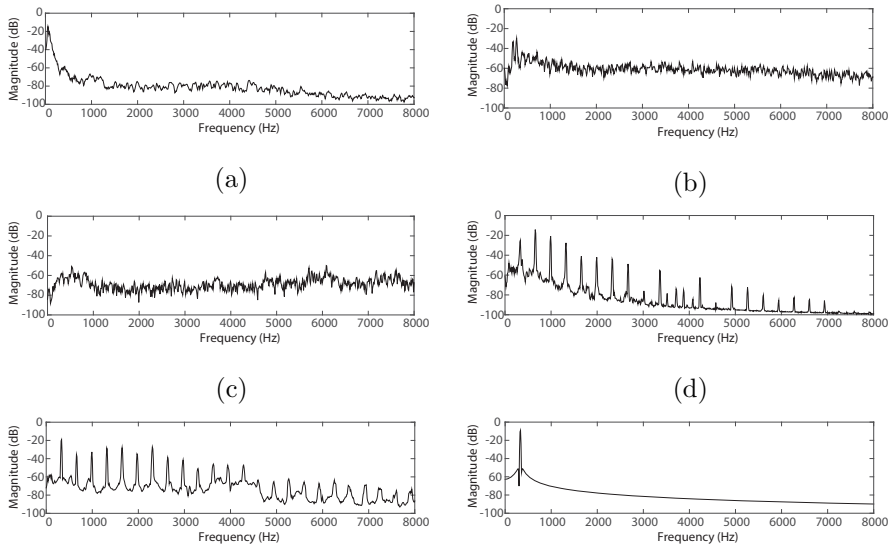


Fig. 4.2 Spectral bases trained from (a) a kick drum, (b) a snare drum, (c) a hi-hat, (d) a piano, (e) a violin and (f) a pure tone sounds.

Exploitation of the entropic prior in the matrix decomposition algorithm was presented by Smaragdis *et al.* [79].

The measures are mathematically defined for a spectral vector $\mathbf{w}$ as

$$\sigma_{L_1-norm} = \frac{\sum_{m=1}^{M} \mathbf{w}_m}{\sqrt{\sum_{m=1}^{M} \mathbf{w}_m^2}}, \tag{4.1}$$

$$\sigma_{Gini} = \frac{M+1}{M} - \frac{2 \sum_{m=1}^{M} (M+1-m)\, \mathbf{w}_m^{(sorted)}}{M \sum_{m=1}^{M} \mathbf{w}_m^{(sorted)}}, \tag{4.2}$$

$$\sigma_{Entropy} = -\sum_{m=1}^{M} \left(\frac{\mathbf{w}_m}{\sum_{m'=1}^{M} \mathbf{w}_{m'}} \log \frac{\mathbf{w}_m}{\sum_{m'=1}^{M} \mathbf{w}_{m'}} \right), \tag{4.3}$$

$$\sigma_{DF} = \frac{\sum_{m=1}^{M-1} |\mathbf{w}_m - \mathbf{w}_{m+1}|}{\sum_{m=1}^{M} \mathbf{w}_m}, \tag{4.4}$$

where $\mathbf{w}^{(sorted)}$ denotes $\mathbf{w}$ sorted in ascending order. Note that the small values of σ_{L_1-norm} and $\sigma_{Entropy}$ imply sparsity, whereas the large value of σ_{Gini} indicates sparsity. A large value of σ_{DF} implies discontinuity between the adjacent vector elements, which the harmonic bases are thought to have.

Fig. 4.3 shows the features computed from the spectral bases of a kick drum, snare drum, hi-hat, piano, violin, and pure tone signal. Dotted lines in the figures are virtual thresholds to distinguish the harmonic and percussive sounds. It can be observed that all the measures – the L1-norm, Gini index, entropy, and DF values – successfully distinguish the piano, violin, and pure tone sounds from the snare drum and hi-hat sounds, confirming our sparsity[continuity] assumption for separating the harmonic and percussive sounds. When comparing a kick drum sound to a piano and a violin sounds, however, all the measures but a proposed DF measure fail to discriminate the sounds.

The results presented in Fig. 4.3 can be easily understood via simplified illustrations presented in Fig. 4.4. Fig. 4.4 (a), (b), (c), and (d) correspond to the normalized spectra of the kick drum, hi-hat, harmonic instruments (piano and violin), and pure tone sounds, respectively. Table 4.1 shows the corresponding values of the illustrations measured using the same sparsity[continuity] measures used in Fig. 4.3. It is interesting that all of the three sparsity measures

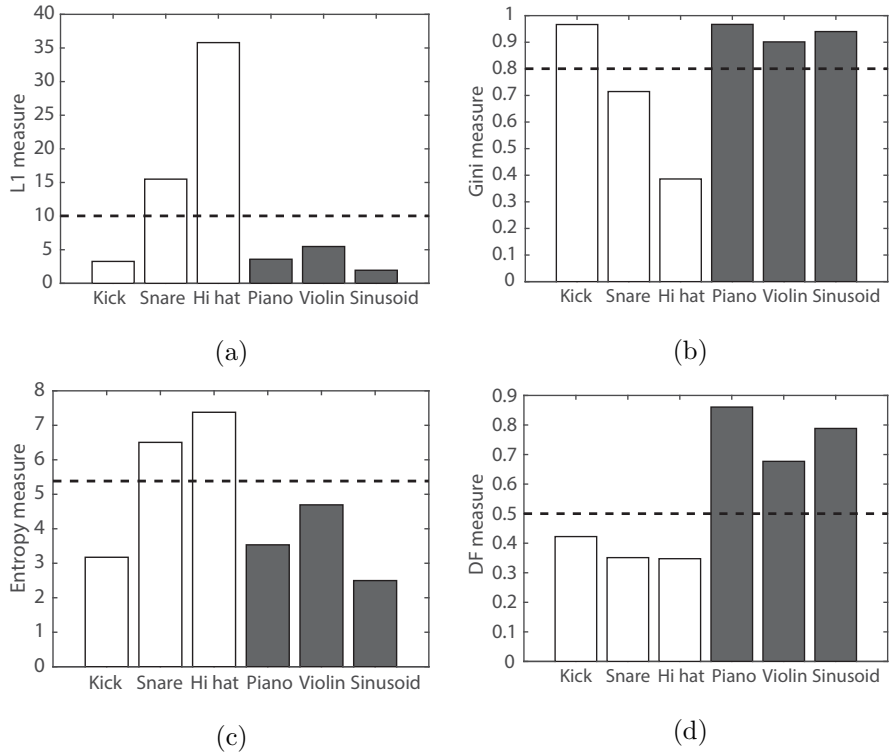


Fig. 4.3 Spectral measures of harmonic and percussive sounds.

showed the same values for illustrations (a) and (c). This is due to the fact that the sparsity measures consider the energy distribution inside the vector, which corresponds here to the number of peaks, regardless of their positions. However, as the DF measure can consider the relative positions of the activated frequency bins – whether they are consecutive or separated – it can contribute to distinguishing illustrations (a) and (c), as shown in Table 4.1. In addition, the DF value of (c) remains the same as (d), which is the maximum DF value. From the results, it can be predicted that the DF value would not be affected by the pitch of harmonic instruments, if the harmonic peaks are sufficiently spaced.

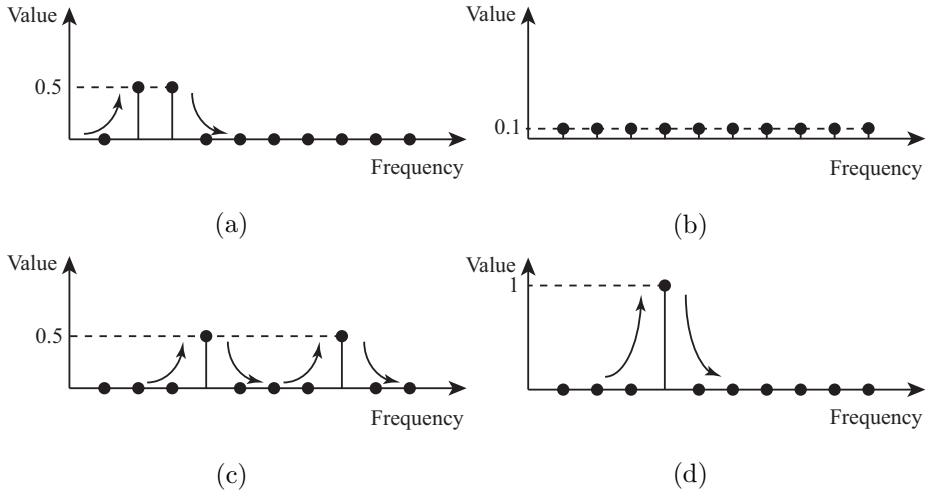


Fig. 4.4 Illustrations of harmonic and percussive spectra.

According to these observations, the sparsity measures might cause inaccurate separation of the harmonic and percussive bases when they are used as costs in a spectrogram decomposition algorithm; only the DF measure can correctly discriminate these bases. The similar observation can be found in the case of temporal bases, the only difference being the lower DF measure for the harmonic components. On the basis of this experiment, the proposed method is derived so as to increase or suppress the DF measure for harmonic-percussive sound separation.

4.2.2 Derivation of the proposed method

In this subsection, we present the proposed method using the theories presented in chapter 2. We utilize matrix decomposition techniques that have been widely used in audio source separation studies: PLCA and NMF. These techniques learn spectral bases and their corresponding temporal activations, which

Table 4.1 Spectral measures of illustrations.

	Percussive		Harmonic	
	(a)	(b)	(c)	(d)
L1-norm measure	1.41	3.16	1.41	1
Gini index	0.8	0	0.8	0.9
Entropy	0.69	2.30	0.69	0
Degree of fluctuation	1	0	2	2

enables one to observe the features of each side. Moreover, we can separately reconstruct the estimated source components using basis selection. The PLCA algorithm is used with the concept of Dirichlet prior imposition first, and then it is extended to the NMF framework. We assume that the harmonic components are *discontinuous* in the spectral domain and *continuous* in the temporal domain. We also assume that the percussive components are continuous in the spectral domain and discontinuous in the temporal domain.

The iterative update equations of PLCA algorithm have been presented in chapter 2. When the Dirichlet prior is imposed, it requires three hyperparameters – $\xi(f|z)$, $\psi(z)$, and $\zeta(t|z)$ – and three corresponding coefficients – c_f, c_z, and c_t – to be determined. The prior information is reflected in the form of a weighted sum in the iterative update formula. The proposed method is a

special case, in which the hyperparameters are determined as

$$\begin{cases} \xi\left(f|z\right) = p\left(f-1|z\right) \\ \psi\left(z\right) = 0 \\ \zeta\left(t|z\right) = p\left(t-1|z\right) \end{cases} . \tag{4.5}$$

The constant coefficients are generally defined to be non-negative. However, we can extend this concept to the point where negative coefficients are used in order to make it possible to impose discontinuity. In such cases, we have to prevent the probability values from being negative by adding an additional formula as

$$p\left(x\right) \leftarrow \max\left(p\left(x\right), \epsilon\right) \tag{4.6}$$

where $p\left(x\right)$ can be $p\left(f|z\right)$, $p\left(z\right)$, or $p\left(t|z\right)$, and ϵ is a small positive number.

The iterative update equations of the proposed method is represented as

$$p\left(z|f,t\right) \leftarrow \frac{p\left(z\right)p\left(f|z\right)p\left(t|z\right)}{\sum\limits_{z} p\left(z\right)p\left(f|z\right)p\left(t|z\right)} \tag{4.7}$$

$$p\left(f|z\right) \leftarrow \frac{c_f p\left(f-1|z\right) + \sum\limits_{t} \mathbf{X}_{f,t} p\left(z|f,t\right)}{\sum\limits_{f}\left\{c_f p\left(f-1|z\right) + \sum\limits_{t} \mathbf{X}_{f,t} p\left(z|f,t\right)\right\}} \tag{4.8}$$

$$p\left(z\right) \leftarrow \frac{\sum\limits_{f}\sum\limits_{t} \mathbf{X}_{f,t} p\left(z|f,t\right)}{\sum\limits_{z}\left\{\sum\limits_{f}\sum\limits_{t} \mathbf{X}_{f,t} p\left(z|f,t\right)\right\}} \tag{4.9}$$

$$p\left(t|z\right) \leftarrow \frac{c_t p\left(t-1|z\right) + \sum\limits_{f} \mathbf{X}_{f,t} p\left(z|f,t\right)}{\sum\limits_{t}\left\{c_t p\left(t-1|z\right) + \sum\limits_{f} \mathbf{X}_{f,t} p\left(z|f,t\right)\right\}} . \tag{4.10}$$

where Eq. 4.7 and 4.9 are identical to the original PLCA's update equations as we do not impose any constraints on $p\left(z|f,t\right)$ and $p\left(z\right)$. Random variables f, t,

and z have the interval of $1 \leq f \leq F$, $1 \leq t \leq T$, and $1 \leq z \leq K$, respectively. Accordingly, it is unable to define $p\left(f-1|z\right)$ and $p\left(t-1|z\right)$ in case of $f=1$ or $t=1$. In order to simplify the notation, we define $p\left(f-1|z\right)=0$ when $f=1$, and $p\left(t-1|z\right)=0$ when $t=1$.

Next, we extend the PLCA-based update formula to the NMF framework. As we have proven in the section 2, assigning basis-related information to the matrix decomposition algorithm can be induced in a probabilistic framework such as PLCA that has a thorough theoretical background. However, the PLCA algorithm requires four marginal distributions for estimation, which is a significant amount of computation, especially compared to the NMF algorithm, which requires the estimation of only two matrices. For this reason, the proposed method is extended to work in the NMF framework.

NMF has duality with the PLCA algorithm and is even equivalent to it when the KL divergence is used as the cost function [47]. According to the PLCA-based derivation, imposing the Dirichlet prior modifies the update formulas to the form of a weighted sum with hyperparameters. Because of its simplicity, the Dirichlet prior can easily be extended to the NMF. We have added the weighted sum formulas after the standard NMF's update equation as follows:

$$\tilde{\mathbf{H}}_{k,t} \leftarrow \frac{\mathbf{H}_{k,t}^{(l)} \sum\limits_{f} \left\{ \mathbf{W}_{f,k}^{(l)} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum\limits_{f'} \mathbf{W}_{f',k}^{(l)}} \tag{4.11}$$

$$\hat{\mathbf{H}}_{k,t}^{(Harm)} \leftarrow \alpha \tilde{\mathbf{H}}_{k,t}^{(Harm)} + \left(1-\alpha\right) \tilde{\mathbf{H}}_{k,t-1}^{(Harm)} \tag{4.12}$$

$$\hat{\mathbf{H}}_{k,t}^{(Perc)} \leftarrow \beta \tilde{\mathbf{H}}_{k,t}^{(Perc)} + \left(1-\beta\right) \tilde{\mathbf{H}}_{k,t-1}^{(Perc)} \tag{4.13}$$

$$\mathbf{H}^{(l+1)} \leftarrow max\left(\hat{\mathbf{H}}, \epsilon\right) \tag{4.14}$$

$$\tilde{\mathbf{W}}_{f,k} \leftarrow \frac{\mathbf{W}^{(l)}_{f,k} \sum_{t} \left\{ \mathbf{H}^{(l+1)}_{k,t} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{t'} \mathbf{H}^{(l+1)}_{k,t'}} \tag{4.15}$$

$$\hat{\mathbf{W}}^{(Harm)}_{f,k} \leftarrow \gamma \tilde{\mathbf{W}}^{(Harm)}_{f,k} + (1-\gamma)\, \tilde{\mathbf{W}}^{(Harm)}_{f-1,k} \tag{4.16}$$

$$\hat{\mathbf{W}}^{(Perc)}_{f,k} \leftarrow \delta \tilde{\mathbf{W}}^{(Perc)}_{f,k} + (1-\delta)\, \tilde{\mathbf{W}}^{(Perc)}_{f-1,k} \tag{4.17}$$

$$\mathbf{W}^{(l+1)} \leftarrow max\left(\hat{\mathbf{W}}, \epsilon\right) \tag{4.18}$$

where α, β, γ, and δ denote weightings, $\widetilde{\mathbf{X}} \left(\triangleq \mathbf{WH} \right)$ denotes the estimated spectrogram, $\tilde{\mathbf{H}}$ and $\hat{\mathbf{H}}$ denote temporarily adopted variables each of which has the same size as $\mathbf{H}$, and $\tilde{\mathbf{W}}$ and $\hat{\mathbf{W}}$ denote temporarily adopted variables each of which has the same size as $\tilde{\mathbf{W}}$. As the variables are not defined when their subscript is 0, we newly define $\tilde{\mathbf{H}}_{\mathbf{k},\mathbf{0}} \triangleq \tilde{\mathbf{H}}_{\mathbf{k},\mathbf{1}}$, and $\tilde{\mathbf{W}}_{\mathbf{0},\mathbf{k}} \triangleq \tilde{\mathbf{W}}_{\mathbf{1},\mathbf{k}}$. By defining so, we can maintain a consistent notation. Note that all matrix elements are calculated before it proceeds to the next equation, and Eq. 4.11 and 4.15 are the update algorithms of the original NMF.

Prior to the iteration, $\mathbf{W}$ and $\mathbf{H}$ are often initialized to randomized values. Here, we consider the bases to be separated into two disjoint groups, i.e., harmonic and percussive groups; the superscripts $(Harm)$ and $(Perc)$ indicates which group the bases belong to. The spectral bases of percussive components can also be initialized to *flat* vectors, of which the components are all equal. It will lead to faster convergence compared to random initialization.

After iteration, the harmonic and percussive spectrograms are reconstructed as

$$\mathbf{X}^{(Harm)} = \mathbf{W}^{(Harm)} \mathbf{H}^{(Harm)} \tag{4.19}$$

$$\mathbf{X}^{(Perc)} = \mathbf{W}^{(Perc)} \mathbf{H}^{(Perc)}. \tag{4.20}$$

The phase of the original spectrogram is directly multiplied by each of the harmonic and percussive spectrograms. Finally, the harmonic and percussive audio signals are reconstructed by the inverse STFT.

4.2.3 Algorithm interpretation

The proposed update equations can be interpreted as a matrix decomposition process that controls the continuity of the basis vectors so as to shift the convergence point. As we have discussed for the case of the PLCA, the constraints prevent the NMF algorithm from converging to the local minima that is closest to the initialized values. Instead, the constraints shift the convergence point to the intended area. Accordingly, the weighting parameters α, β, γ, and δ affect the degree of the convergence point transition.

When the values of α and δ become smaller, the time activations of the harmonic components and the spectral bases of the percussive components are more likely to be shaped such that they have a continuous structure where the fluctuation is minimized. On the other hand, when β and γ become larger, the spectral bases of the harmonic components and the time activations of the percussive components may become discontinuous.

The role of the parameters can be explained using the mathematical formulas. Eq. 4.18, 4.19, 4.22, and 4.23 can be directly switched to alternative forms as follows:

$$\hat{\mathbf{H}}_{k,t}^{(Harm)} \leftarrow \tilde{\mathbf{H}}_{k,t-1}^{(Harm)} + \alpha \left(\tilde{\mathbf{H}}_{k,t}^{(Harm)} - \tilde{\mathbf{H}}_{k,t-1}^{(Harm)} \right) \tag{4.21}$$

$$\hat{\mathbf{H}}_{k,t}^{(Perc)} \leftarrow \tilde{\mathbf{H}}_{k,t-1}^{(Perc)} + \beta \left(\tilde{\mathbf{H}}_{k,t}^{(Perc)} - \tilde{\mathbf{H}}_{k,t-1}^{(Perc)} \right) \tag{4.22}$$

$$\hat{\mathbf{W}}_{f,k}^{(Harm)} \leftarrow \tilde{\mathbf{W}}_{f-1,k}^{(Harm)} + \gamma \left(\tilde{\mathbf{W}}_{f,k}^{(Harm)} - \tilde{\mathbf{W}}_{f-1,k}^{(Harm)} \right) \tag{4.23}$$

$$\hat{\mathbf{W}}_{f,k}^{(Perc)} \leftarrow \tilde{\mathbf{W}}_{f-1,k}^{(Perc)} + \delta \left(\tilde{\mathbf{W}}_{f,k}^{(Perc)} - \tilde{\mathbf{W}}_{f-1,k}^{(Perc)} \right). \tag{4.24}$$

The above formulas are interpreted as forming continuities[discontinuities] by suppressing[amplifying] the differences between the adjacent elements. If the weighting parameter is larger than 1, the differences are amplified. Considering our principle that the spectral bases of the harmonic components and the temporal bases of the percussive components are discontinuous, we can set the scopes of the parameters as

$$\beta > 1, \quad \gamma > 1. \tag{4.25}$$

In this case, the matrix elements in the left side of the Eq. 4.13 and 4.16 may become negative. The Eq. 4.14 and 4.18 can prevent the components from being negative.

On the other hand, weighting parameters smaller than 1 shape the bases such that they have less fluctuation. Based on our principle, the scopes of the parameters are determined as

$$\alpha < 1, \quad \delta < 1. \tag{4.26}$$

4.3 Performance Evaluation

In this section, we evaluate the proposed method by comparing it to conventional methods. To this end, the Signal Separation Evaluation Campaign (SiSEC) dataset [80] and the QUASI dataset [81] are used to investigate the objective and quantitative analysis results. To complement the weakness of the objective measures, the result of a subjective scoring test is provided. Also, toy

examples and audio demos will help intuitive understanding about the proposed method.

In each subsection, we compare the performance of the proposed method induced in the NMF-based framework using Ono *et al.*'s hard-mixing-based method, Ono *et al.*'s MAP-estimation-based method, Fitzgerald's median-filter-based method, Canadas-Quesada *et al.*'s NMF-based method, and Fitzgerald *et al.*'s KAM-based method. Experimental parameters for the conventional methods are set to the values presented in the literature, including the frame size, hop size, and sampling frequency, so as to provide the evaluation environment suitable for each method. For example, the sound mixtures and the ground truths of the harmonic and percussive sounds are down-sampled to 16 kHz for Ono *et al.*'s method.

The objective performances are measured using the BSS EVAL 3.0 toolbox, which is supported by the reference [61]. The toolbox is generally used for the evaluation of source separation algorithms. The considered measures include SIR, SAR, and SDR.

4.3.1 Parameter setting

Prior to the performance comparison with conventional methods, we have determined the evaluation parameters for the proposed method using the SiSEC 2015 dataset for professionally-produced music recordings (MUS) task [80]. The dataset is composed of a development set, and a test set, each of which contains 50 songs. Here, only the development set was used for the parameter setting. In the dataset, four sources—vocals, bass, drums, and other—are included for each song. The sound sources are provided in stereo, but they have been con-

Table 4.2 Evaluation parameters.

Parameter	Value
Frame size	4096
Hop size	1024
Window	Hamming
Sampling frequency (Hz)	44,100
Number of iterations	100
Number of bases (H/P)	750 (500/250)

verted into mono sounds before the experiment, by averaging the left and right channels. In the experiment, we assume that only drum sounds have percussive characteristics; and that the others are harmonic. The harmonic and percussive sounds are peak normalized and added to make a mixture.

Our main focus is to determine parameters that control continuity and discontinuity, whereas other variables are set to reasonable values. This is due to the fact that joint optimization of all these variables requires too much computation quantity. Table 4.2 summarizes the parameters and their values used for the experiment.

We have set the maximum iteration number to 100 by considering the computation time of the NMF algorithm. More iteration will lead to more stable performance by enabling the NMF algorithm to use sufficient time to converge. However, since the increase in the number of iterations is linearly proportional

to the increase in computation time, it has to be determined considering the computation cost/performance trade-off. According to our observations, 100 iterations is sufficient for the NMF algorithm to converge, and more iterations did not lead to better performance. This corresponds with the Canadas-Quesada *et al.*'s work, which also used 100 iterations for the NMF.

Additionally, the frame size and the hop size when performing the short-time Fourier transform (STFT) are set to 4,096 and 1,024, respectively. Not only does setting the hop size to 1/4 of the frame size allow perfect reconstruction of the separated audio signal when a Hamming window is used, but it also provides fine time resolution, which provides clearer separation of the vocal components. The sampling rate of the input data is 44,100 Hz.

The numbers of harmonic and percussive bases are set to 500 and 250, respectively, which is the same condition as in Canadas-Quesada *et al.*'s method. The optimal number of bases is difficult to estimate, even when the ground truth signals are given. For that reason, the number of bases is often determined heuristically in the NMF-based source separation methods. If the number of bases is too small, the estimation error will still remain large even after convergence. Hence, sufficient number of bases were used.

The basis vectors are initialized with random numbers that follow the uniform distribution of values defined over the interval $(0, 1)$, except for the percussive spectral bases: they are initialized to be *flat* vectors, i.e., the components of the vectors are all equal. The effect of the initialization of percussive bases will be discussed later.

In order to find the parameters that show the best SDR performance, we have altered each continuity/discontinuity parameter with small step sizes;

the parameter set with the maximum average SDR performance was finally selected. The harmonic and percussive components' SDRs are calculated for each song and are averaged in a dB scale. For such grid search, the following values were considered: $\alpha = \{0.6, 0.7, 0.8, 0.9\}$, $\beta = \{1.01, 1.05, 1.1, 1.2\}$, $\gamma = \{1.01, 1.05, 1.1, 1.2\}$, $\delta = \{0.8, 0.9, 0.95, 0.99\}$. The parameters were initially tested in a fixed interval with a step size of 0.1, and were further investigated with the values 1 ± 0.05 and 1 ± 0.01 for β, γ, and δ, since they showed highest SDR with the values close to 1. Among the parameters, α, which indicates temporal continuity, is proven to be most influential. The lower the value is, the less time-varying components are classified as harmonic components. If all parameters are set to 1, then the proposed algorithm works exactly the same as the standard NMF. Hence, they have to be finely tuned around 1. The maximum SDR is achieved when α, β, γ, and δ are 0.7, 1.05, 1.05, and 0.95, respectively.

4.3.2 Toy examples

Performance comparison

In this subsection, we evaluate the HPSS methods with two mixture examples of 10 seconds in duration. The first example consists of the piano sound and the hi-hat cymbal sound. These sounds match the assumptions of the conventional HPSS methods about sparsity well; thus, not only the proposed method but also the conventional methods are expected to successfully separate the harmonic components from the percussive components. The second example consists of the singing voice sound and the kick drum sound. Because the singing voice sound components often lack the temporal-side non-sparsity and the kick drum

components lack the spectral-side non-sparsity, it is relatively difficult to clearly separate them using the conventional methods. We observe the performance for each case, and compare the gaps between the two cases to investigate the robustness.

Note that the evaluation parameters were set to the same values used for the SiSEC 2015 dataset evaluation. Here, we have tested with two initialization methods: random initialization and flat initialization. "Random initialization" refers to the case where the basis components are initialized with random numbers that follow the uniform distribution defined over the interval $(0, 1)$. Meanwhile, "flat initialization" refers to the case where the vector components are all equal.

Fig. 4.5 shows the ground truths and the separation results of the first toy example. The piano attacks are mixed in the percussive spectrogram because of their wideband characteristics. Nevertheless, the separation results are relatively clear for all methods, as is confirmed in Table 4.3 (a), which shows the corresponding quantitative performance measures. As for the proposed method, flat initialization showed better performance in both harmonic and percussive SDR, but the performance difference was not significant.

The results for the second toy example are presented in Fig. 4.6; the corresponding performance measures are presented in Table 4.3 (b). These data show that the overall performances are lowered, as compared to the first example. Because the vocal components lack temporal continuity and the kick drum components lack spectral continuity, the conventional methods based on the continuity principle—Ono *et al.*'s methods, Fitzgerald's median filter-based method, and Fitzgerald *et al.*'s KAM-based method—show significantly de-

graded results. Canadas-Quesada *et al.*'s method partly covers the problem by employing the sparsity assumptions. However, as the L1-norm sparsity is not a suitable measure to distinguish between kick drums from other non-percussive spectra, it does not result in sufficient performance improvement. However, the proposed method still has higher performance by employing the continuity control of bases, as can be confirmed from the figure and the table.

When the percussive spectral bases are flat initialized, the proposed method shows improved performance of over 6dB compared to the conventional methods in both harmonic and percussive cases. However, we can further improve the performance with random initialization. This is due to the fact that the kick drum's spectra are not flat but rather sparse as shown previously. It results in the increase of the required time to converge. According to our investigation, the average SDR was improved to a level similar to that of random initialization when the algorithm was iterated 200 times. Hence, this can be used to verify that random initialization is more relevant to separating kick drum sounds, since it requires fewer iterations.

As can be seen in the first example, a typical percussive sound, such as a hi-hat drum, exhibits a clear separation performance regardless of the initialization method. As can be seen in the second example, kick drum components are well explained by a small number of iterations when randomly initialized, and more iteration is required when flat initialization is used.

Parameter analysis

We can extract some important information about the continuity/discontinuity parameters; how they affect the SDR, SIR, and SAR values of the har-

monic/percussive results. The second example was chosen to be used because it is appropriate to show performance variations. The first example is excluded because it can be separated well regardless of the parameter values. We aim to examine the effects of continuity-related parameters, α and δ. The discontinuity-related parameters have less effect on the performance because a perfectly continuous spectrum can be part of the percussive sound, but a perfectly discontinuous spectrum is not part of the harmonic sound.

Fig. 4.7 (a) shows the effect of α, which controls the continuity of the harmonic temporal bases. Strong imposition of the continuity prior (reduction of α) results in the increase of harmonic SIR, which affects the decrease of percussive SIR. It also decreases harmonic SAR because the ill-separated harmonic components generate artifacts. Fig. 4.7 (b) shows the relations between δ that regulates the continuity of percussive spectral bases and the SIR and SAR performances. Contrary to the case of α, the SIR of the percussive estimates increased as δ became smaller (when the continuity of the percussive basis was strongly imposed). Besides, these results also show the trade-off relations of SIR and SAR in both harmonic and percussive cases. Using the characteristics of the parameters derived from these analyses, the proposed method can be applied for various purposes, especially for tasks that require high SIR results.

4.3.3 SiSEC 2015 dataset

Performance comparison

In Table 4.4 (a), the performance of the maximal average SDR point is compared to that in other methods when they are measured in the SiSEC development set.

Table 4.3 Performances measured with the toy examples [dB].

	SDR		SIR		SAR	
	H	P	H	P	H	P
Ono (hard mix.)	7.70	7.40	10.91	18.46	10.85	7.81
Ono (MAP esti.)	12.70	9.02	17.37	16.62	14.60	9.94
Fitzgerald (med filt.)	11.84	17.60	14.97	22.90	14.86	19.14
Canadas-Quesada	**16.07**	15.31	22.06	20.64	**17.36**	16.88
Fitzgerald (KAM)	5.62	13.02	7.34	26.91	11.21	13.21
Proposed (flat)	15.48	**21.00**	30.93	**28.18**	15.61	**21.93**
Proposed (random)	13.79	14.95	**31.43**	16.86	13.87	19.94

(a) Example1

	SDR		SIR		SAR	
	H	P	H	P	H	P
Ono (hard mix.)	-15.15	7.61	-14.75	22.74	10.29	7.77
Ono (MAP esti.)	-15.03	4.85	-14.92	18.62	**15.79**	5.10
Fitzgerald (med filt.)	-15.33	7.71	-14.73	21.87	8.43	7.91
Canadas-Quesada	-14.45	7.16	-14.21	**31.59**	12.63	7.18
Fitzgerald (KAM)	-14.23	7.78	-13.75	26.11	9.49	7.85
Proposed (flat)	-8.72	14.78	-6.75	21.47	3.41	15.88
Proposed (random)	**1.63**	**19.41**	**25.16**	20.44	1.66	**26.24**

(b) Example2

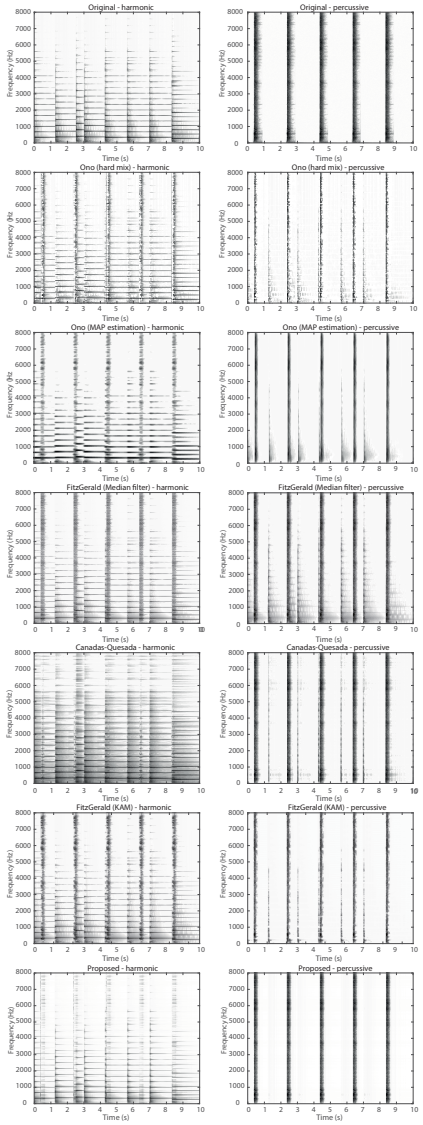


Fig. 4.5 HPSS results of the conventional methods and the proposed method with the mixture of piano and hi-hat sound. The harmonic spectrograms are aligned on the left, and the percussive spectrograms are aligned on the right.

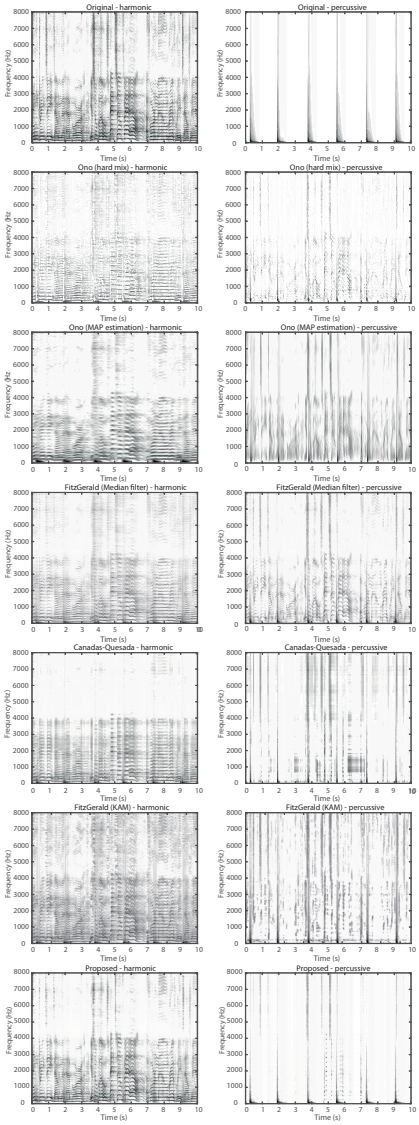


Fig. 4.6 HPSS results of the conventional methods and the proposed method with the mixture of singing voice and kick drum sound. The harmonic spectrograms are aligned on the left. The percussive spectrograms are aligned on the right.

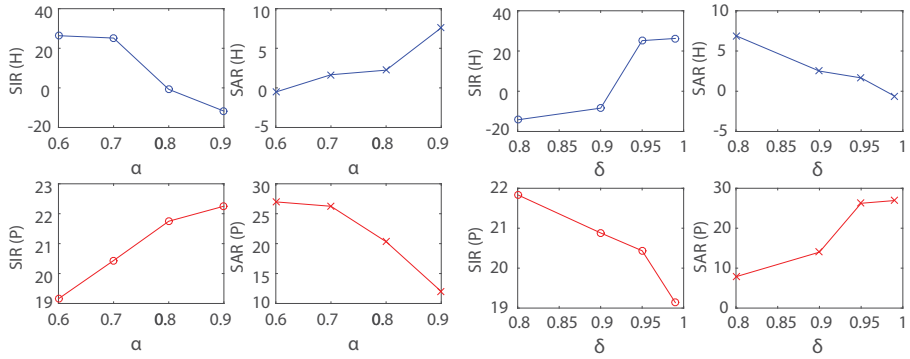


Fig. 4.7 Effect of continuity parameters to SIR and SAR values.

The harmonic and percussive SDR, SIR, and SAR are calculated for each song, and are then averaged. The proposed method shows the highest performance in the average SDR values for both harmonic and percussive cases.

It is not fair to objectively compare the performance from the previous experiment, because the parameters of the proposed method are the ones that are already trained in the development set. Therefore, we performed another experiment using the SiSEC test set.

Table 4.4 (b) shows the performance estimated in the SiSEC test set. Similarly, we compared the SDR, SIR, and SAR of each method. The evaluation parameters of the proposed method are set to the same values as those we used for the development set evaluation. A similar tendency in performance obtained in the previous experiment can be observed; the proposed method shows the best SDR performance by preserving good SIR and SAR values.

Ono *et al.*'s algorithms clearly demonstrate the trade-off relationship between SIR and SAR. Ono *et al.*'s hard-mixing-based method shows the lowest percussive SAR value among the algorithms. On the other hand, the percussive SIR shows the second highest performance after the proposed method. Ono *et*

al.'s MAP estimation-based method has the lowest percussive SIR, but it shows high percussive SAR performance. This tendency is found in both development and test sets. Thus, it is easy for us to obtain high values for either SIR or SAR, but it is difficult to obtain high SDR values by obtaining both high values. HPSS methods based on smoothness in a local area can use the SIR/SAR trade-off by broadening or narrowing the range of the area.

The most prominent advantage of the proposed method is clear separation of percussive sound. The percussive SDR of the proposed method has outperformed that of Fitzgerald's median filter-based method by 1.85dB in the test set. Such result can be regarded as a significant difference, considering that Fitzgerald's median filter-based method showed the second highest performance and the energy of harmonic components is greater than that of percussive components.

Two reasons can be suggested why the proposed method far outperforms the conventional methods. First, the conventional methods, except for the Canadas-Quesada *et al.*'s method, did not consider the frequency and time-side characteristics at the same time. In the proposed method, the NMF has allowed simultaneous observation of the frequency and time axis properties. Therefore, different properties could be imposed separately to the bases of each side. This has a strong effect on sources that have mismatch in the time domain (continuity in the time axis) and that we have to compensate the performance loss by using characteristics in the frequency domain (discontinuity in the frequency axis), such as vocal sounds. Secondly, we can impose the characteristics of harmonic/percussive bases most successfully by using the DF measure. This differentiates the proposed method from the conventional methods that uses

sparsity measures including the Canadas-Quesada *et al.*'s method.

Effect of Wiener filter

We have investigated the effect of Wiener filtering, when it is applied at the end of the spectrogram estimation step. In this case, $\mathbf{X}^{(Harm)}$ and $\mathbf{X}^{(Perc)}$ work as the harmonic and percussive masks. The Wiener-filtered spectrograms are defined as

$$\mathbf{G}_{f,t}^{(Harm)} = \frac{\left\{\mathbf{X}_{f,t}^{(Harm)}\right\}^2}{\left\{\mathbf{X}_{f,t}^{(Harm)}\right\}^2 + \left\{\mathbf{X}_{f,t}^{(Perc)}\right\}^2} \times \mathbf{X}_{f,t} \tag{4.27}$$

$$\mathbf{G}_{f,t}^{(Perc)} = \frac{\left\{\mathbf{X}_{f,t}^{(Perc)}\right\}^2}{\left\{\mathbf{X}_{f,t}^{(Harm)}\right\}^2 + \left\{\mathbf{X}_{f,t}^{(Perc)}\right\}^2} \times \mathbf{X}_{f,t} \tag{4.28}$$

where $\mathbf{G}^{(Harm)}$ and $\mathbf{G}^{(Perc)}$ denote the Wiener-filtered spectrograms of harmonic and percussive components, respectively.

As we can see in Table 4.4, Wiener filtering has improved average SIR values in both harmonic and percussive cases. This also leads to a decrease in SAR, according to the SIR/SAR trade-off. Overall, this has a positive impact on the harmonic SDR, but the SDR of the percussive components is decreased.

Effect of removing vocal components

We also analyze the case where vocal components are removed from the harmonic side. We can expect that the SIR values of the percussive side will be increased due to the fact that the estimated percussive sound will not contain the vocals that used to be not fully separated. Besides, the SAR values of the harmonic components will be also increased for the same reason.

Table 4.4 Performances of HPSS methods with the SiSEC dataset [dB].

	SDR		SIR		SAR	
	H	P	H	P	H	P
Ono (hard mix.)	7.36	-2.80	10.65	8.10	11.05	-1.45
Ono (MAP esti.)	7.55	-3.32	10.50	0.17	11.73	2.86
Fitzgerald (med filt.)	8.77	-0.57	**12.68**	4.48	12.00	2.86
Canadas-Quesada	8.08	-1.72	9.22	8.04	**16.00**	0.36
Fitzgerald (KAM)	8.07	-2.17	9.90	6.41	13.91	-0.01
Proposed (flat)	9.39	**1.44**	11.22	9.24	15.28	**3.09**
Proposed (Wiener filt.)	**9.56**	0.89	12.46	**9.97**	13.94	2.31

(a) Development set

	SDR		SIR		SAR	
	H	P	H	P	H	P
Ono (hard mix.)	6.93	-3.90	10.32	6.07	10.57	-2.04
Ono (MAP esti.)	7.20	-4.35	10.10	-1.30	11.29	**2.86**
Fitzgerald (med filt.)	8.34	-1.38	**12.20**	3.13	11.66	2.77
Canadas-Quesada	7.57	-3.04	8.86	6.32	**15.58**	-0.11
Fitzgerald (KAM)	7.79	-2.94	9.78	5.19	13.44	-0.33
Proposed (flat)	8.87	**0.47**	10.41	7.63	15.38	2.69
Proposed (Wiener filt.)	**9.28**	0.14	11.55	**8.83**	14.13	1.89

(b) Test set

Table 4.5 shows the experimental results when the vocal sound is removed from the mixtures. It can be seen that the proposed method still outperforms the conventional methods; however, the SDR performance differences have been reduced. Hence, it can be inferred that the proposed method has strengths in terms of separating vocal components. Moreover, the harmonic SAR and the percussive SIR tend to be increased as expected previously.

Effect of initialization

As we have verified in the previous subsection, initialization of percussive spectral bases can affect the performance. Moreover, the choice of initialization method depends on instrument composition. We aim to investigate the general effects and implications of altering initialization methods.

The previous experiment on the second example revealed that the randomly initialized bases converge faster than the flat initialized bases, when the percussive components contains kick drum components. If so, we can assume that replacing a part of flat initialized bases with randomly initialized bases will improve performance. Out of 250 spectral bases for percussive components, we have initialized k_{rand} bases with random numbers and the rests with flat vectors, and observed the changes occurred in the performance.

Fig. 4.8 (a) and (b) present the effect of presence of randomly initialized bases in the percussive spectral basis group on the performance metrics when investigated with the development set and the test set, respectively. It can be seen that the percussive SDR shows insignificant difference when $k_{rand} = 0, 5, 10$ in both cases. Meanwhile, the performance variation of harmonic SDR is much greater, and is maximized when $k_{rand} = 10$ in the development set and $k_{rand} =$

Table 4.5 Performances of HPSS methods with the SiSEC dataset in the absence of vocal sound [dB].

	SDR		SIR		SAR	
	H	P	H	P	H	P
Ono (hard mix.)	8.06	-2.51	11.13	9.43	11.85	-1.43
Ono (MAP esti.)	8.21	-3.14	11.1	0.5	12.35	2.79
Fitzgerald (med filt.)	9.59	-0.07	**13.28**	5.7	12.89	2.81
Canadas-Quesada	8.82	-0.41	9.86	**11.90**	**17.07**	0.67
Fitzgerald (KAM)	8.30	-2.20	10.29	6.19	13.95	0.25
Proposed (flat)	**9.90**	**1.48**	11.85	9.52	15.59	**3.04**

(a) Development set

	SDR		SIR		SAR	
	H	P	H	P	H	P
Ono (hard mix.)	7.61	-3.60	10.78	7.64	11.40	-2.14
Ono (MAP esti.)	7.73	-4.13	10.65	-0.95	11.77	**2.90**
Fitzgerald (med filt.)	9.13	-0.81	**12.78**	4.42	12.56	2.68
Canadas-Quesada	8.45	-1.62	9.28	**11.01**	**17.36**	-0.29
Fitzgerald (KAM)	8.29	-3.18	10.08	5.19	14.22	-0.79
Proposed (flat)	**9.35**	**0.55**	10.99	8.33	15.64	2.62

(b) Test set

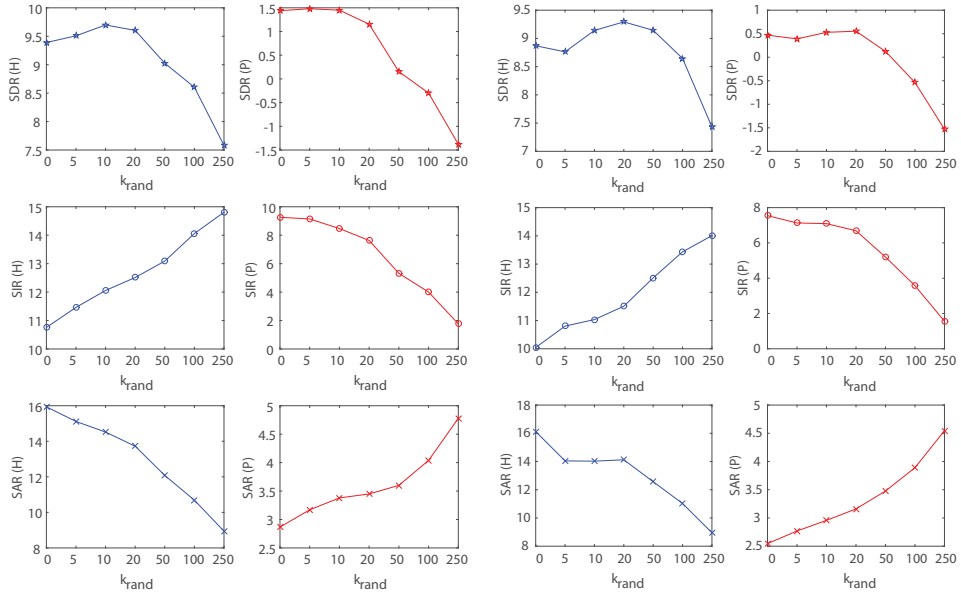


Fig. 4.8 Effect of randomly initialized bases on performance: (a) development set and (b) test set.

20 in the test set. When the k_{rand} is greater than 20, the SDR declines sharply in both harmonic and percussive cases. Hence, we can see that it is better to use 10 to 20 randomly initialized percussive bases when applying the proposed method to analyze real recordings. This is because some percussive components that cannot be separated with the flat initialized bases can be explained with randomly initialized bases. However, the randomly initialized bases show lower performance in overall separation performance, hence they are inappropriate to be used to replace the roll of flat initialized bases.

Another interesting effect of the initialization can be observed in the SIR and SAR values. As the number of randomly initialized bases increases, the harmonic SIR increases because fewer percussive components remain on the harmonic side. On the other hand, the percussive SIR decreases as the amount

of harmonic interference in the percussive side increases. Due to the SIR/SAR trade-off relationship, the harmonic SAR decreases and the percussive SAR increases. In addition to adjusting the continuity/discontinuity-related parameters, we can obtain various harmonic and percussive separation results by adjusting the basis initialization.

Effect of number of bases

As mentioned in the Chapter 1, the optimal number of bases cannot be predicted in advance of the experiment and we often set it to be sufficient. This is why we have used 750 bases (500 for harmonic, 250 for percussive) for the previous experiments. Here, we investigate the effect of the number of bases. To this end, we have set all parameters equivalent to *Proposed (flat)* in Table 4.4. For the test data, "AM Contra - Heart Peripheral" in SiSEC dataset is used. We have set the number of bases to 30, 45, 75, 150, 300, 450, 600, 750, 1500, 3000 and observed the transition in the performance. Note that the ratio between the number of harmonic and percussive bases is fixed to 2:1 in all cases as in the case of 750 bases.

Fig. 4.9 (a) and (b) show the performance transition according to the number of bases. It can be seen that the percussive SDR is not degraded severely when the total number of bases is larger than or equal to 300 (100 bases for the percussive sound). Hence, it can be inferred that about 150 bases out of 250 percussive bases were redundant in the former experiments. In case of harmonic components, more bases guarantee higher SDR performance due to the description of vocal components requires sufficient number of bases. However, it does not show meaningful performance gain when the harmonic basis number

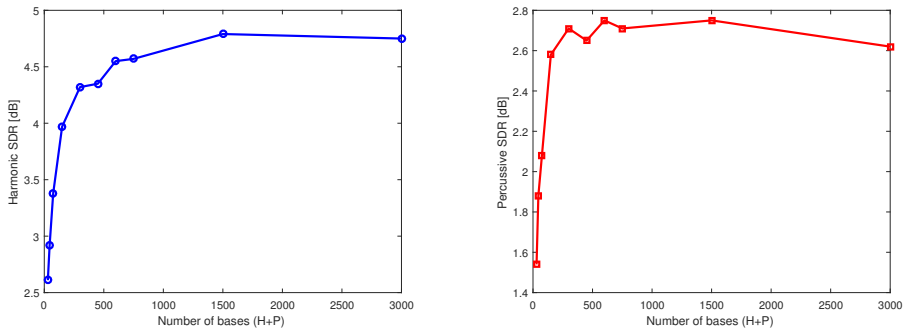


Fig. 4.9 Effect of number of bases.

exceeds 1000.

Imposing the constraints on the bases can affect the energy distribution of the bases. Different from the standard spectrogram decomposition algorithms that aim to find arbitrary bases that well describe the spectrogram, constrained decomposition targets to find a small number of meaningful bases. Thus, the constraint imposition makes the basis energy distribution be sparser. Inversely, we can check the meaningfulness of the bases with their energy distribution.

Table 4.6 shows the sparsity of the energy of the NMF bases measured in Gini index. It can be seen that the proposed method guarantees sparser energy distribution, which may indicate it enforces NMF to learn a smaller number of bases that have the meaningful structure. Note that the Gini index is larger in the case of the percussive components than the harmonic components because we have allocated fewer number of bases.

Iterative update process

Fig. 4.10 shows the change in the harmonic and percussive spectrograms, where the iteration count i is 1, 5, 20, and 100, respectively. In the experiment, a snip-

Table 4.6 Gini index of the energy distribution of the bases.

	Harmonic	Percussive	Overall
Proposed NMF	0.58	0.33	0.52
Standard NMF	-	-	0.33

pet from "Skelpolu–Resurrection" is used. The spectral and temporal bases are generally initialized to randomized values. Accordingly, the estimated spectrograms of the harmonic and percussive components show no directivity ($i = 1$). As the iteration proceeds, however, the bases are shaped such that they have spectral and temporal features that match our principle; thus, the spectrograms show horizontal/vertical directivity.

Computation time

Average computation time of each algorithm measured in the SiSEC developement and test set is presented in Table 4.7. NMF-based algorithms shows slower performance as predicted. This is because they estimate the bases using the entire spectrogram. The difference between the speed of the Canadas-Quesada *et al.*'s method and that of the proposed method is mainly due to the sampling rate. Fitzgerald *et al.*'s KAM-based method also contains iterative update formula, causing large computation amount. Note that the average length of the 100 experimental songs is 250.4 seconds.

The speed of the proposed method depends on the number of bases and the spectrogram size. Our experiment reveals that the computation time is linearly

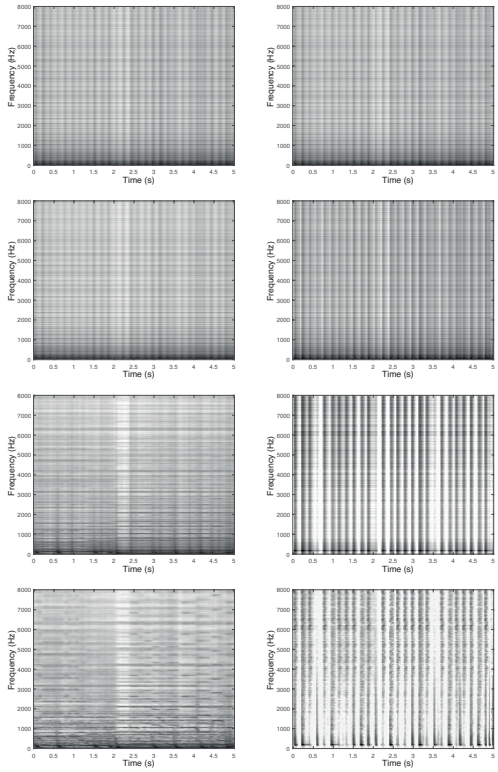


Fig. 4.10 The estimated spectrograms of the harmonic components (left) and percussive components (right) at the iteration number $i = 1, i = 5, i = 20, i = 100$ from top to bottom.

proportional to the time duration of the input sound. Nevertheless, since splitting the input sound into small segments also produces multiple input signals, it does not guarantee reduction of computation time. However, less number of bases can be used for shorter segments, which can lead to the computation time decrease. The trade-off relation between the performance and the speed exists because the bases can be estimated more accurately with the entire spectrogram, but the performance decrease is observed to be not severe when the

Table 4.7 Average computation time of HPSS algorithms in the SiSEC development and test set.

Algorithm	Computation time (sec)
Ono(hard mix.)	17.7
Ono (MAP esti.)	26.7
Fitzgerald (med. filt.)	13.1
Canadas-Quesada	115.7
Fitzgerald (KAM)	473.1
Proposed	311.3

input signal is divided into 60 second length, and the numbers of harmonic and percussive bases are set to 250 and 125, respectively. In this case, the average computation time can be reduced to 35 seconds.

4.3.4 QUASI dataset

For further generalization of the performance, we evaluated the HPSS methods using the QUASI dataset [81]. The QUASI dataset is composed of 11 songs mixed by a professional sound engineer, in addition to their individual tracks. Among the songs included, *"Parting friends"* by *Emily Hurst* is excluded in this experiment, because it is a vocal-only song and does not contain any percussive component. The considered percussive source list includes *Drums, Drms,*

Table 4.8 Performances of HPSS methods with QUASI dataset [dB].

	SDR		SIR		SAR	
	H	P	H	P	H	P
Ono (hard mix.)	3.94	-1.53	7.09	7.39	8.30	0.31
Ono (MAP esti.)	4.18	-3.09	6.05	0.43	10.55	2.41
Fitzgerald (med filt.)	5.46	0.90	**8.73**	4.91	9.51	**4.78**
Canadas-Quesada	4.03	-2.15	4.72	**9.11**	**14.88**	-0.30
Fitzgerald (KAM)	4.83	-0.12	6.21	7.93	12.38	1.66
Proposed (flat)	**6.17**	**2.50**	7.95	9.06	12.50	4.65

Drums_loop, kick, snare, kick_snare, hihat, tamb_blip, wood, 909, drms_loop1, drms_loop2, OH1, OH2, shaker, tamb, drum_kit, talk_drum. Here, we compare performance with other algorithms using only flat initialization.

The measured performances are summarized in Table 4.8. Fitzgerald's median filtering-based method and Canadas-Quesada *et al.*'s method clearly show the trade-off relation of SIR/SAR. Even though they have defeated the proposed method in the SIR and SAR, the proposed method shows the best harmonic and percussive SDR by achieving the high SIR and preserving a high SAR as in the previous experiments. In general, we conclude that the proposed method outperforms the conventional methods.

4.3.5 Subjective performance evaluation

In previous subsections, the objective scores of the proposed method were investigated. However, the perceived quality of the separated sounds can be assessed

differently, according to Emiya *et al.* [82]. In this subsection, we describe the methods of the subjective scoring test and discussions of the results.

A total of 33 normal-hearing subjects participated in the test. They were requested to score the separated harmonic and percussive results on a 1 to 5 scale, which is largely known as the mean opinion score (MOS) test [83]. The original mixture sound and the ground truth of the harmonic and percussive sounds were provided along with the separated results. The English-translated version of the test sheet is publicly accessible at `http://marg.snu.ac.kr/hpss_test_form/`. Also, as indicated on the test sheet, sound examples used for the test are provided on the website `http://marg.snu.ac.kr/hpss_test/`. Randomly selected 10 songs from the SiSEC 2015 database were used for the test, none of which belonged to the same genre. The songs were cut at random positions to have duration of 20 seconds.

Table 4.9 shows the average subjective scores and the corresponding SDR values of each algorithm. The average SDR values of the test songs show the similar tendency to the total mean presented in Table 4.4, proving randomness of the test song selection. The p-values in Table 4.9 are calculated to examine the statistical difference between the conventional method and the proposed method.

Focusing on the subjective scores of harmonic sounds, it can be observed that the proposed method shows the best performance. A paired t-test revealed that there is a significant difference ($p < 0.06$) between the proposed method and other methods, except for the case of Fitzgerald *et al.*'s median filtering-based method ($p = 0.066$). In the results of percussive sounds, the proposed method outperforms all other methods with the significant difference.

Table 4.9 Subjective scores and corresponding objective measures (SDR).

	Average SDR [dB]		Subjective score (p-value)	
	H	P	H	P
Ono (hard mix.)	6.89	-3.86	3.15 (1.09×10^{-2})	2.54 (1.65×10^{-10})
Ono (MAP esti.)	6.90	-3.28	3.08 (9.27×10^{-4})	2.40 (2.66×10^{-9})
Fitzgerald (med filt.)	**8.14**	-1.31	3.22 (6.61×10^{-2})	2.45 (7.09×10^{-8})
Canadas-Quesada	7.43	-2.56	2.84 (2.38×10^{-9})	2.59 (2.22×10^{-10})
Fitzgerald (KAM)	7.68	-3.36	3.02 (8.62×10^{-5})	2.26 (3.66×10^{-13})
Proposed (flat)	8.13	**0.56**	**3.34**	**3.44**

4.3.6 Audio demo

The demo audio clips of the proposed method and conventional methods are provided on the website `http://marg.snu.ac.kr/hpss_audio_demo/`. It also consists of original mixture and harmonic/percussive sources of seven songs of various genres. We believe a perceptual evaluation from these demos supports the quantitative results presented in the previous sections.

4.4 Summary

In this chapter, we proposed a novel method for HPSS that exploits the continuity and discontinuity of the bases. Previous HPSS research studies have claimed that the harmonic components have spectral sparsity and temporal smoothness, whereas the percussive components have the spectral smoothness and temporal sparsity. However, most of the conventional methods fail to fully

consider these characteristics, which results in low performance.

A closer examination of the spectra of harmonic and percussive sounds reveals that continuity measurement, rather than sparsity, is proven to be a more appropriate feature to distinguish harmonic and percussive components. Based on the observation, we proposed a novel HPSS algorithm that exploits continuity control in the iterative update formula of the PLCA algorithm and reformulated it in a NMF framework to reduce the computational cost.

The performance of the proposed method was verified both qualitatively and quantitatively in comparison with the conventional methods. The results showed that the proposed method outperforms conventional methods. Since we can remove all kinds of percussive sound from the mixture using this method using the proposed methods, we now focus on HISS problem in the next chapters.

Chapter 5

Informed Approach to Harmonic Instrument sound Separation

5.1 Introduction

In this chapter, we present an informed approach to HISS problem. Music source separation aims to restore original source sounds from a mixture sound. It can be used as a pre-processing step for many music signal processing techniques such as music transcription, remixing, up-mixing, instrument identification, and equalization [6]. In particular, the task has been studied extensively in under-determined scenarios where the number of channels is smaller than the number of instruments, but still shows limited performance.

Recently, many studies that use additional information have been presented such as music score [20], user-guided audio signals [25], and manually provided annotations [23]. These types of information help overcome poor performance, but the environment in which side-information is available is extremely limited.

Thus, it is desirable that source separation methods use less side-information.

When only musical instrument sounds exist in the music, the use of side-information use can be reduced by considering the timbre characteristics of the musical instruments. Especially in cases in which harmonic instruments and drums coexist, their characteristics that appear in the time-frequency representation have been used for separation. Our previous work focused on the spectral features of harmonic and percussive sounds [51], and it was also extended to simultaneously consider the time and frequency domain aspects of the instruments [63].

The separation of harmonic instrument sounds is a more challenging problem as the differences of the spectral and temporal characteristics are less obvious compared to the harmonic-percussive source separation. Spiertz and Gnann presented the basis clustering algorithm as a post-processing of NMF [84]. Fitzgerald *et al.* used shift-invariant non-negative tensor factorization [85]. Ozerov and Fevotte focused on the mixing procedure [11].

Other studies on harmonic instrument sound separation adopt *source-filter model*. Heittola *et al.* trained bases for instruments and used them for separation [86]. Rodriguez-Serrano *et al.* also made instrument-dependent models and used them for separation [87]. However, the pre-training process is not always available. Klapuri *et al.*'s work extended Heittola *et al.*'s method to separately estimate approximated spectral envelopes and their corresponding excitations without the pre-training process [88]. But their work fails to precisely assess spectral envelopes, since it roughly approximates the envelopes using band-pass filter banks. Ozerov *et al.* developed flexible audio source separation toolbox (FASST), in which spectra are also split into excitations and

filter parts [89] [90].

In this chapter, an informed approach to HISS problem is presented for situations in which audio segments of the used instruments can be obtained. From the segments, we can extract the spectral envelopes of the instruments, which are used for the instrument sound separation. Clearly, this method is based on the source-filter model with linear predictive coding (LPC), since the envelopes are extracted via linear prediction. The spectral bases of the NMF algorithm are partitioned ahead of the iteration and then forced to resemble the envelopes. To this end, Dirichlet prior is also used for this basis shaping. The comparative evaluation reveals that the proposed method outperforms the other conventional methods.

The rest of the chapter is organized as follows. Section 5.2 describes the proposed method in the matrix factorization framework. Section 5.3 shows the experimental results with real recordings. Conclusions are presented in Section 5.4.

5.2 Proposed method

In this section, we present detailed description about the proposed HISS method. This section is composed of three subsections that present excitation-filter model, linear predictive coding, and the proposed NMF-based spectrogram decomposition procedure.

5.2.1 Excitation-filter model

The time domain mixture signal is generated by summing individual sources as

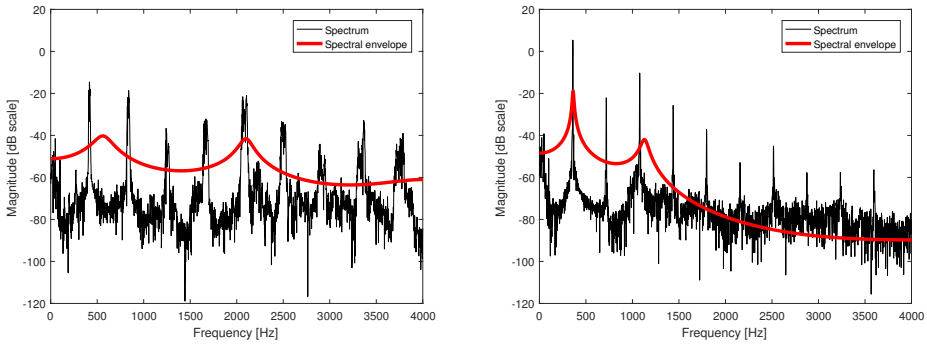
$$x = \sum_{i=1}^{I} x_i \tag{5.1}$$

where x denotes the mixture sound, x_i denotes the sound of the i-th instrument, and I denotes the number of instruments. When this time domain signal is converted into a spectrogram, it can be similarly represented as

$$\begin{aligned} \mathbf{X} &= \sum_{i=1}^{I} \mathbf{X}_i \\ &\approx \sum_{i=1}^{I} \sum_{k \in \Phi_i} \mathbf{w}_k \mathbf{h}_k \end{aligned} \tag{5.2}$$

where $\mathbf{X_i}$ denotes the magnitude spectrogram of instrument i, Φ_i denotes the index set of bases that explain $\mathbf{X}_i$, and $\mathbf{w}_k$ and $\mathbf{h}_k$ denote the k-th spectral basis and its time activation, respectively. The spectrogram conversion error is assumed to be small and negligible. For the convenience of description, we assume that $||\mathbf{w}_k||_1 = 1$ for all k. $\mathbf{w}$ and $\mathbf{h}$ can be estimated with the matrix decomposition algorithms such as PLSA, PLCA [41], and NMF [91].

We also assume that the instrument sounds can be represented using the source-filter model, which we alternatively address as *excitation-filter model* to avoid term collision. The excitation-filter model has been widely used to analyze the speech production mechanism [92]. According to the excitation-filter model, the timbre of an instrument is determined by its *filter*, whereas the pitch is determined by the excitation signal. Fig. 5.1 illustrates the spectrum and the corresponding LPC spectral envelope of violin and clarinet plotted on a log-scale. As the excitation signal is filtered by the instrument's resonant



(a) (b)

Fig. 5.1 Spectrum and corresponding spectral envelope computed via linear prediction of (a) violin and (b) clarinet.

structure, a spectral basis (or a spectrum of an instrument, equivalently) $\mathbf{w}_k$ has to be represented as

$$\mathbf{w}_k = \mathbf{v}_i \odot \mathbf{e}_k \tag{5.3}$$

where $\mathbf{v}_i$ is the filter's frequency response of instrument i that can be alternatively referred to as *spectral envelope*, $\mathbf{e}_k$ is the spectrum of the k-th excitation signal, the operator $\odot$ denotes the element-wise multiplication, and $k \in \Phi_i$.

Note that the above Eq. 5.3 can be satisfied only if a proper source separation is applied. During the ongoing matrix decomposition iteration, the spectral envelopes vary for each k. However, the proposed method focuses on the reverse direction of this theory: *Can a group of bases successfully reconstruct the sound of an instrument, if we can make the basis envelopes equal to the true envelope of the instrument?* Since the proposed method requires the estimation of the true spectral envelopes of the instruments and the basis envelopes, the representative spectral envelope extraction method is presented in the next subsection.

5.2.2 Linear predictive coding

Spectral envelope can be obtained using LPC, which assumes that the filter can be approximated by a finite number of poles. In this subsection, we describe how we can obtain LPC coefficients and how they can be extensively applied to the spectral bases of the NMF algorithm.

Calculation of LPC coefficients

LPC aims to calculate the infinite impulse response (IIR) filter coefficients $\mathbf{a} = (a_1, a_2, \ldots, a_M)^T$ that best predict the signal value minimizing the energy of the error signal, represented as

$$\begin{aligned} \mathbf{a} &= \arg\min_{\mathbf{a}} E\left\{\|err\,(n)\|^2\right\} \\ &= \arg\min_{\mathbf{a}} E\left\{\left\|y\,(n) - \sum_{m=1}^{M} a_m y\,(n-m)\right\|^2\right\}, \end{aligned} \tag{5.4}$$

where y is a real, time domain signal, err is the error signal, and M is the number of filter coefficients. The accurate estimation of LPC coefficients is important, since the filter's frequency response, namely spectral envelope, is represented as

$$H\,(z) = \frac{1}{1 - \sum_{m=1}^{M} a_m z^{-m}}. \tag{5.5}$$

The problem of computing LPC coefficients can be converted into an alternative form, which is referred to as *autocorrelation method*. It can be mathematically represented as

$$\mathbf{Ra} = \mathbf{r} \tag{5.6}$$

where $\mathbf{r}$ is a vector of autocorrelations of y defined as $\mathbf{r} = (r_{yy}\,(1)\,, r_{yy}\,(2)\,, \ldots\,, r_{yy}\,(M))^T$, and $\mathbf{R}$ is an autocorrelation matrix defined

as

$$\mathbf{R} = \begin{pmatrix} r_{yy}(0) & r_{yy}(1) & \cdots & r_{yy}(M-1) \\ r_{yy}(1) & r_{yy}(0) & & r_{yy}(M-2) \\ \vdots & & \ddots & \vdots \\ r_{yy}(M-1) & r_{yy}(M-2) & \cdots & r_{yy}(0) \end{pmatrix} \tag{5.7}$$

where $r_{yy}(m) = E\{y(n)\, y(n-m)\}$. From the above formula, we can observe that the calculation of LPC coefficients does not necessarily require the original time domain signal y and that it can also be calculated by the autocorrelations. As $\mathbf{R}$ is a Toeplitz matrix, $\mathbf{a}$ can be easily obtained using Levinson-Durbin recursion [93].

Envelope of spectral bases

Consider the spectrum $\mathbf{Y}$ of a time domain signal $y(n)$, and its magnitude $|\mathbf{Y}|$. The spectral envelope of the magnitude spectrum can be directly obtained without the spectrum-to-time domain signal conversion process. According to the Wiener-Khinchin theorem, the computation of autocorrelation is simplified as

$$\begin{aligned} \mathbf{r}_{yy} &= IFFT[\mathbf{S}_{yy}] \\ &= IFFT[\mathbf{Y}\mathbf{Y}^*] \\ &= IFFT\left[|\mathbf{Y}|^2\right] \end{aligned} \tag{5.8}$$

where $\mathbf{S}_{yy}$ is the power spectral density of $y(n)$, $\mathbf{r}_{yy}$ is defined as $\mathbf{r}_{yy} = (r_{yy}(0), r_{yy}(1), \ldots, r_{yy}(M))^T$, $IFFT[\bullet]$ denotes the inverse fast Fourier transform, and $(\bullet)^*$ denotes the complex conjugate. Here, we can see that the magnitude spectrum $|\mathbf{Y}|$ has sufficient information to attain the autocorrelations, which in turn can be used to estimate LPC coefficients. These LPC coefficients are finally used to estimate the spectral envelope.

5.2.3 Spectrogram decomposition procedure

According to the equivalence of PLCA and NMF that we have proven in the chapter 2, we do not present how we can implement on the PLCA framework. Instead, we describe our method on the NMF framework. Fig. 5.2 illustrates the overall procedure of how the proposed method works. The input mixture audio is transformed into a magnitude spectrogram and is decomposed using the proposed modified NMF. In so doing, the bases are randomly initialized first, and the spectral bases and their corresponding time activations are estimated iteratively afterwards. After the estimation, $\mathbf{w}$ is divided into two parts, envelope $\mathbf{v}$ and excitation $\mathbf{e}$, by means of LPC. The envelopes of the bases that belong to an instrument's index set are replaced by the true envelope of the instrument. The spectral bases are then reconstructed by multiplying the new envelope and the excitation followed by the next iteration. After the iteration is finished, the spectrograms are reconstructed for each instrument. Finally, the audio signals are reconstructed.

The proposed method modifies the NMF update equations that minimize the KL divergence. It can be mathematically represented as

$$\hat{\mathbf{H}}_{k,t} \leftarrow \frac{\mathbf{H}^{(l)}_{k,t} \sum_{f} \left\{ \mathbf{W}^{(l)}_{f,k} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{f'} \mathbf{W}^{(l)}_{f',k}} \tag{5.9}$$

$$\tilde{\mathbf{W}}_{f,k} \leftarrow \frac{\mathbf{W}^{(l)}_{f,k} \sum_{t} \left\{ \mathbf{H}^{(l+1)}_{k,t} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{t'} \mathbf{H}^{(l+1)}_{k,t'}} \tag{5.10}$$

$$\mathbf{H}^{(l+1)}_{k,t} \leftarrow \hat{\mathbf{H}}_{k,t} \sum_{f} \tilde{\mathbf{W}}_{f,k} \tag{5.11}$$

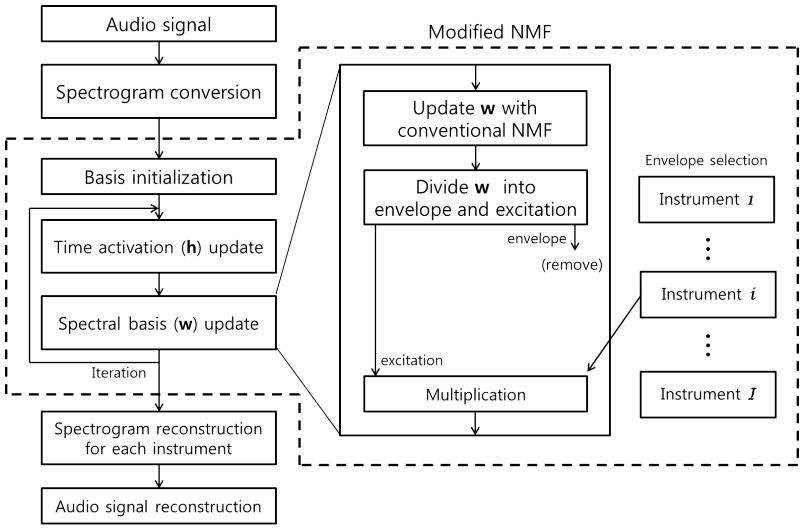


Fig. 5.2 Overview of the proposed method.

$$\hat{\mathbf{W}}_{f,k} \leftarrow \frac{\tilde{\mathbf{W}}_{f,k}}{\sum_{f} \tilde{\mathbf{W}}_{f,k}} \tag{5.12}$$

$$\mathbf{w}_k^{(l+1)} \leftarrow \breve{\mathbf{v}}^{(i)} \odot \mathbf{e}^{(k)}, k \in \Phi_i \tag{5.13}$$

where $\mathbf{H}$ is the $K \times N$ matrix of time activations, of which k-th row is $\mathbf{h}_k$, $\mathbf{W}$ is the $M \times K$ matrix of spectral bases, of which k-th column is $\mathbf{w}_k$, $\hat{\mathbf{W}}$ and $\tilde{\mathbf{W}}$ denote the temporarily adopted variables, which have the same size as $\mathbf{W}$, $\hat{\mathbf{H}}$ denotes the temporarily adopted variable that has the same size as $\mathbf{H}$, $\tilde{\mathbf{X}}$ is the estimated spectrogram reconstructed with up-to-date $\mathbf{W}$ and $\mathbf{H}$, $\mathbf{v}^{(k)} = \left(v_1^{(k)}, \ldots, v_F^{(k)}\right)^T$ and $\mathbf{e}^{(k)} = \left(e_1^{(k)}, \ldots, e_F^{(k)}\right)^T$ denote the spectral envelope and excitation spectrum of the k-th column of $\hat{\mathbf{W}}$ represented as $\hat{\mathbf{w}}_k$, respectively, $\breve{\mathbf{v}}^{(i)}$ is the true spectral envelope of instrument i, and f and t denote the index of frequency bin and time frame, respectively. Note that Eq. 5.11 and 5.12 are the normalization stages for the spectral bases.

Spectral envelope of bases

The spectral envelope of $\hat{\mathbf{w}}_k$ can be obtained as

$$\mathbf{r}^{(k)} \leftarrow IFFT\left[\{\hat{\mathbf{w}}_k\}^2\right] \tag{5.14}$$

$$\begin{aligned}\mathbf{a}^{(k)} &= \left(a_0^{(k)},\, a_1^{(k)},\, \ldots,\, a_M^{(k)}\right)^T \\ &\leftarrow LevinsonDurbin\left(\mathbf{r}^{(k)}\right)\end{aligned} \tag{5.15}$$

$$v_f^{(k)} \leftarrow \left| \frac{\eta^{(k)}}{1 - \sum_{m=1}^{M} \left\{ a_m \exp\left(-\mathbf{i} 2\pi \frac{f}{F}\right)\right\}} \right| \tag{5.16}$$

$$e_f^{(k)} \leftarrow \frac{\hat{\mathbf{W}}_{f,k}}{v_f^{(k)}} \tag{5.17}$$

where $\mathbf{r}^{(k)}$ contains the autocorrelations of $IFFT\left[\hat{\mathbf{w}}_k\right]$, $\eta^{(k)}$ is the normalization constant to make $||\mathbf{v}^{(k)}||_1 = 1$, and $LevinsonDurbin\left(\bullet\right)$ is the function that calculates the $M+1$ dimensional vector $\mathbf{a}^{(k)}$ of LPC coefficients by means of Levinson-Durbin recursion. Note that the imaginary number $\mathbf{i}$ is differentiated from the instrument index i, and $a_0^{(k)} = 1$ by the definition of LPC.

True spectral envelope of an instrument

The true spectral envelope $\breve{\mathbf{v}}^{(i)}$ is computed through the similar step. We assume that an audio segment $\breve{x}_i$ is given for all i. First, it is converted to a magnitude spectrogram $\breve{\mathbf{X}}$ and then the spectral envelopes $\breve{\mathbf{v}}_t$ of each frame $\breve{\mathbf{x}}_t$ is calculated. Then they are averaged as

$$\breve{\mathbf{v}}^{(i)} = \frac{\sum_t \left\|\breve{\mathbf{x}}_t\right\|_1 \breve{\mathbf{v}}_t}{\left\|\sum_t \left\|\breve{\mathbf{x}}_t\right\|_1 \breve{\mathbf{v}}_t\right\|_1} \tag{5.18}$$

Signal reconstruction

After the iteration, the estimated spectrograms of each instrument are reconstructed as

$$\hat{\mathbf{X}}_i = \sum_{k \in \Phi_i} \mathbf{w}_k \mathbf{h}_k \tag{5.19}$$

where $\hat{\mathbf{X}}_i$ denotes the estimated spectrogram of instrument i. These spectrograms are converted to time domain signals by means of inverse short-time Fourier transform.

5.3 Performance evaluation

In this section, the comparative evaluation is performed with the conventional HISS methods. For the objective comparison of the performances, we have used audio dataset of real recordings and the representative performance indicators.

5.3.1 Experimental settings

The performance of the proposed method is evaluated and compared with the method of Klapuri *et al.* and the FASST of Ozerov *et al.* Bach 10 dataset which consists of 10 pieces is used for the evaluation [94]. The dataset contains real recordings of four instruments; violin, clarinet, saxophone, and bassoon. We analyze a total of six cases where two out of four instrument sounds are linearly mixed. Note that the sounds are amplified or suppressed in advance to have same energies in the mixture.

Klapuri *et al.*'s method is based on two important assumptions; the number of notes in a frame should remain constant for all frames, and the fundamental

Table 5.1 Experimental parameters.

Parameter	Value
Sampling rate (Hz)	44,100
Frame size / Hop size	4096 / 1024
Number of iterations	100
Number of bases (per instrument)	40
LPC order (M)	3

frequencies of the notes are known. Bach 10 dataset satisfies the conditions because all the instruments in the dataset are monophonic and the ground truth fundamental frequencies are provided in the dataset. FASST 2.0 provides a number of options to perform sound separation. Among the options, we assumed instantaneous mixing scenario with the fixed adaptability, while the rest options are set adaptive. Two types of time-frequency representations, *erb* and *stft*, are tested.

Table 5.1 shows the evaluation parameters used for the experiment of the proposed method. The optimal number of bases may correspond to the number of notes of an instrument. However, it is impossible to recognize it in advance. Consequently, we set the number of bases so that it is sufficient to account for every note in the music. Considered evaluation metrics are SDR, SIR, and SAR as in the previous chapters. Note that SDR is the representative performance measure.

The true spectral envelope is learned in a two different ways. At first, a 5 second audio clip of the mixed instruments are randomly picked and cut

among the 9 other pieces. Because this can be considered not fair to the other methods, we have also used an external dataset for the true envelope extraction: real world computing (RWC) music database [95]. The various sounds of the instruments that are mixed in the Bach 10 dataset are contained in the RWC database. Among the available playing styles and dynamics, we have selected the sounds with normal-playing style and piano-level dynamics. The sounds are concatenated in advance and the true envelope is extracted with it.

5.3.2 Performance comparison

Table 5.2 shows the performance metrics averaged over the 10 pieces and 6 cases of instrument combinations. It can be seen that the Klapuri *et al.*'s method shows the highest SAR whereas the FASST with erb shows the lowest SAR. Nevertheless, 8.95dB can be considered to be a high value hence no methods seem to show defects in this result. Due to the SIR-SAR trade-off, FASST and Klapuri *et al.*'s method shows the low SIR value considering the SAR value. Especially, FASST shows the lowest performance among the methods. One possible reason for the low performance is its blindness: it lacks the aid of additional side-information. The proposed method shows better performance than the conventional methods regardless of the method used to get the true envelope. When the RWC database is used to compute true envelope, the proposed method's performance further increases compared to the case where Bach 10 dataset is used. This is because the lengths of the concatenated audio segments are longer than 5 seconds, hence it might have gotten more chance to obtain the envelopes close to the real one. This is meaningful in that it shows the possibility to accurately estimate the true envelope even with the external dataset.

Table 5.2 Performances measured with the Bach 10 dataset (dB).

	SDR	SIR	SAR
FASST (stft)	0.50	1.80	9.17
FASST (erb)	0.18	1.44	8.95
Klapuri	1.37	1.79	**15.29**
Proposed (Bach 10)	4.00	7.01	9.80
Proposed (RWC)	**4.52**	**7.23**	10.63

Further investigation on the true envelope extraction method is presented in the next subsection.

5.3.3 Envelope extraction

In the previous experiment, we have obtained the true envelope with the entire audio clip in the RWC database. In this subsection, we investigate the effect of the audio clip length and pitch that is used to calculate the true envelope. This is important since how much data is required to precisely estimate the true envelope has to be examined before we apply it to realistic environments.

Table 5.3 shows the performance transition according to the pitch of the training data. We have performed four experiments each uses an audio segment of 5 second for the envelope training. In the first experiment (*Random pitch*), the training data is obtained via cutting the original audio clip used for the envelope training in the previous experiment to 5 second at random position. Other experiments divide the original audio clip into low, middle, and high pitch parts and select one of them to cut the 5 second audio segment from the

Table 5.3 Effect of pitch in envelope training.

	SDR	SIR	SAR
Random pitch	2.74	5.61	10.64
Low pitch	**3.30**	**5.62**	10.51
Middle pitch	2.93	5.61	10.64
High pitch	1.93	5.49	**11.18**

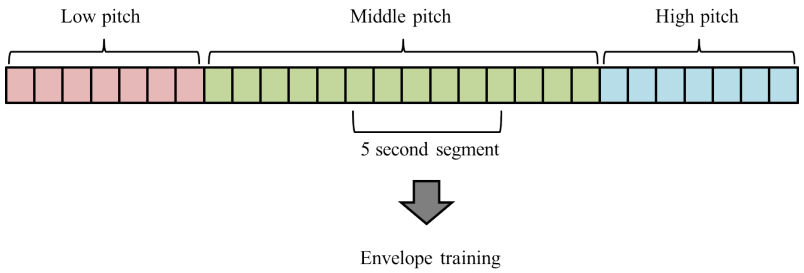


Fig. 5.3 Method to select an audio clip for envelope training.

random position as can be seen in Fig. 5.3. First, we can see that the SDR has decreased severely (from 4.52 to 2.74) when we only use a partial training data compared to the previous experiment. Then we can confirm that using low pitch part to train envelope shows better SDR performance even though the SIR and SAR do not show meaningful difference in performance. This is due to the fact that the high pitched sounds show sparser spectral energy distribution and it can distort the envelope shape.

5.4 Summary

In this chapter, we proposed a novel approach to separate harmonic instrument sounds by applying envelope constraints to spectral bases in NMF. The proposed approach focused on the spectral envelope which assigns distinct timbre to a sound. On the basis of the excitation-filter model, we could decompose a spectrum (or a spectral basis) to its spectral envelope and its excitation. The modified iterative update equations of the NMF with the true envelope constraints led to the successful separation of the harmonic instrument sounds. Performance evaluation with real recordings proved that it showed the highest performance compared to the conventional methods. In the next chapter, this method is extended to the blind scenario where no additional information is available.

Chapter 6

Blind Approach to Harmonic Instrument sound Separation

6.1 Introduction

In this chapter, the previous approach to HISS problem is extended to the case where no additional information is provided. The informed approach showed high performance compared to the conventional ones, however, it requires users to provide the sounds of each instrument. This enables the bases to have the true envelope, which in turn makes the excitation parts adapted to represent the pitch of the instrument.

In addition to this, the blind approach aims to let both spectral envelope and excitation separately converge to the real ones. To this end, the bases in a group are forced to have the same spectral envelope in performing the matrix decomposition. The spectral envelope of each basis is calculated through LPC and all envelopes of each basis in the same group are averaged. This is because

the spectral envelope is determined for each instrument whereas the excitations can differ from other bases that belong to the same group. As the iteration proceeds, the average spectral envelopes of each group converge to the true spectral envelopes of the instruments.

The strengths of the proposed method are its simplicity and flexibility; it requires neither additional musical information such as musical scores and time-frequency annotations nor assistance of f_0-estimation techniques. Since it only requires the provision of the number of instruments, it can be applied to the various cases where the musical information is not available. Consequently, the proposed method is applicable to multi-pitch scenario, regardless of the number of simultaneously existing musical notes.

The rest of the chapter is organized as follows. Section 6.2 depicts the proposed method in the NMF framework. Section 6.3 shows the experimental results with the dataset and compares the performance with the conventional methods. Section 6.4 summarizes the chapter.

6.2 Proposed method

In this section, we describe how we extend the informed approach to the blind scenario. As similar to the previous approach, it is mainly based on the NMF with the generalized Dirichlet prior. Fig. 6.1 shows the structural overview of the proposed method. Similar to the informed approach, the input signal is converted into a spectrogram and decomposed using the NMF to minimize the KL divergence between $\mathbf{X}$ and $\tilde{\mathbf{X}}$. During the iteration, the spectral envelopes of the bases belong to the same group are averaged. Then the mean envelope

is applied to the bases in the next step. This process can be represented in the mathematical formula as

$$\hat{\mathbf{H}}_{k,t} \leftarrow \frac{\mathbf{H}^{(l)}_{k,t} \sum_{f} \left\{ \mathbf{W}^{(l)}_{f,k} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{f'} \mathbf{W}^{(l)}_{f',k}} \tag{6.1}$$

$$\tilde{\mathbf{W}}_{f,k} \leftarrow \frac{\mathbf{W}^{(l)}_{f,k} \sum_{t} \left\{ \mathbf{H}^{(l+1)}_{k,t} \mathbf{X}_{f,t} \Big/ \tilde{\mathbf{X}}_{f,t} \right\}}{\sum_{t'} \mathbf{H}^{(l+1)}_{k,t'}} \tag{6.2}$$

$$\mathbf{H}^{(l+1)}_{k,t} \leftarrow \hat{\mathbf{H}}_{k,t} \sum_{f} \tilde{\mathbf{W}}_{f,k} \tag{6.3}$$

$$\hat{\mathbf{W}}_{f,k} \leftarrow \frac{\tilde{\mathbf{W}}_{f,k}}{\sum_{f} \tilde{\mathbf{W}}_{f,k}} \tag{6.4}$$

$$\overline{\mathbf{v}}^{\Phi_i} \leftarrow \frac{\sum_{k \in \Phi_i} \nu_k \mathbf{v}^{(k)}}{\sum_{k \in \Phi_i} \sum_{m=1}^{M} \nu_k v^{(k)}_m} \tag{6.5}$$

$$\mathbf{w}^{(l+1)}_{k} \leftarrow \overline{\mathbf{v}}^{\Phi_i} \odot \mathbf{e}^{(k)} \tag{6.6}$$

where $\overline{\mathbf{v}}^{\Phi_i}$ is the average spectral envelope for instrument i, ν_k is the weight of $\mathbf{v}^{(k)}$ for the weighted mean. ν_k is heuristically determined as $\{\|\mathbf{h}_k\|_1\}^5$, which assumes the bases of which spectral envelopes are close to the actual envelope have larger time activity.

Note that the Eq. 6.1 to 6.4 are identical to Eq. 5.9 to 5.12. The true spectral envelope extracted from the given instrument sound is replaced by the group-wise average envelope $\overline{\mathbf{v}}^{\Phi_i}$. Because the proposed method is an unsupervised separation method, the ground truth spectral envelopes are not assumed to be given in advance. We have made the bases in a group have a unified envelope and the envelope converge to the ground truth envelope. Technically, there has been

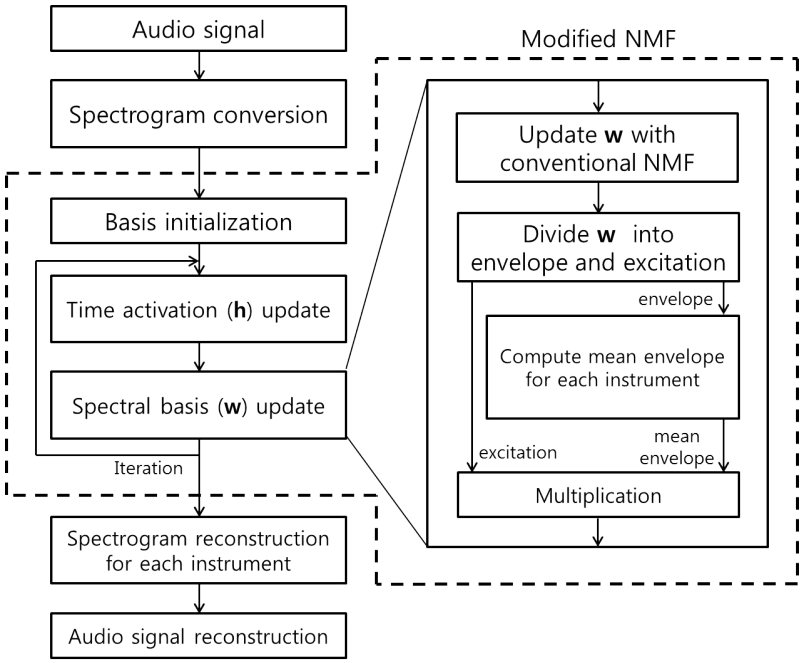


Fig. 6.1 Overview of the proposed method.

no technique that only estimates the envelopes of the mixed instruments, hence, we have utilized the NMF algorithm that simultaneously learns the envelope and the excitation. Enforcing the spectral envelopes in a group to be identical can be interpreted as the imposition of Dirichlet prior, with the hyperparameter and its weight in Eq. 2.35 determined as

$$\xi_k = \overline{\mathbf{v}}^{\Phi_i} \odot \mathbf{e}^{(k)} \tag{6.7}$$

$$w_{freq} = 0 \tag{6.8}$$

where ξ_k denotes the k-th column of $\boldsymbol{\Xi}$.

Finally, the spectrograms of each instrument can be reconstructed in the

same way we did in the previous chapter as

$$\hat{\mathbf{X}}_i = \sum_{k \in \Phi_i} \mathbf{w}_k \mathbf{h}_k, \tag{6.9}$$

followed by the inverse short-time Fourier transform.

6.3 Performance evaluation

In this section, we compare the performance of the proposed method in the same environment that we have tested the our informed approach. Before we present the SDR performances, we first optimize ν_k in a heuristic manner.

6.3.1 Weight optimization

In this subsection, the weight ν_k in Eq. 6.5 is optimized using the first piece in the dataset; *AchGottundHerr*. We analyzed the case where weight is a function of mean activation $\|\mathbf{h}_k\|_1$. The assumption is that the bases of which spectral envelopes are close to the actual ones will have larger time activity in average.

Fig. 6.2 shows the performance transitions of the average SDR with the increase of exponent p. Here, it is assumed that $\nu_k = \{\|\mathbf{h}_k\|_1\}^p$. The envelopes are averaged with equal weights when $p = 0$. It can be observed that the average SDR is maximized when $p = 5$. We use this value as the weight in the rest of the experiments.

6.3.2 Performance comparison

Table 6.1 shows the performance metrics averaged over the 10 pieces and 6 cases of instrument combinations. It can be seen that the Klapuri *et al.*'s method

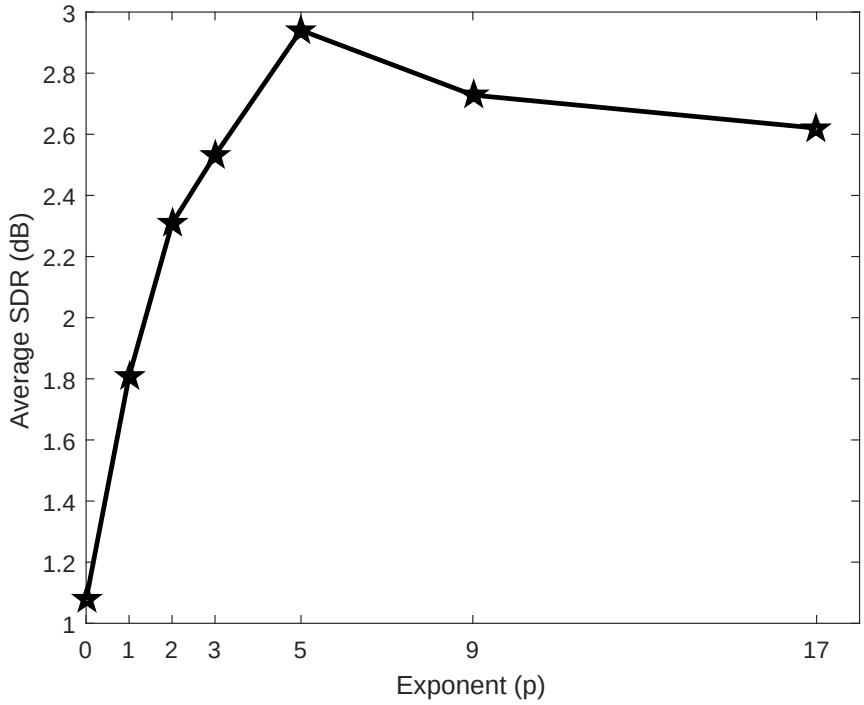


Fig. 6.2 Average SDR value with the increase of exponent p.

shows the highest SAR whereas the proposed method shows the lowest SAR. Still, however, 7.77dB is considered as a high value in general, and artifacts are hardly noticeable in the reconstructed audio. When the performances of the informed approach that were already presented in the section 5 are excluded, the proposed method shows the highest SIR value with a difference of more than 4dB from the second-highest. This difference is quite noticeable and it also causes the difference in SDR. Klapuri *et al.*'s method shows better performance as it uses the fundamental frequency and initializes the excitation prior to the expectation-maximization algorithm. However, it still shows lower performance compared to the proposed method.

The bottom two results show the performance of the proposed method when pure sounds of the instruments are added in front of the mixture. The proposed method was applied to test how the performance of the proposed method changes when the solo part of a musical instrument is partially present. The numbers in parentheses indicate the duration of each instrument's solo part added in front of the mixture. Note that the solo parts are the randomly selected music segments in the dataset. It can be seen that the SAR remains constant, whereas SIR increases as the length of the solo parts increases. This also leads to the increase of the SDR. From this observation, we can claim that the performance of the proposed method increases if solo parts exist in the mixture.

When compared with the informed approach, it can be guessed that the proposed blind approach's performance has been degraded because of the lack of side-information. However, it is difficult to distinguish whether it is due to the existence of solo part or the effect of hand labeling. To answer to this curiosity, we compare the performance of the informed approach (with Bach 10) to the blind approach with the concatenated 5 second solo parts. Then it can be observed that the informed approach can be considered equal to the case where the solo parts of the blind approach are labled by a human. It is interesting in that the SDR performances of the two (the fourth and the last) approaches make a gap of 0.82dB. This difference can be interpreted as the influence of labeling.

On the other hand, the effect of the presence of the solo part can be investigated via comparing two blind approaches: the one without the solo part and the one with the 5 second solo part. It can be seen that the presence of the

Table 6.1 Performances measured with the Bach 10 dataset (dB).

	SDR	SIR	SAR
FASST (stft)	0.50	1.80	9.17
FASST (erb)	0.18	1.44	8.95
Klapuri	1.37	1.79	**15.29**
Informed approach (Bach 10)	4.00	7.01	9.80
Informed approach (RWC)	**4.50**	**7.22**	10.60
Blind approach	2.57	5.91	7.77
Blind approach (+1s)	2.75	6.24	7.77
Blind approach (+5s)	3.18	7.08	7.86

solo part generates a SDR gap of 0.61dB. Also, it is interesting to observe that the presence of solo parts mainly improves SIR whereas the labeling of the solo part improves SAR.

Fig. 6.3 shows the average SDR of each case. FASST shows the lowest SDR when separating bassoon and violin sounds. Meanwhile, Klapuri *et al.*'s method and the proposed method work poorly when separating saxophone and violin, whereas they work best in clarinet-violin separation. It can be also seen that the proposed method shows the highest SDR for all combinations.

6.3.3 Effect of envelope similarity

In this subsection, we investigate the possibility to separate the instrument sounds with little difference in the spectral envelope. To this end, we have used RWC music instrument database as in the chapter 5. To generate mixtures, we

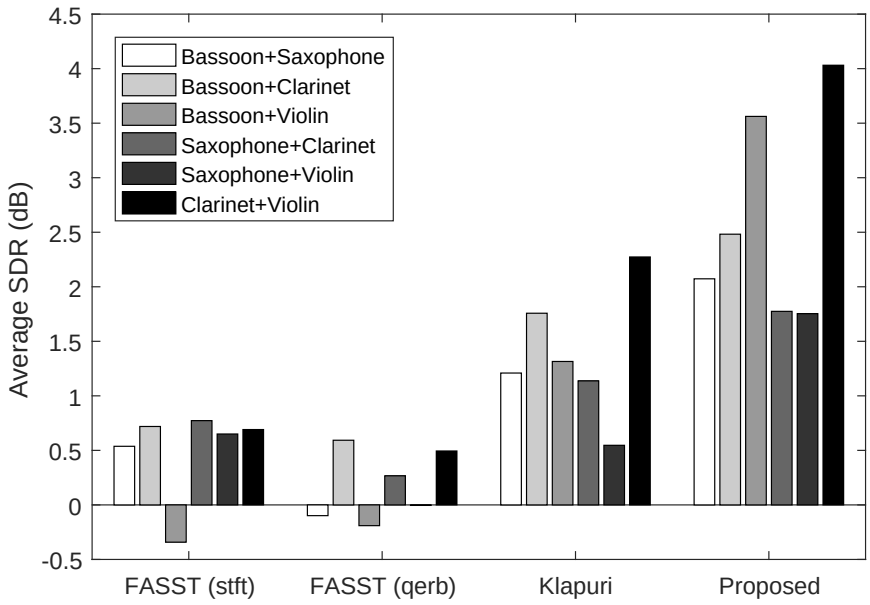


Fig. 6.3 Average SDR values with varying instrument combinations.

have cut an audio clip to multiple segments, each of which contains the sound of a single note. Then we have mixed the sounds of two instruments to have 10-second duration with 20 notes (10 notes per an instrument) placed at random positions.

Table 6.2 shows the separation performances. It can be seen that the SDR is the lowest when we separate the mixture of violin and viola sounds. This may be due to the fact that the spectral envelopes of violin and viola are similar as can be seen from their shapes. However, cello has much bigger size compared to the two, it is relatively easy to separate cello sound from the others. Especially, the performance of separating cello sound from violin sound is the highest because their envelopes are least similar. Thus, we can confirm that the envelope shape

Table 6.2 Performance of separating instruments with similar envelope.

	SDR	SIR	SAR
Violin+Viola	-0.08	1.40	8.62
Violin+Cello	2.69	4.78	9.09
Viola+Cello	2.01	3.92	9.22

affects critically to the separation performance.

6.4 Summary

In this chapter, we proposed a novel blind approach to separate harmonic instrument sounds by applying envelope constraints to spectral bases in NMF. Conventional research efforts regarding instrument sound separation often used side-information such as musical score and pitch. By extending the spectral envelope constraint-based informed approach presented in the previous chapter, the proposed method successfully minimized the required information by constraining the spectral bases, which were accurately obtained via LPC. Also, the proposed method outperformed the conventional methods in the comparative evaluation with real recordings.

Chapter 7

Conclusion and Future Work

7.1 Contributions

The final purpose of this thesis was to present the possibility of reducing the side-information in performing the source separation. This was enabled by adopting the Dirichlet prior which is a powerful method especially to impose spectral and temporal characteristics in spectrogram decomposition framework. Even though it has been used in some conventional works, its application was limited to several tasks such as sparsity imposition. We have discovered its appropriateness to be applied for various tasks that require basis shaping and proved it through the studies presented in this thesis. On the basis of the theoretical backgrounds (Chapter 2), we have applied the generalized Dirichlet prior to the harmonic-percussive sound separation task in order to impose sparsity and harmonicity characteristics (Chapter 3) or continuity and discontinuity characteristics (Chapter 4). Also, focusing on the spectral envelope which is

shared among the sounds generated by an instrument, we exploited envelope constraints on the basis vectors for harmonic instrument sound separation. Considering the difficulty of the task, we first presented informed approach which assumes the sounds of each instrument are obtainable (Chapter 5), followed by its extension to the blind case where only the number of instruments are known (Chapter 6).

The major contributions of this thesis can be summarized as follows:

- **Generalization of Dirichlet prior to linear algebraic framework**: We have extended the concept of Dirichlet prior to the NMF framework. Even though it had been applied to matrix decomposition techniques in the probabilistic framework, computational complexity was one of the bottlenecks that interfere its broader applications. Focusing on the effect of the Dirichlet prior on the iterative update equations of PLCA algorithm, we could generalize it to the NMF framework. By doing so, we could dramatically reduce the amount of required computations.

- **Harmonic-percussive sound separation based on their spectral characteristics**: We exploited the spectral characteristics of the harmonic and percussive sounds to separate them through spectrogram decomposition. Since conventional methods assumed that harmonic components are sustained for a certain amount of time, they missed to separate time-varying harmonic components such as vibrato, glissando, and human voice. Our method overcame this problem by focusing on the spectral aspects of the sounds.

- **Harmonicity constraint**: Sparsity (or sparseness, equivalently) con-

straint has been widely studied in the conventional studies to shape the bases to have sparse structure. However, none of them have succeeded in making the bases to have harmonic structure. We found that we can enforce a spectrum to have harmonic structure if we can make its spectrum (i.e., *spectrum of spectrum*) sparse. The harmonicity constraint is expected to be widely applied to tasks that require to separate harmonically-distributed components.

- **Harmonic-percussive sound separation based on continuity of basis vectors**: Spectral/temporal sparseness has been the major assumptions in HPSS research in conjunction with continuity. Our further investigation on harmonic and percussive spectra revealed that discontinuity rather than sparsity is more appropriate to distinguish the harcmonic and percussive spectra. Based on this observation, we presented a novel spectrogram decomposition algorithm that controls the degree of continuity of the spectral and temporal bases. Comparative evaluations with the conventional methods showed the outstanding performance of this method.

- **Harmonic instrument sound separation based on spectral envelope constraint**: Spectral envelope is one of the most important features that characterize the timbre of a sound. Accordingly, conventional works tried to utilize it in their source separation framework. However, none of them successfully utilized the spectral envelope because it did not seem to harmonize well with the spectrogram decomposition framework. We have overcome this problem by presenting a simple way to calculate the envelope that uses linear prediction. For the first step, we verified the validity

of our method by exploiting the pre-trained envelopes of the instruments. Then we applied the same idea to the blind scenario that makes the envelope and the excitation of bases separately converge. The experimental results revealed that the proposed HISS method works better than the conventional methods.

Despite our contributions to the field of source separation research, our works have limitations for the following parts. First, this study only considers single-channel case. Even though the single-channel source separation is one of the most important ground research for the multi-channel source separation, how to jointly manage the spatial information has to be furtherly studied. Second, application of our approach to vocal sound separation is necessary. Since the vocal components do not show consistent spectro-temporal characteristics because of the pronunciation that changes continuously, further study is expected to be necessary. Finally, based on the inspiration that we have obtained about the spectro-temporal characteristics, it is necessary to extend our research to the general audio source separation tasks such as audio event separation. The need for detection and classification of audio scenes and events is rapidly rising in conjunction with the increasing attention to the artificial intelligence. In the next section, we suggest where to start further research with some important points regarding the aforementioned problems.

7.2 Future work

7.2.1 Application to multi-channel audio environment

Our research shows the new possibility of utilizing the spectro-temporal characteristics in the single-channel environment. In addition to the works, joint consideration of spatial information is necessary to further improve the performance. Indeed, many music source separation algorithms are known to successfully improve the performance by effectively utilize spatial information in addition to the single-channel source separation algorithms. In our case, we can directly extend the proposed methods to factorize the tensor in which the spectrogram is three-dimensionally stacked, by replacing the NMF with the non-negative tensor factorization (NTF).

However, as our methods assume that each instrument has a common spectro-temporal characteristic, post-processings like mastering can affect the performance in a negative way. This problem is expected to be severe with the HISS methods presented in chapter 5 and chapter 6. When the microphones to record the sounds are separately installed as in the case of analyzing multiple youtube videos of a concert [96, 97, 98], the similar problem can happen because they all go through the different filters. Compensation of these distortions and effects is expected to be necessary.

7.2.2 Application to vocal separation

Human voice often contains the important information especially about the melody. However, it shows intermediate characteristics in both spectral and temporal side [99, 100], and does not show consistent spectral envelopes. Thus,

it is expected that the characteristics containing meaningful musical information such as MFCC can be alternatively used. However, the feature has to be inversed to the spectral domain in order to be able to directly apply it to the Dirichlet prior framework.

7.2.3 Application to various audio source separation tasks

Based on the knowledge gained from the instrument sound separation, we can extend the research to more complicated sounds such as audio events that show various spectro-temporal characteristics. Moreover, since audio events occur diverse and occasionally, it is impossible to pre-train all signals. Nevertheless, analyzing such events can give meaningful information to people because they often contain crucial information about the surroundings. Especially, emergency-related sound events like glass breaking sound and gun shot sound are even more important [101]. The HPSS method presented in the chapter 4 has been contributed to achieving rank 2 in the task1 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 [102]. As our approaches have assumed, focusing on the spectro-temporal characteristics can be a fundamental basis for the further research since such characteristics are maintained even for sound events.

Bibliography

[1] T. Adali, C. Jutten, A. Yeredor, A. Cichocki, and E. Moreau, "Source separation and applications," *IEEE Signal Processing Magazine*, 2014.

[2] A. Delorme and S. Makeig, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[3] C. W. Hesse and C. J. James, "On semi-blind source separation using spatial constraints with applications in eeg analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2525–2534, 2006.

[4] R. R. Vázquez, H. Velez-Perez, R. Ranta, V. L. Dorr, D. Maquin, and L. Maillard, "Blind source separation, wavelet denoising and discriminant analysis for eeg artefacts and noise cancelling," *Biomedical Signal Processing and Control*, vol. 7, no. 4, pp. 389–400, 2012.

[5] L. T. Duarte, S. Moussaoui, and C. Jutten, "Source separation in chemical analysis: Recent achievements and perspectives," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 135–146, 2014.

[6] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[7] N. J. Bryan, "Interactive sound source separation," Ph.D. dissertation, Stanford University, 2014.

[8] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Applied Sciences*, vol. 6, no. 2, p. 57, 2016.

[9] Y. Li, S.-I. Amari, A. Cichocki, D. W. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 2, pp. 423–437, 2006.

[10] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "Convolutive blind source separation methods," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 1065–1094.

[11] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[12] R. Aichner, H. Buchner, M. Zourub, and W. Kellermann, "Multi-channel source separation preserving spatial information," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1. IEEE, 2007, pp. I–5.

[13] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE*

transactions on audio, speech, and language processing, vol. 15, no. 3, pp. 1066–1074, 2007.

[14] M. E. Davies and C. J. James, "Source separation using single channel ica," *Signal Processing*, vol. 87, no. 8, pp. 1819–1832, 2007.

[15] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Transactions on signal processing*, vol. 45, no. 2, pp. 434–444, 1997.

[16] V. Zarzoso and A. Nandi, "Blind source separation," in *Blind Estimation Using Higher-Order Statistics*. Springer, 1999, pp. 167–252.

[17] G. R. Naik, W. Wang *et al.*, *Blind source separation*. Springer, 2014.

[18] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[19] S. Ewert and M. Muller, "Using score-informed constraints for nmf-based source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 129–132.

[20] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 888–891.

[21] J. Park and K. Lee, "Separation of monophonic music signal based on user-guided onset information," *International Congress on Sound and Vibration*, 2014.

[22] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected f0 track." in *LVA/ICA*. Springer, 2012, pp. 438–445.

[23] A. Lefevre, F. Bach, and C. Févotte, "Semi-supervised nmf with time-frequency annotations for single-channel source separation," in *ISMIR 2012: 13th International Society for Music Information Retrieval Conference*, 2012.

[24] N. J. Bryan, G. J. Mysore, and G. Wang, "Isse: an interactive source separation editor," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 257–266.

[25] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*. IEEE, 2009, pp. 69–72.

[26] M. Parvaix, L. Girin, and J.-M. Brossier, "A watermarking-based method for informed source separation of audio signals with a single sensor," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 6, pp. 1464–1475, 2010.

[27] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation: source coding meets source separation," in *Applications of Signal*

Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on. IEEE, 2011, pp. 257–260.

[28] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[29] W. Lim and T. Lee, "Harmonic and percussive source separation using a convolutional auto encoder," in *European Signal Processing Conference*, 2017.

[30] K. Osako, Y. Mitsufuji, R. Singh, and B. Raj, "Supervised monaural source separation based on autoencoders," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on.* IEEE, 2017, pp. 11–15.

[31] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," *arXiv preprint arXiv:1703.08019*, 2017.

[32] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "Nmf-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, 2015.

[33] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2136–2147, 2015.

[34] Y. Han, J. Kim, K. Lee, Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 1, pp. 208–221, 2017.

[35] 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*, 2011, pp. 23–24.

[36] A. Diment, P. Rajan, T. Heittola, and T. Virtanen, "Modified group delay feature for musical instrument recognition," in *International Symposium on Computer Music Multidisciplinary Research*, 2013.

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

[38] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[39] M. Kim and P. Smaragdis, "Manifold preserving hierarchical topic models for quantization and approximation," in *International Conference on Machine Learning*, 2013, pp. 1373–1381.

[40] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 1, pp. 70–79, 2007.

[41] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational intelligence and neuroscience*, vol. 2008, 2008.

[42] A. Cichocki, R. Zdunek, and S.-i. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.

[43] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative nmf and its application to single-channel source separation." in *INTERSPEECH*, 2014, pp. 865–869.

[44] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[45] J. M. Becker, C. Sohn, and C. Rohlfing, "Nmf with spectral and temporal continuity criteria for monaural sound source separation," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 316–320.

[46] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *Digital*

Signal Processing (DSP), 2011 17th International Conference on. IEEE, 2011, pp. 1–6.

[47] C. Ding, T. Li, and W. Peng, "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing," *Computational Statistics & Data Analysis*, vol. 52, no. 8, pp. 3913–3927, 2008.

[48] F. J. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, and P. Cabanas-Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 26, 2014.

[49] M. Shashanka, "Latent variable framework for modeling and separating single-channel acoustic sources," *Department of Cognitive and Neural Systems, Boston University*, 2007.

[50] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.

[51] J. Park and K. Lee, "Harmonic-percussive source separation using harmonicity and sparsity constraints." in *ISMIR*, 2015, pp. 148–154.

[52] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *Proc. ICA*, 2003, pp. 843–847.

[53] O. Gillet and G. Richard, "Drum track transcription of polyphonic music using noise subspace projection." in *ISMIR*, 2005, pp. 92–99.

[54] M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Signal Processing Conference, 2005 13th European*. IEEE, 2005, pp. 1–4.

[55] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[56] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Signal Processing Conference, 2008 16th European*. IEEE, 2008, pp. 1–4.

[57] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals." in *ISMIR*, 2008, pp. 139–144.

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

[59] A. Gkiokas, V. Papavassiliou, V. Katsouros, and G. Carayannis, "Deploying nonlinear image filters to spectrogram for harmonic/percussive separation," in *Proceedings of the International Conference on Digital Audio Effects (DAFx), York, UK*, 2012, pp. 17–21.

[60] M. Vinyes, "MTG MASS database," http://www.mtg.upf.edu/static/mass/resources, 2008.

[61] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[62] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 141–145.

[63] J. Park, J. Shin, and K. Lee, "Exploiting continuity/discontinuity of basis vectors in spectrogram decomposition for harmonic-percussive sound separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1061–1074, 2017.

[64] A. Elowsson and A. Friberg, "Modeling the perception of tempo," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3163–3177, 2015.

[65] M. Tian and M. B. Sandler, "Towards music structural segmentation across genres: Features, structural hypotheses, and annotation principles," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 2, p. 23, 2016.

[66] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama, "Hmm-based approach for automatic chord detection using refined acoustic features," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* IEEE, 2010, pp. 5518–5521.

[67] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE Transactions on audio, speech, and language processing*, vol. 21, no. 10, pp. 2096–2107, 2013.

[68] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in *Acoustics speech and signal processing (icassp), 2010 ieee international conference on.* IEEE, 2010, pp. 425–428.

[69] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 1, pp. 228–237, 2014.

[70] J. Driedger, M. Muller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2014.

[71] W. Buyens, B. van Dijk, J. Wouters, and M. Moonen, "A harmonic/percussive sound separation based music pre-processing scheme for cochlear implant users," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European.* IEEE, 2013, pp. 1–5.

[72] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 529–540, 2008.

[73] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative evaluations of various harmonic/percussive sound separation algorithms based on anisotropic continuity of spectrogram," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 465–468.

[74] B. Thoshkahna and R. R. Kalpathi, "A postprocessing technique for improved harmonic/percussion separation for polyphonic music." in *ISMIR*, 2011, pp. 251–256.

[75] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals." in *ISMIR*, 2014, pp. 611–616.

[76] D. FitzGerald, A. Liukus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/percussive separation using kernel additive modelling," *Irish Signals and Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies*, 2014.

[77] N. Q. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 205–208.

[78] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[79] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Advances in neural information processing systems*, 2009, pp. 1705–1713.

[80] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 387–395.

[81] "QUASI database: A musical audio signal database for source separation,"
http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi.

[82] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[83] 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.

[84] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proceedings of the 12th International Conference on Digital Audio Effects*, 2009.

[85] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, vol. 2008, 2008.

[86] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation." in *ISMIR*, 2009, pp. 327–332.

[87] F. Rodriguez-Serrano, J. Carabias-Orti, P. Vera-Candeas, T. Virtanen, and N. Ruiz-Reyes, "Multiple instrument mixtures source separation evaluation using instrument-dependent nmf models," *Latent Variable Analysis and Signal Separation*, pp. 380–387, 2012.

[88] A. Klapuri, T. Virtanen, and T. Heittola, "Sound source separation in monaural music signals using excitation-filter model and em algorithm," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5510–5513.

[89] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[90] Y. Salaün, E. Vincent, N. Bertin, N. Souviraa-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot, "The flexible audio source separation toolbox version 2.0," in *ICASSP*, 2014.

[91] 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.

[92] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[93] S. V. Vaseghi, *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

[94] 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.

[95] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Music genre database and musical instrument sound database," 2003.

[96] Y. Uehara, T. Kawamura, S. Egami, Y. Sei, Y. Tahara, and A. Ohsuga, "Linked data collection and analysis platform for music information retrieval," in *Joint International Semantic Technology Conference*. Springer, 2016, pp. 127–135.

[97] M. Airoldi, D. Beraldo, and A. Gandini, "Follow the algorithm: An exploratory investigation of music on youtube," *Poetics*, vol. 57, pp. 1–13, 2016.

[98] A. Bagri, F. Thudor, A. Ozerov, and P. Hellier, "A scalable framework for joint clustering and synchronizing multi-camera videos," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013, pp. 1–5.

[99] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," *ISAST Transactions on Electronic and Signal Processing*, vol. 4, no. 1, pp. 62–73, 2010.

[100] I.-Y. Jeong and K. Lee, "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1197–1200, 2014.

[101] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1d convolutional recurrent neural networks," *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[102] Y. Han and J. Park, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

초 록

음악 신호의 음원 분리는 개별 음원의 소리들을 추출해내고 재구성하는 것을 목표로 한다. 최근 음원 분리는 오디오 신호 처리 분야에서의 중요성과 영향력으로 인해 많은 관심을 받아왔다. 잡음 제거나 악기 별 이퀄라이징과 같은 적용분야와 더불어, 음원 분리는 전처리로 사용되었을 경우에 다양한 음악 정보 분석 알고리즘들의 성능에도 직접적인 영향을 미칠 수 있다. 그러나 현재까지의 음원 분리 알고리즘은 만족스러운 성능을 보여주지 못하고 있으며, 이러한 현상은 음원에 대한 공간적인 또는 음악적인 정보가 주어지지 않았을 경우 더욱 심화된다. 우리는 음원에 대한 정보가 주어지지 않은 블라인드 환경에서 스펙트로그램에 표현되는 주파수 축과 시간 축 특성을 활용하였다. 스펙트로그램 분해 알고리즘은 시간/주파수 특성을 활용하기 적합하여 널리 쓰이지만, 그 과정에 제약 조건을 주는 것은 몇몇의 특수한 특성들에 대해서만 가능하였다. 본 논문의 주요 목표는 스펙트로그램 분해 알고리즘의 기저들을 제약하기 위한 방법으로서의 일반화된 디리클레 사전확률의 가능성을 살펴보는 것이다. 우리는 화성악기와 타악기 소리의 분리부터 화성악기들 간의 음원 분리까지 다양한 과업에 일반화된 디리클레 사전확률을 적용하였으며, 디리클레 사전확률의 유연한 응용 가능성과 함께 높은 수준의 성능까지 확인할 수 있었다.

주요어: 음원 분리, 비음수 행렬 분해, 확률적 은닉 성분 분석, 디리클레 사전 확률
학 번: 2012-31246