A Study on Utilization of Prior Knowledge in Underdetermined Source Separation and Its Application

Shogo Seki

Contents

Abstract

In the field of environmental sound recognition, source separation is one of core technologies, used to extract individual sound sources from mixed signals. Source separation is closely related to other acoustic technologies and is used to develop various applications such as automatic transcription systems for meetings, active music listening systems, and music arranging systems for composers. When a mixed signal is composed of more sources than the number of microphones, i.e., in an underdetermined source separation scenario, separation performance is still limited and there remains much room for improvement. Moreover, depending on the method used to extract the source signals, subsequent systems using the acoustic features calculated from the estimated source information can suffer from performance degradation. Supervised learning is a promising method which can be used to alleviate these problems. Training data composed of source signals, as well as mixed signals, is used to obtain as much prior information about the sound sources as possible into account. Supervised learning is essential for improving the performance of underdetermined source separation, however there are problems which remain to be addressed.

In this dissertation, I address two problems with the supervised learning approach for underdetermined source separation and its application. The first is how to improve the use of prior information, and the second is how to improve the representation ability of source models. To deal with the first problem, I focus on, 1) the characteristics of individual source signals in the spectral and feature domains, and 2) the temporal characteristics implicitly considered in time-frequency analysis. Furthermore, I also explore the use of deep generative models for prior information, to deal with the second problem.

Since synthesized music signals are often stereophonic signals and are generated as linear combinations of many individual source signals and their respective mixing gains, information about phase, or its differential, between each channel, which represents the spatial characteristics of recording environments, cannot be utilized as acoustic clues for source separation. In order to address this problem, this dissertation proposes a supervised source separation method for stereophonic music signals based on an extension of non-negative matrix factorization (NMF). NMF-based decomposition is applied to approximate the amplitude spectrogram of a music signal as linear combinations of mixing gains and the spectrograms of individual sources, in which source spectrograms are further decomposed into a set of spectral templates and respective activations. In addition to the conventional supervised approach, cepstral distance regularization (CDR) is further introduced to regularize the timbre information of each source. Experimental evaluations demonstrate that CDR yields significant performance improvements and provides better estimation for mixing gains.

While time-frequency masking is a powerful approach for source separation and speech enhancement in terms of signal recovery accuracy, e.g., signal-to-noise ratio, it can over-suppress and damage speech components, leading to limited performance in succeeding speech processing systems. To overcome this problem, this dissertation proposes a method of restoring missing components of time-frequency masked speech spectrograms using direct estimation of a time domain signal based on time-domain spectrogram factorization (TSF). This TSF-based method allows us to take the local interdependencies of the components of a complex spectrogram, derived from the redundancy of a time-frequency representation, into account, as well as the global structure of the magnitude spectrogram. Experimental results show that the proposed TSF-based method significantly out-performs conventional methods, and has the potential to estimate both phase and magnitude spectra simultaneously and precisely.

Multichannel non-negative matrix factorization (MNMF) is a well-known method used for underdetermined audio source separation, which adopts the NMF concept to model and estimate the power spectrograms of the sound sources in a mixed signal. While MNMF works reasonably well for particular types of sound sources, one limitation is that it can fail to work for sources with spectrograms that do not comply with NMF. In contrast to underdetermined cases, an improved variant of determined source separation methods, called the multichannel variational autoencoder (MVAE) method, was recently proposed, in which a conditional VAE (CVAE) is used instead of the NMF model for expressing source power spectrograms. While the original MVAE method was formulated for use in determined mixing scenarios, we propose a generalized version, combining the features of MNMF and MVAE so that it can also be used for underdetermined source separation. We call this method the generalized MVAE (GMVAE) method. Experimental evaluations reveal that GMVAE outperformed baseline methods, including MNMF.

1 Introduction

1.1 Background

Human beings can perceive where sound signals come from, discriminate and identify various sounds, and recognize related events or meanings. A number of studies have attempted to replicate these complicated functions of humans using technology, as part of a field of study known as computational auditory scene analysis (CASA) [1, 2]. Recently, researchers have been focusing on environmental sounds, as well as on speech and music.

Source separation [3] is one of the most important CASA technologies, which allows the separation of individual source signals from a mixture of signals. Although humans can easily pick out various sounds from a noisy background, such as birds chirping, people talking, traffic noises, rain falling, etc. this is a very difficult task to automate. Source separation is used in conjunction with other technologies to develop various applications. By combining source separation with automatic speech recognition (ASR) [4], we can develop systems that can automatically identify different speakers in a meeting and transcribe each person's utterances. By combining source separation with binaural techniques [5], it is possible to develop new sound systems that allow music listeners to adjust a wide variety of variables in order to achieve their favorite allocation of sound sources, i.e., active music listening systems. Using source separation with voice conversion (VC) [6], could potentially allow composers to extracting the vocal components of a music signal and convert them by giving them different or additional attributes.

For decades, source separation has been studied and developed under blind conditions, i.e., blind source separation (BSS) [3, 7, 8], where no information about the number or location of source signals, or the mixing process, are given. However, if we can use the same number of microphones as the number of source signals, or a larger number of microphones, which are called determined and overdetermined source separation, respectively, impressive source separation performance can be achieved [9, 10, 11, 12, 13]. In contrast, when a mixture of signals is composed of more sources than the number of microphones, which is called underdetermined source separation [14, 15], separation performance is still limited, so there is room for improvement.

Supervised learning [16, 17], which uses training data composed of the source signals contained within the mixed signals, is a promising way to alleviate this problem, by taking as much information about the sources, i.e., prior information, as possible into account. The source estimation method which is used is important because subsequent processing can be impaired if the source estimation data is inaccurate. Time-frequency masking [18] is one example of such a source estimation approach. Since non-target components are over-suppressed and target components remain only sparsely, the acoustic features calculated from the masked sources degrade during subsequent processing.

1.2 Thesis Scope

This thesis addresses two problems which are encountered when using a supervised approach for underdetermined source separation. One is how to utilize prior information more efficiently, and the other is how to improve the representation ability of a

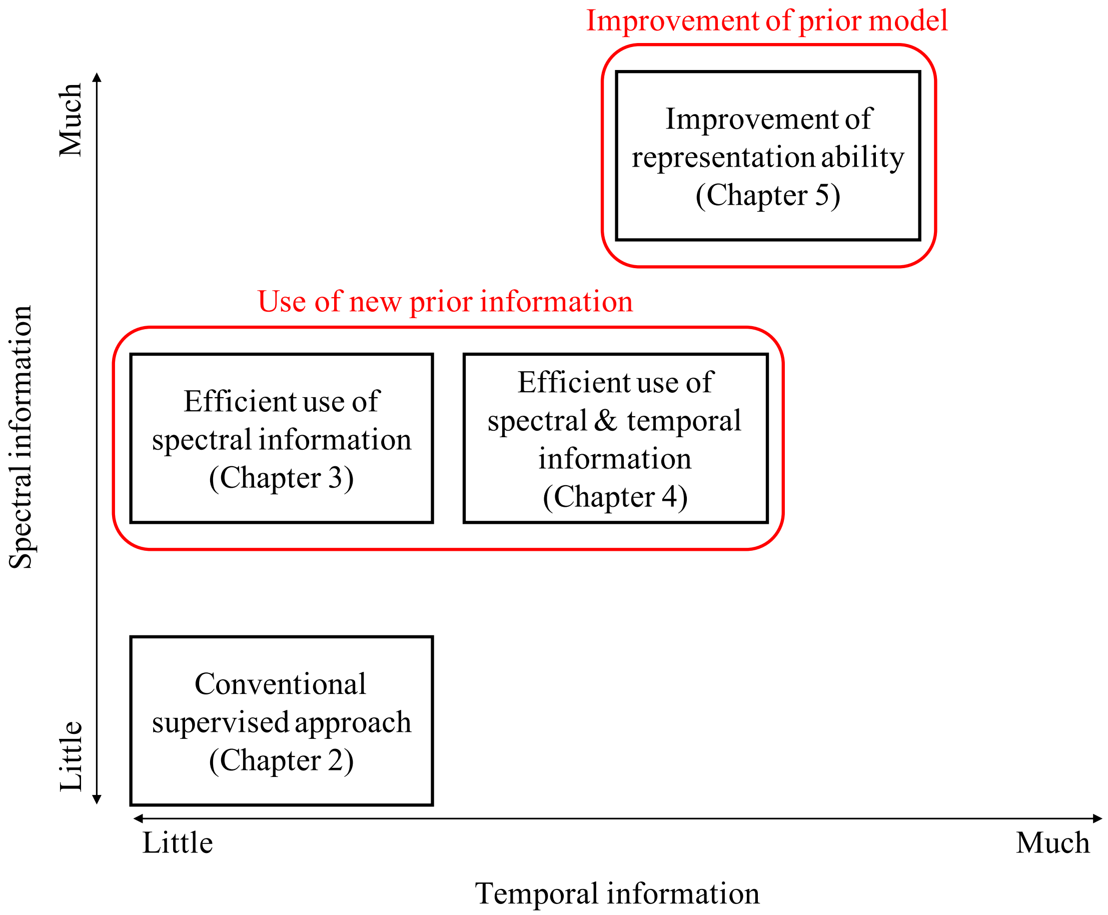

Improvement of prior model
Improvement of representation ability (Chapter 5)
Use of new prior information
Efficient use of spectral information (Chapter 3)
Efficient use of spectral & temporal information (Chapter 4)
Conventional supervised approach (Chapter 2)
Spectral information
Much
Little
Temporal information
Little
Much

Figure 1.1: Problems addressed in this thesis.

model when using prior information, as shown in Figure 1.1.

1.2.1 Improvement of the Use of Prior Information

Synthesized music signals, such as the music distributed on CDs or through online music websites are generally stereophonic signals composed of linear combinations of many individual source signals and their mixing gains, in which spatial information, i.e., phase information, or its differential between each channel, cannot be utilized as acoustic clues for source separation. To separate with these stereophonic music signals, this thesis employs the concept of non-negative matrix factorization (NMF) [19, 20] and propose a supervised source separation method based on non-negative tensor factorization (NTF) [21], a multi-dimensional extension of NMF. In order to reflect prior information of each source efficiently, this thesis further introduces the cepstral distance regularization (CDR) [22] method to consider and regularize the timbre information of the sources. Experimental results show that CDR yields significant improvement in separation performance and provides better estimation for mixing gains. Experimental results show that CDR yields significant performance improvements and provides better estimation for mixing gains.

After estimating the source signals in a mixed signal, time-frequency masking is a well-known approach used to extract source signals for source separation and speech enhancement [23]. While it is very effective in terms of signal recovery accuracy, e.g., signal-to-noise ratio, one drawback is that it can over-suppress and damage speech components, resulting in limited performance when used with succeeding speech processing systems. To overcome this flaw, this thesis proposes a method to restore the missing components of time-frequency masked speech spectrograms, which is based on direct estimation of a time-domain signal, referred to as time-domain spectrogram factorization (TSF) [24, 25]. TSF-based missing component restoration allows us to take into account the local inter-dependencies of the elements of complex spectrograms derived from the redundancy of a time-frequency representation, as well as the global structure of the magnitude spectrogram. Experimental results demonstrate that the proposed TSF-based method significantly outperforms conventional methods, and has the potential to estimate both phase and magnitude spectra simultaneously and precisely.

1.2.2 Improvement of Prior Model

When solving underdetermined source separation problems, multichannel non-negative matrix factorization (MNMF) [14, 15], a multichannel extension of NMF, adopts the NMF concept to model and estimate the power spectrograms of the sound sources in a mixed signal. Although MNMF works reasonably well for particular types of sound sources, it can fail to work for sources with spectrograms that do not comply with NMF, resulting in limited performance. However, a supervised source separation method called a multichannel variational autoencoder (MVAE) [26, 27], which is an improved variant of determined source separation methods, has been proposed, in which a conditioned variational autoencoder (VAE) [28], i.e., conditional VAE [29] is used instead of the NMF model for modeling source power spectrograms. This thesis proposes a generalized method of MVAE called the GMVAE method, which is constructed by combining the features of an MNMF and an MVAE so that it can also deal with underdetermined source separation problems. Experimental evaluations demonstrate that GMVAE outperforms baseline methods including MNMF.

1.3 Dissertation Overview

This dissertation is organized as follows: In Chapter 2, the basic framework of blind source separation (BSS) and its fundamental techniques are described. In Chapter 3, stereophonic music source separation using the timbre information of sources is described. In Chapter 4, missing component restoration by considering the redundancy of time-frequency representation is described. In Chapter 5, multichannel source separation based on a deep generative model is described. In Chapter 6, the relationship between source separation and real-world data circulation (RWDC) is mentioned. In Chapter 7, the contributions of this dissertation and future work are discussed.

2 Source Separation

Depending on the number of sources and microphones present, mixing processes, and types of signals, various source separation methods have been proposed. This chapter begins by categorizing source separation methods and describes fundamental components of separation algorithms. Source separation methods using supervised learning are also described and various methods of source separation are evaluated from a statistical point of view.k

2.1 Introduction

Source separation is a technology focusing on distinguishing individual sources of sound within a mixture of audio signals, replicating the human ability to identify various sounds in a noisy environment. This selective listening mechanism is called the cocktail-party effect [30, 31, 32], in reference to the human ability to focus on particular conversations going on simultaneously at the same party, a phenomenon which has been studied for many years. The most basic, but most difficult, source separation problems are those involving blind source separation (BSS) [3, 7, 8], in which an observed mixture of signals received by a microphone or microphone array are separated without any additional information about the component signals. Source separation problems are generally divided according to how many sources and microphones are present, whether additional training data is available, and the method used to solve these problems.

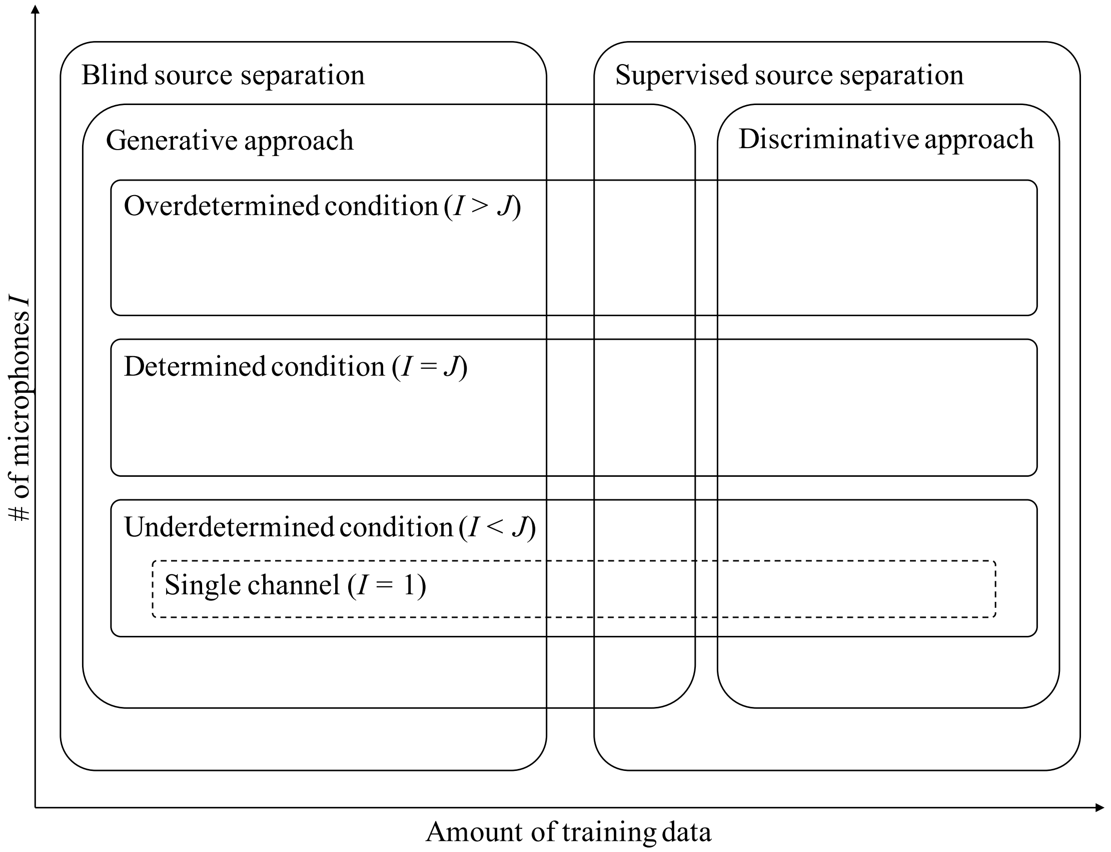


Figure 2.1: Categorization of source separation problems, where J represents the number of sources to be estimated and I represents the number of microphones.

Various source separation scenarios are categorized in Figure 2.1.

Early source separation methods assumed a determined condition, where the number of sources in a mixed signal are observed using the same number of microphones. Independent component analysis (ICA) [7, 33] is an appropriate method for performing determined source separations; source signals are separated by estimating a linear separation filter, under the assumption that the source signals present in the observed signal are statistically independent each other. ICA was initially applied in the time domain, assuming an instantaneous mixing where source signals arrive at the microphones without any time-delay or reverberation. Frequency domain ICA (FDICA) [34, 35] was then developed to resolve more realistic, convolutive mixing problems, in which source signals are mixed and observed with time-delay and reverberation. FDICA was further developed, and a natural extension of ICA called independent vector analysis (IVA) [9, 10, 12] was proposed, which solved the permutation problems which occurred when using FDICA. IVA was then further developed, and independent low-rank matrix analysis (ILRMA) [36] was then proposed. More recently, overdetermined source separation problems, in which the source signals are observed with a larger number of microphones, has been studied [36]. These methods estimate a linear separation filter to perform source separation, and separation performance tends to be sufficient.

In contrast, when considering more realistic situations, in which a mixture of signals is composed of more source signals than the number of microphones, i.e., in underdetermined conditions, non-negative matrix factorization (NMF) [19, 20] is a popular approach for handling source separation problems. NMF was originally proposed for single-channel source separation [37, 38, 39, 40]. After the development of several metrics used for optimization [41, 42], the relationship between NMF and generative models was discovered [43]. NMF was then extended into multichannel NMF (MNMF) [14, 15] to solve separation problems involving multi-channel signals. Although NMF and MNMF are applicable even for underdetermined source separation, separating the source signals requires the estimation of non-linear filters, e.g., multichannel Wiener filters [44]. Since these filters generally cause distortion, separation performance of underdetermined source separation methods tends to be insufficient.

Based on methods used for BSS, supervised source separation methods using supervised learning were then proposed [45, 46, 47]. By using the source signals contained in mixed signals as training data, supervised source separation methods attempt to take as much information about the sources (prior information) as possible into account. Moreover, thanks to the availability of increasing levels of computational power and the large amount of signal data now stored in databases [48, 49, 50, 51, 52, 53, 54, 55, 56], supervised source separation methods based on deep neural networks (DNNs) have recently been developed [57, 58, 59, 60, 61, 62], which can be categorized into generative and discriminative approaches. The former approach is regarded as an improved variant of conventional supervised source separation methods, in which the generative models for sources are represented as DNNs instead of NMFs. This allows us to use the flexible representation capacity of DNNs for source modeling. The latter is an approach in which DNNs are used to train models and infer target sources from a mixture of signals [26, 27, 63, 64]. Although discriminative approaches have shown adequate performance, large amounts of training data are required for the models to learn the mapping functions needed to separate signal mixtures into their component sources, and it is difficult to handle sources which do not appear in training data.

The rest of this chapter is organized as follows. Section 2.2 describes BSS problems and an efficient optimization technique called the Majorization-Minimization (MM) [65, 66] principle. Section 2.3 reviews supervised methods, including recent neural network-based methods. In Section 2.4, several performance evaluation criteria related to source separation are described. This chapter is then summarized in Section 2.5.

2.2 Blind Source Separation (BSS) in Underdetermined Conditions

2.2.1 General Formulation

Suppose that there are J source signals and that a mixed signal from these sound sources is captured by I microphones. We assume a convolutive mixing in the time domain, which is equivalent to instantaneous mixing in the frequency domain. Let $s_j(f,n)$ and $x_i(f,n)$ be the short-time Fourier transform (STFT) coefficient of the j-th source signal, and that of the i-th observed signal, respectively, where f and n are the frequency and time indices, respectively. We denote the vectors containing the STFT coefficients of all the sources, $s_1(f,n),\dots,s_J(f,n)$ and observed signals $x_1(f,n),\dots,x_I(f,n)$ as:

$$\mathbf{s}(f,n) = [s_1(f,n),\dots,s_J(f,n)]^\mathsf{T} \in \mathbb{C}^J, \tag{2.1}$$

$$\mathbf{x}(f,n) = [x_1(f,n),\dots,x_I(f,n)]^\mathsf{T} \in \mathbb{C}^I, \tag{2.2}$$

where $(\cdot)^\mathsf{T}$ represents the transpose and $\mathbb{C}$ denotes complex numbers. We assume that $s_j(f,n)$ independently follows a zero-mean complex Gaussian distribution with variance $v_j(f,n)$:

$$p(s_j(f,n)) = \mathcal{N}_\mathbb{C}(s_j(f,n)|0, v_j(f,n)), \tag{2.3}$$

where the complex Gaussian distribution $\mathcal{N}_\mathbb{C}(z|\mu,\sigma^2)$ for a complex random variable z with mean μ and variance σ^2 is defined as:

$$\mathcal{N}_\mathbb{C}(z|\mu,\sigma^2) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{(z-\mu)^2}{\sigma^2}\right). \tag{2.4}$$

(2.3) is usually called the local Gaussian model (LGM) [67, 68, 69]. When $s_j(f,n)$ and $s_{j'}(f,n)$ are mutually independent for $j \neq j'$, $\mathbf{s}(f,n)$ follows a complex Gaussian distribution:

$$p(\mathbf{s}(f,n)) = \mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|\mathbf{0}, \mathbf{V}(f,n)), \tag{2.5}$$

where $\mathbf{V}(f,n)$ is a diagonal covariance matrix with diagonal entries $v_1(f,n)$, ..., $v_J(f,n)$.

In an underdetermined condition, a mixing system is given as follows:

$$\mathbf{x}(f,n) = \mathbf{A}(f)\mathbf{s}(f,n), \tag{2.6}$$

which describes the relationship between $\mathbf{s}(f,n)$ and $\mathbf{x}(f,n)$, where $\mathbf{A}(f) = [\mathbf{a}_1(f), \ldots, \mathbf{a}_J(f)] \in \mathbb{C}^{I\times J}$ is referred to as a mixing matrix. From (2.5) and (2.6), $\mathbf{x}(f,n)$ is shown to follow:

$$p(\mathbf{x}(f,n)) = \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, \mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f)), \tag{2.7}$$

where $(\cdot)^{\mathsf{H}}$ represents the conjugate transpose. Thus, given an observed mixed signal $\mathcal{X} = \{\mathbf{x}(f,n)\}_{f,n}$, using the mixing matrices $\mathcal{A} = \{\mathbf{A}(f)\}_f$ and variance in source signals $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$, the log-likelihood is given as:

$$\begin{aligned}&\log p(\mathcal{X}|\mathcal{A},\mathcal{V})\\ =&\log\prod_{f,n} p(\mathbf{x}(f,n)|\mathbf{A}(f),\mathbf{V}(f,n))\\ \overset{c}{=}&-\textstyle\sum_{f,n}\left[\mathrm{tr}(\mathbf{x}(f,n)^{\mathsf{H}}(\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f))^{-1}\mathbf{x}(f,n)) + \mathrm{logdet}\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f)\right],\end{aligned} \tag{2.8}$$

where $\overset{c}{=}$ denotes the equality that holds when constant terms are ignored. If there is no constraint imposed on $v_j(f,n)$, (2.8) will be split into multiple frequency-wise source separation problems. This indicates that there is a permutation ambiguity in the separated components for each frequency, since permutation of j does not affect the value of the log-likelihood. Thus, permutation alignment is generally required after $\mathcal{A}$ is obtained.

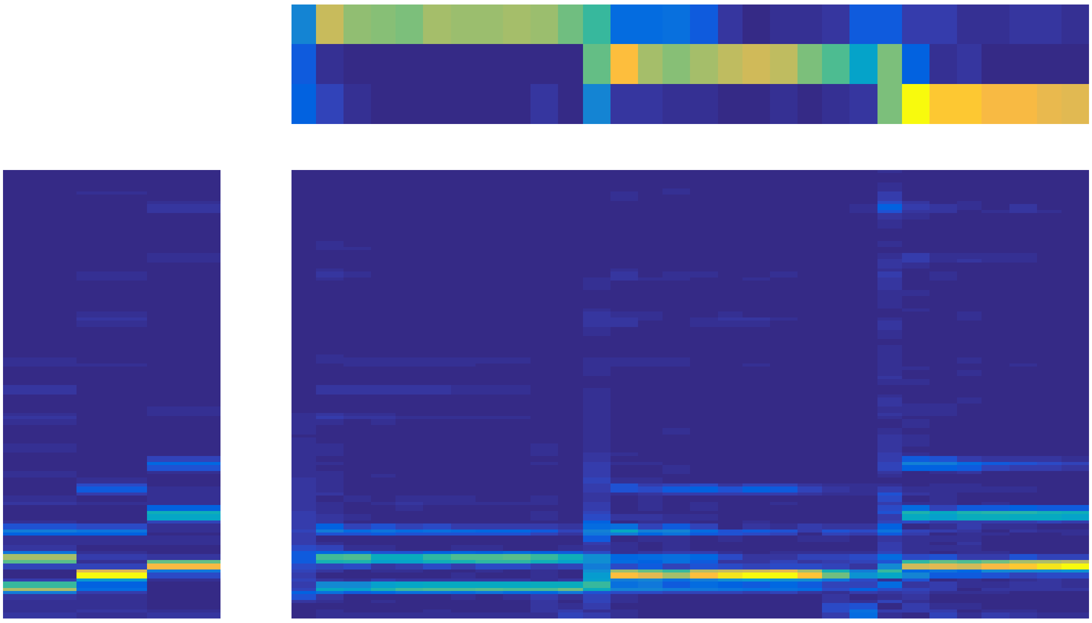


Figure 2.2: Visualization of NMF decomposition.

2.2.2 Non-negative Matrix Factorization (NMF)

When using a single-channel microphone ($I = 1$), $\mathbf{A}(f)$ is written as $\mathbf{A}(f) = \mathbf{1}^{1\times J}$ and (2.8) can be rewritten as follows:

$$\begin{aligned}&\log p(\mathcal{X}|\mathcal{A},\mathcal{V})\\ =&\log\prod_{f,n} p(x(f,n)|\mathbf{V}(f,n))\\ \overset{c}{=}&-\sum\nolimits_{f,n}\left[\frac{|x(f,n)|^2}{\sum_j v_j(f,n)}+\log\sum\nolimits_j v_j(f,n)\right],\end{aligned} \tag{2.9}$$

which is equivalent to NMF based on Itakura-Saito divergence (IS-NMF) [43].

As shown in Figure 2.2, NMF models $v_j(f,n)$ as the sum of K_j spectral templates $h_{j,1}(f)$, ..., $h_{j,K_j}(f) \geq 0$ scaled by time-varying activations $u_{j,1}(n),\ldots,u_{j,K_j}(n) \geq 0$:

$$v_j(f,n) = \sum\nolimits_{k=1}^{K_j} h_{j,k}(f)u_{j,k}(n). \tag{2.10}$$

It is also possible to share all the spectral templates of every source and let the contribution of the k-th spectral template to source j be determined in a data-driven manner. Thus, $v_j(f,n)$ can also be expressed as:

$$v_j(f,n) = \sum\nolimits_{k=1}^{K} b_{j,k} h_k(f) u_k(n), \tag{2.11}$$

where $b_{j,k} \in [0,1]$ is a continuous indicator variable satisfying $\sum_k b_{j,k} = 1$. Here $b_{j,k}$ can be interpreted as the expectation of a binary indicator variable that describes the index of the source to which the k-th template is assigned.

2.2.3 Multichannel NMF (MNMF)

The covariance matrix of the observed signal $\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f)$ can be written as the linear sum of the outer products of a steering vector $\mathbf{a}_j(f)$ multiplied by a source variance $v_j(f,n)$. MNMF treats the outer product of a steering vector, namely the spatial covariance of the j-th source, denoted by $\mathbf{R}_j(f)$, as a full-rank matrix:

$$\begin{aligned}\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f) &= \sum\nolimits_j v_j(f,n)\mathbf{a}_j(f)\mathbf{a}_j^{\mathsf{H}}(f) \\ &= \sum\nolimits_j v_j(f,n)\mathbf{R}_j(f),\end{aligned} \tag{2.12}$$

while employing NMF to source variance $v_j(f,n)$.

Recently, several variants of MNMF have been proposed. Full-rank spatial covariance analysis (FCA) [70, 71, 72, 73] is a simple version of MNMF, where a source variance $v_j(f,n)$ is not modeled as an NMF. FastMNMF [74] is a fast variant of MNMF, which imposes joint diagonalizability on the spatial covariance $\mathbf{R}_j(f)$ and avoids calculating matrix inversions.

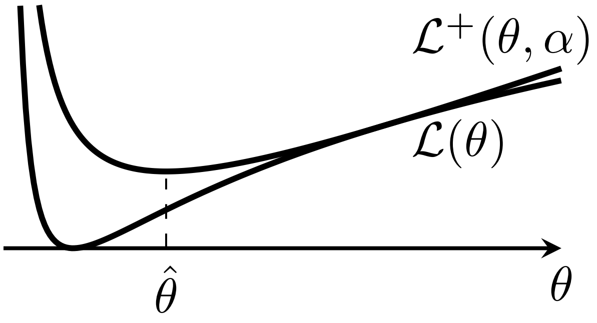


Figure 2.3: Illustration of majorization-minimization algorithm.

2.2.4 Majorization-Minimization (MM) Algorithm

The MM algorithm is an iterative algorithm that searches for a stationary point of an objective function by iteratively minimizing an auxiliary function called a “majorizer” that is guaranteed to never go below the objective function. When constructing an MM algorithm for a particular minimization problem, the criticalissue is to designing the majorizer. If a majorizer is properly designed, the algorithm is guaranteed to converge to a stationary point of the cost function. If we can build a tight majorizer/minorizer that is easy to optimize, we can generally expect to obtain a fast-converging algorithm.

Figure 2.3 shows an illustration of the MM algorithm. Let $\mathcal{L}(\theta)$ and θ be an objective function to be minimized and a parameter, respectively. A majorizer of $\mathcal{L}(\theta)$ is given as a function that satisfies

$$\mathcal{L}(\theta) = \min_{\alpha} \mathcal{L}^+(\theta, \alpha) \tag{2.13}$$

where α is an auxiliary variable. Then, we can show that the objective function is non-increasing under the following iterative updates:

$$\theta \leftarrow \operatorname*{argmin}_{\theta} \mathcal{L}^+(\theta, \alpha), \tag{2.14}$$

$$\alpha \leftarrow \operatorname*{argmin}_{\alpha} \mathcal{L}^+(\theta, \alpha). \tag{2.15}$$

2.2.5 Parameter Estimation

The optimization algorithm of MNMF consists of iteratively updating the spatial covariances $\mathcal{R} = \{\mathbf{R}_j(f)\}_{j,f}$, and the source variance parameters $\mathcal{H}_1 = \left\{h_{j,k_j}(f)\right\}_{j,k_j,f}$, $\mathcal{U}_1 = \left\{u_{j,k_j}(n)\right\}_{j,k_j,n}$ or $\mathcal{B} = \{b_{j,k}\}_{j,k}$, $\mathcal{H}_2 = \{h_k(f)\}_{k,f}$, $\mathcal{U}_2 = \{u_k(n)\}_{k,n}$. We can derive update equations using the principle of the MM algorithm. The optimal update of $\mathcal{R}$ is analytically obtained as:

$$\mathbf{R}_j(f) \leftarrow \mathbf{\Lambda}_j^{-1}(f)\#(\mathbf{R}_j(f)\mathbf{\Omega}_j(f)\mathbf{R}_j(f)), \tag{2.16}$$

where # denotes the geometric mean of two positive definite matrices [75]:

$$\mathbf{A}\#\mathbf{B} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}. \tag{2.17}$$

$\mathbf{\Lambda}_j(f)$, $\mathbf{\Omega}_j(f)$ are given as follows:

$$\mathbf{\Lambda}_j(f) = \sum\nolimits_n v_j(f,n)\hat{\mathbf{X}}^{-1}(f,n), \tag{2.18}$$

$$\mathbf{\Omega}_j(f) = \sum\nolimits_n v_j(f,n)\hat{\mathbf{X}}^{-1}(f,n)\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n), \tag{2.19}$$

where $\mathbf{X}(f,n)$ and $\hat{\mathbf{X}}(f,n)$ represent:

$$\mathbf{X}(f,n) = \mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n), \tag{2.20}$$

$$\hat{\mathbf{X}}(f,n) = \sum\nolimits_j v_j(f,n)\mathbf{R}_j(f). \tag{2.21}$$

The update rules for $\mathcal{H}_1$ and $\mathcal{U}_1$ can be derived as:

$$h_{j,k_j}(f) \leftarrow h_{j,k_j}(f)\sqrt{\frac{\sum_n u_{j,k_j}(n)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}{\sum_n u_{j,k_j}(n)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}}, \tag{2.22}$$

$$u_{j,k_j}(n) \leftarrow u_{j,k_j}(n)\sqrt{\frac{\sum_f h_{j,k_j}(f)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}{\sum_f h_{j,k_j}(f)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}}. \tag{2.23}$$

Similarly, the update rules for $\mathcal{B}$, $\mathcal{H}_2$, and $\mathcal{U}_2$ can be derived as:

$$b_{j,k} \leftarrow b_{j,k}\sqrt{\frac{\sum_{f,n} h_k(f)u_k(n)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}{\sum_{f,n} h_k(f)u_k(n)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}}, \tag{2.24}$$

$$h_k(f) \leftarrow h_k(f)\sqrt{\frac{\sum_{j,n} b_{j,k}u_k(n)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}{\sum_{j,n} b_{j,k}u_k(n)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}}, \tag{2.25}$$

$$u_k(n) \leftarrow u_k(n)\sqrt{\frac{\sum_{j,f} b_{j,k}h_k(f)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}{\sum_{j,f} b_{j,k}h_k(f)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}}. \tag{2.26}$$

Note that the update rules for a single-channel case can be obtained by:

$$h_{j,k_j}(f) \leftarrow h_{j,k_j}(f)\sqrt{\frac{\sum_n u_{j,k_j}(n)x^2(f,n)\hat{x}^{-2}(f,n)}{\sum_n u_{j,k_j}(n)\hat{x}^{-1}(f,n)}}, \tag{2.27}$$

$$u_{j,k_j}(n) \leftarrow u_{j,k_j}(n)\sqrt{\frac{\sum_f h_{j,k_j}(f)x^2(f,n)\hat{x}^{-2}(f,n)}{\sum_f h_{j,k_j}(f)\hat{x}^{-1}(f,n)}}, \tag{2.28}$$

$$b_{j,k} \leftarrow b_{j,k}\sqrt{\frac{\sum_{f,n} h_k(f)u_k(n)x^2(f,n)\hat{x}^{-2}(f,n)}{\sum_{f,n} h_k(f)u_k(n)\hat{x}^{-1}(f,n)}}, \tag{2.29}$$

$$h_k(f) \leftarrow h_k(f)\sqrt{\frac{\sum_{j,n} b_{j,k}u_k(n)x^2(f,n)\hat{x}^{-2}(f,n)}{\sum_{j,n} b_{j,k}u_k(n)\hat{x}^{-1}(f,n)}}, \tag{2.30}$$

$$u_k(n) \leftarrow u_k(n)\sqrt{\frac{\sum_{j,f} b_{j,k}h_k(f)x^2(f,n)\hat{x}^{-2}(f,n)}{\sum_{j,f} b_{j,k}h_k(f)\hat{x}^{-1}(f,n)}}. \tag{2.31}$$

2.2.6 Separation Process

After optimizing the spatial covariances and source variance parameters, source estimation is needed. The estimate of each source signal can be obtained as the minimum mean square error (MMSE) estimator. Let $\mathbf{c}_j(f,n) \in \mathbb{C}^I$ be j–th source image. Using the optimized parameters, the source image is given as:

$$p(\mathbf{c}_j(f,n)) = \mathcal{N}_{\mathbb{C}}(\mathbf{c}_j(f,n)|\mathbf{0}, v_j(f,n)\mathbf{R}_j(f)). \tag{2.32}$$

Since the joint distribution of j-th source image $\mathbf{c}_j(f,n)$ and mixture signal $\mathbf{x}(f,n)$ is also a complex Gaussian:

$$p(\mathbf{c}_j(f,n), \mathbf{x}(f,n)) = \mathcal{N}_{\mathbb{C}}\left(\begin{bmatrix}\mathbf{c}_j(f,n)\\ \mathbf{x}(f,n)\end{bmatrix}\middle|\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}v_j(f,n)\mathbf{R}_j(f) & v_j(f,n)\mathbf{R}_j(f)\\ v_j(f,n)\mathbf{R}_j(f) & \sum_j v_j(f,n)\mathbf{R}_j(f)\end{bmatrix}\right), \tag{2.33}$$

the conditional distribution becomes:

$$p(\mathbf{c}_j(f,n)|\mathbf{x}(f,n)) = \mathcal{N}_{\mathbb{C}}(\mathbf{c}_j(f,n)|\mathbf{W}_j(f,n)\mathbf{x}(f,n), v_j(f,n)(\mathbf{I}-\mathbf{W}_j(f,n))\mathbf{R}_j(f)). \tag{2.34}$$

The MMSE estimator of $\mathbf{c}_j(f,n)$ is the conditional expectation $\mathbb{E}\left[\mathbf{c}_j(f,n)|\mathbf{x}(f,n)\right] = \mathbf{W}_j(f,n)\mathbf{x}(f,n)$ and $\mathbf{W}_j(f,n)$ is known as a multichannel Wiener filter [44]:

$$\mathbf{W}_j(f,n) = v_j(f,n)\mathbf{R}_j(f)\left(\sum\nolimits_j v_j(f,n)\mathbf{R}_j(f)\right)^{-1}. \tag{2.35}$$

Similarly, in a single-channel scenario, the joint distribution of j-th source $s_j(f,n)$ and mixture signal $x(f,n)$ and the conditional distribution are expressed as:

$$p(s_j(f,n), x(f,n)) = \mathcal{N}_{\mathbb{C}}\left(\begin{bmatrix}s_j(f,n)\\ x(f,n)\end{bmatrix}\middle|\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}v_j(f,n) & v_j(f,n)\\ v_j(f,n) & \sum_j v_j(f,n)\end{bmatrix}\right), \tag{2.36}$$

$$p(s_j(f,n)|x(f,n)) = \mathcal{N}_{\mathbb{C}}(s_j(f,n)|W_j(f,n)x(f,n), v_j(f,n)(1-W_j(f,n))), \tag{2.37}$$

where the estimate of $s_j(f,n)$ can be given by $\mathbb{E}\left[s_j(f,n)|x(f,n)\right] = W_j(f,n)x(f,n)$ and $W_j(f,n)$ is single-channel Wiener filter:

$$W_j(f,n) = \frac{v_j(f,n)}{\sum_j v_j(f,n)}. \tag{2.38}$$

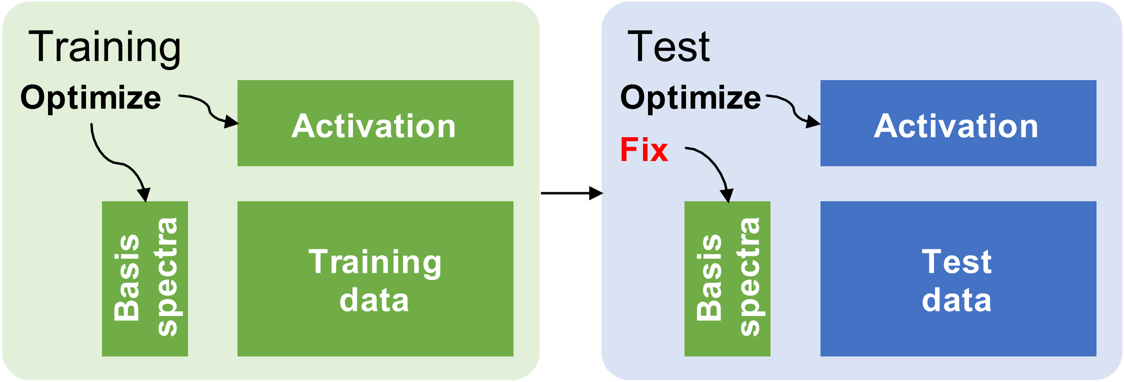


Figure 2.4: Block diagram of supervised NMF.

2.3 Supervised Source Separation

2.3.1 Supervised NMF

Although BSS can separate mixed signals under blind conditions, its separation performance is limited since prior information about the sources and the spatial characteristics of the environment are not considered. Supervised learning is a technique that improves the performance of BSS methods by taking as much information about the sources in a mixture of signals into account as possible.

Supervised NMF [45] is a representative method of supervised source separation, which consists of two steps, training and testing (Figure 2.4). During training, using training data for each source signal, the power (or amplitude) spectrogram is approximated using NMF and the spectral templates of the sources are learned. During testing, the learned spectral templates are then stacked with those of the other sources, and used as “fixed” or “initialized” in parameter estimation. Thanks to our prior knowledge about each source, the other parameters, e.g., the corresponding activations, are more accurately estimated, resulting in improvement in performance.

2.3.2 Neural Network-based methods

Recent advances in deep neural networks (DNNs) [76] have provided us with various new methods of handling source separation problems. These neural network-based methods can be divided into two types of approaches: generative approaches and discriminative approaches. The difference between these approaches is whether networks are used to directly represent conditional distributions (2.34) and (2.37) or mapping functions from mixture signals into source signals.

Discriminative Approaches

When using a typical discriminative approach, networks are used to represent mapping from a mixed signal to source signals, or corresponding time-frequency masking [77, 78] is used when the types of sources are fundamentally different, e.g., speech vs noise, singing voice vs accompaniments, etc. Thanks to the powerful representation capabilities derived from its activation functions and stacked layer architecture, mapping functions can easily learn to distinguish different sources.

An improved variant of discriminative approaches, deep clustering [79, 80, 81], has also been proposed. During training, the network uses a time-frequency bin to train an embedding to perform mapping. Each embedding of a time-frequency bin is trained so that embeddings from the same source converge, while those from different sources diverge. Since the network does not learn mapping functions for specific source signals, we can set an arbitrary number as the pre-defined number of sources at testing time. Source separation is performed by clustering these embeddings and building time-frequency masks which only pass the components belonging to the same clusters. As a result of continuous development, several extensions have been proposed [82, 83].

Another discriminative approach is to train a network to map mixture signals to the target sources, then the outputs are used as source variances in LGM [67, 68, 69]. While allowing us to use the powerful denoising abilities of the neural network, one shortcoming of this approach is that the separation algorithm does not guarantee an increase in the log-likelihood.

Several methods based on discriminative approaches have demonstrated impressive separation performance, however their performance tends to be negatively affected when used in environments containing unfamiliar noises, and they require large amounts of training data to achieve good performance.

Generative Approaches

With the development of deep generative models, e.g., variational autoencoder (VAE) [28], generative adversarial network (GAN) [84], and generative flow [85], several methods have attempted to alternate the source variances represented by NMF using neural networks [26, 27, 63, 64, 86, 87, 88, 89, 90, 91]. Since the networks are trained in advance to represent the source signals, mixed signal data is not required during training.

Among these deep generative models, VAE is the most frequently used since network training is easier than GAN, and the network architecture is more flexible than that of generative flow methods. VAE-NMF is a method which has been successfully used for speech enhancement [86, 87, 88, 89, 90, 91], where the speech signals are modeled with a VAE while noise is represented using an NMF. Let $\tilde{\mathbf{s}} = \{\tilde{s}(f)\}_f$ be a source spectrum in training data. The VAE consists of an encoder network $q_\phi(\mathbf{z}|\tilde{\mathbf{s}})$ and a decoder network $p_\theta(\tilde{\mathbf{s}}|\mathbf{z})$, where the encoder is expressed as a Gaussian distribution:

$$q_\phi(\mathbf{z}|\tilde{\mathbf{s}}) = \prod\nolimits_d \mathcal{N}(z(d)|\mu_\phi(d;\tilde{\mathbf{s}}), \sigma_\phi^2(d;\tilde{\mathbf{S}})). \tag{2.39}$$

$\mathbf{z}$ denotes a latent variable, and $z(d)$, $\mu_\phi(d;\tilde{\mathbf{s}})$, and $\sigma_\phi^2(d;\tilde{\mathbf{s}})$ represent the d–th elements of $\mathbf{z}$, $\mu_\phi(\tilde{\mathbf{s}})$, and $\sigma_\phi^2(\tilde{\mathbf{s}})$, respectively. The decoder distribution $p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, g)$ is expressed as a zero-mean complex Gaussian distribution, i.e., an LGM:

$$p_\theta(\tilde{\mathbf{s}}|\mathbf{z}, g) = \prod\nolimits_f \mathcal{N}_{\mathbb{C}}(\tilde{s}(f)|0, v(f)), \tag{2.40}$$

$$v(f) = g\sigma_\theta^2(f; \mathbf{z}), \tag{2.41}$$

where $\sigma_\theta^2(f; \mathbf{z})$ represents the (f)–th element of the decoder output $\sigma_\theta^2(\mathbf{z})$, and g is the global scale of the generated spectrogram. Encoder and decoder network parameters ϕ and θ are trained using the following objective function:

$$\mathcal{J}(\phi, \theta; \tilde{\mathbf{s}}) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\tilde{\mathbf{s}})}[\log p_\theta(\tilde{\mathbf{s}}|\mathbf{z})] - \mathrm{KL}[q_\theta(\mathbf{z}|\tilde{\mathbf{s}})||p(\mathbf{z})], \tag{2.42}$$

where $p(\mathbf{z})$ is a standard Gaussian distribution and $\mathrm{KL}[\cdot||\cdot]$ is the Kullback-Leibler divergence. Thus, source variance $\mathbf{v}(f)$ represents decoder output $\sigma_\theta^2(f; \mathbf{z})$ as scaled by g.

2.4 Evaluation Metrics for Source Separation

The performance of source separation methods is evaluated using objective metrics such as signal-to-distortion ratio (SDR), which represents the overall error between the reference signals and the estimated signals. SDR can be broken down into three components, the source image-to-spatial distortion ratio (ISR), the signal-to-interference ratio (SIR), and the signal-to-artifacts ratio (SAR) [92], where these metrics represent the amount of spatial distortion, interference and artifacts, respectively. Let $\mathbf{c}_{i,j}$ and $\hat{\mathbf{c}}_{i,j}$ be a true source image and an estimated source image and suppose that $\hat{\mathbf{c}}_{i,j}$ is decomposed as follows:

$$\hat{\mathbf{c}}_{i,j} = \mathbf{c}_{i,j} + \mathbf{e}_{i,j}^{\mathrm{spat}} + \mathbf{e}_{i,j}^{\mathrm{interf}} + \mathbf{e}_{i,j}^{\mathrm{artif}}, \tag{2.43}$$

where $\mathbf{e}_{i,j}^{\mathrm{spat}}$, $\mathbf{e}_{i,j}^{\mathrm{interf}}$, $\mathbf{e}_{i,j}^{\mathrm{artif}}$ are errors representing spatial distortion, interference, artifacts, respectively. SDR, ISR, SIR, and SAR of j–th image are respectively expressed in decibels, and are respectively given by:

$$\mathrm{SDR} = 10\log_{10}\frac{\sum_i \|\mathbf{c}_{i,j}\|_2^2}{\sum_i \|\mathbf{e}_{i,j}^{\mathrm{spat}} + \mathbf{e}_{i,j}^{\mathrm{interf}} + \mathbf{e}_{i,j}^{\mathrm{artif}}\|_2^2}, \tag{2.44}$$

$$\mathrm{ISR} = 10\log_{10}\frac{\sum_i \|\mathbf{c}_{i,j}\|_2^2}{\sum_i \|\mathbf{e}_{i,j}^{\mathrm{spat}}\|_2^2}, \tag{2.45}$$

$$\mathrm{SIR} = 10\log_{10}\frac{\sum_i \|\mathbf{c}_{i,j} + \mathbf{e}_{i,j}^{\mathrm{spat}}\|_2^2}{\sum_i \|\mathbf{e}_{i,j}^{\mathrm{interf}}\|_2^2}, \tag{2.46}$$

$$\mathrm{SAR} = 10\log_{10}\frac{\sum_i \|\mathbf{c}_{i,j} + \mathbf{e}_{i,j}^{\mathrm{spat}} + \mathbf{e}_{i,j}^{\mathrm{interf}}\|_2^2}{\sum_i \|\mathbf{e}_{i,j}^{\mathrm{artif}}\|_2^2}. \tag{2.47}$$

While these metrics are all calculated in time-frequency domain, distortions in feature domain are also sometimes important. The Mel-frequency cepstral coefficient (MFCC) distance is a metric used to measure error in the MFCC domain. Given true and estimated MFCCs $c_0, ..., c_D$ and $\hat{c}_0, ..., \hat{c}_D$, MFCC distance is defined as follows:

$$\text{MFCC distance [dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(c_d - \hat{c}_d)^2}. \tag{2.48}$$

2.5 Summary

In this chapter, a brief overview of source separation methods was provided. Categorization of source separation methods and fundamental components of separation algorithm were also described. Various supervised source separation methods were then reviewed.

Although supervised source separation methods have boosted performance, much room for improvement still remains.

3 Stereophonic Music Source Separation Using Timbre Information of Sources

This chapter describes a supervised source separation method for stereophonic music signals containing multiple recorded or processed signals, where synthesized music is focused on the stereophonic music. As the synthesized music signals are often generated as linear combinations of many individual source signals and their respective mixing gains, phase or phase difference information between inter-channel signals, which represent spatial characteristics of recording environments, cannot be utilized as acoustic clues for source separation. Non-negative tensor factorization (NTF) is an effective technique which can be used to resolve this problem by decomposing amplitude spectrograms of stereo channel music signals into basis vectors and activations of individual music source signals, along with their corresponding mixing gains. However, it is difficult to achieve sufficient separation performance using this method alone, as the acoustic clues available for separation are limited. To address this issue, this chapter proposes a cepstral distance regularization (CDR) method for NTF-based stereo channel separation, which involves making the cepstrum of the separated source signals follow Gaussian mixture models (GMMs) of the corresponding the music source signal. These GMMs are trained in advance using available samples. Experimental evaluations separating three and four sound sources are conducted to investigate the effectiveness of the proposed method in both supervised and partially supervised separation frameworks, and performance is also compared with that of a conventional NTF method. Experimental results demonstrate that the proposed method yields significant improvements within both separation frameworks, and that CDR provides better separation parameters.

3.1 Introduction

Music signals are widely available through various types of music media, such as CDs and download services via the internet, and can be listened to using devices such as CD players, portable audio players, computers and smartphones. These music signals are usually composed of multiple source signals collected from various instrumental sounds and vocals, and are often presented as two-channel, stereophonic signals corresponding to the left and right ears of listeners. An effective source separation technique for breaking up stereophonic music signals into its various component source signals would be useful in several applications, such as automatic music transcription [93] and extraction of vocals [94].

BSS is a popular framework used to separate mixed observation signals into individual source signals using only the mixed observation signals, and has been the focus of much study for many years [3, 7, 8]. BSS is classified into some problems, depending on the relationship between the number of the observed signals and that of the source signals. ICA [7, 33] is an effective method for solving BSS problems in overdetermined conditions, in which the number of the observation signals is larger than the number of source signals. ICA is used to build a time-invariant linear separation filter by assuming independence between the source signals, leading to high separation performance.

Table 3.1: Overview of BSS methods including the proposed method

Method	# of channels	Mixing condition	Spatial clues
ICA [7, 33]	Multichannel	Overdetermined	Yes
IVA [9, 10, 12]	Multichannel	Overdetermined	Yes
NMF [37]	Single channel	Underdetermined	No
MNMF [14, 15]	Multichannel	Underdetermined	Yes
Proposed	Multichannel	Underdetermined	No

IVA [9, 10, 12], which is one of the extensions of ICA, solves the permutation problem in FDICA [34, 35] and can thus achieve better performance. However, BSS problems in underdetermined conditions, in which the number of observed signals is fewer than the number of source signals, cannot be satisfactorily resolved using linear filters, and as a result separation performance is insufficient.

One effective source separation technique for such underdetermined BSS problems is NMF [37, 38]. NMF approximates a magnitude/power spectrogram of the observed signal as the product of two non-negative matrices by assuming the additive in the magnitude/power spectral domain. Since Wiener filters can be built by estimating the prior signal-to-noise ratio for every time-frequency slot of the observation signal obtained from the approximation, NMF can be used to solve underdetermined BSS problems. As an NMF extension for signal separation involving multiple observation signals, e.g., microphone array signals, MNMF has been proposed [14, 15]. Although source separation for multi-channel signals has also been achieved as shown in Table 1, most multi-channel source separation techniques require MNMF can introduce spatial information gathered from microphone locations as additional acoustic clues in addition to source information considered with conventional NMF, high initial value dependency has been observed. Source separation for multi-channel signals has also been achieved, however most multi-channel source separation techniques require phase information from the observation signals to perform separation, and it remains challenging to solve BSS problems in underdetermined conditions without such phase information. Separation of music signals composed of many source signals, i.e., synthesized music, whose source signals are recorded or processed individually, is an example of a source separation problem in underdetermined conditions which must be solved without phase information. Unlike recorded music in which all of the source signals are played in concert, we cannot utilize the spatial characteristics of synthesized music signals as additional acoustic clues, leaving only magnitude information to help us.

In the source separation problems on the synthesized music in which only the magnitude information of the observation signals is available, if prior information about the source signals to be separated can be obtained, this can provide additional clues for separation. In contrast to BSS, source separation methods using training data (prior information), known as supervised source separation, have also been proposed. Supervised NMF [45, 46, 47] is one of these supervised source separation techniques. Pitch information (spectral harmonic structures of the source signals) and timbre information (spectral envelopes of the source signals) from training data are trained simultaneously and then used to separate the observation signals by fixing the trained parameters. Supervised NMF can precisely separate source signals whose spectral structures are similar to those of the training data. However, when there are some mismatches between the training data and the observation signals, separation becomes difficult, leading to insufficient separation performance, thus special techniques are needed to compensate for such mismatches. This can be avoided by using the trained parameters as initial values in the separation algorithm, which is referred to as lightly-supervised

NMF. Since lightly-supervised NMF updates the parameters used for the separation, it can adapt to differences in the source signals between the training data and the target data. On the other hand, it easily suffers from the overfitting problem. To address this issue, cepstral distance regularization (CDR) [22] which is used as speech enhancement and can enhance both spectra and features of sources signals has been proposed. CDR does not constrain each parameter to be estimated, but instead constrains the estimated source signals, and forces them to follow the spectral envelopes of the training data.

This chapter proposes a stereophonic music separation method for the synthesized music which models the music generation process using a NTF [21]-based technique and assumes that the spectrograms of the observed stereo channel signals have low-rank structures in the magnitude/power spectral domain, as in conventional NMF methods (Table 1). In addition, the proposed method applies either a supervised or lightly-supervised separation framework and also introduces soft constraints for the timbre information for each source using CDR.

This chapter is organized as follows. Section 3.2 begins from basic formulation of NMF and describes the concept of CDR. Our proposed stereophonic music separation method, our assumptions about the stereo channel signal mixing process and our CDR adaptation are described in Section 3.3. Section 3.4 describes our experimental evaluation of the proposed method using three- and four-source signal separation tasks, and our results are reported. The effectiveness of CDR when used with music signals is also evaluated. Section 3.5 summarizes this chapter.

3.2 Cepstral Distance Regularization (CDR)

Let $\mathbf{X} \in \mathbb{R}_{\geq 0}^{K \times N}$, $\mathbf{T} \in \mathbb{R}_{\geq 0}^{K \times B}$, and $\mathbf{U} \in \mathbb{R}_{\geq 0}^{B \times N}$ be the amplitude or the power spectrogram of a mixture signal recorded with a microphone, a basis matrix expressing spectral patterns, and an activation matrix which represents time-varying gains corresponding $\mathbf{T}$. NMF approximates the amplitude or the power spectrogram $\mathbf{X}$ as

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{T}\mathbf{U}, \tag{3.1}$$

where $k \in \{1, \ldots, K\}$, $b \in \{1, \ldots, B\}$, and $n \in \{1, \ldots, N\}$ are indices denoting frequency bins, the basis spectra, and time frames, respectively. An estimate $\hat{x}_{kn}$ corresponding to x_{kn} of the observation matrix $\mathbf{X}$ is given by linear combination of the basis spectra B:

$$\hat{x}_{kn} = \sum_{b} t_{kb} u_{bn}. \tag{3.2}$$

CDR is a regularization process used in the field of speech enhancement which constrains the estimated speech to be enhanced so that its distribution in a feature space follows the distribution of the speech in the training data. A CDR term is defined as follows:

$$\mathcal{K}(\hat{\mathbf{X}}) = -\log \prod_{n} \sum_{p} w_p \prod_{q} \mathcal{N}(E_{qn}; \mu_{pq}, \sigma_{pq}^2), \tag{3.3}$$

$$E_{qn} = \sum_{r} c_{qr} \log \sum_{k} f_{rk} \hat{x}_{kn}, \tag{3.4}$$

where E_{qm} is the mel-frequency cepstral coefficient (MFCC) of $\hat{x}_{kn}$. $\mathbf{f} = \{f_{rk}\} \in \mathbb{R}^{R \times K}$ is an R-dimensional filter-bank matrix and $\mathbf{c} = \{c_{qr}\} \in \mathbb{R}^{(Q+1) \times R}$ is the 0-through-Qth part of an inverse cosine transform matrix. (3.3) denotes the negative log-likelihood of GMM with a parameter $\{w_p, \mu_p, \mathbf{\Sigma}_p\}_{1 \leq p \leq P}$, where w_p, $\mu_p = (\mu_{p0}, \ldots, \mu_{pQ})^{\mathsf{T}}$, and $\mathbf{\Sigma}_p = \mathrm{diag}(\sigma_{p0}^2, \ldots, \sigma_{pQ}^2)$ are a mixture component weight, a mean vector, and a covariance matrix, respectively. The parameter is trained in advance using available samples of target speech, and then used with fixed. Thus, CDR enhances estimated speech by forcing it to adopt the same, specific spectral envelope, i.e., voice timbre as the training speech. Since GMMs are used as probabilistic models of the spectral envelopes, CDR also provides soft clustering criteria for speech in the feature space.

3.3 Proposed Method

3.3.1 Stereophonic Music Signal Mixing Process

Since it is not helpful to use inter-channel phase information as clues for the separation of synthesized music, we assume that the observed stereophonic music signals are created by controlling the amplitude of individual music source signals to the left and right channels, i.e., panning and then mixing the resulting stereo channel signals of the individual source signals. In NTF-based separation, we further assume that this mixing process is also applied in a similar manner to the amplitude/power spectral domain. Let $\mathcal{S}_{\mathcal{C}} \in \mathbb{C}^{K \times N \times C}$ be sets of the complex spectrograms of the observation signals, composed of $c \in \{1, \ldots, C\}$ channels ($C = 2$ in this study), while $\mathcal{X}_{\mathcal{C}} \in \mathbb{C}^{K \times N \times M}$ represents the complex spectrograms of the $m \in \{1, \ldots, M\}$ source signals. Given gain matrix $\mathbf{G} \in \mathbb{R}_{\geq 0}^{M \times C}$ which controls the panning operations, the mixing process in Figure 3.1 can be represented as:

$$\mathcal{S}_{\mathcal{C}} = f(\mathbf{G}, \mathcal{X}_{\mathcal{C}}), \tag{3.5}$$

where f is a function satisfying linear operation for a matrix and a multi-dimensional array, i.e., tensor. We then introduce the following approximation in the magnitude spectra domain:

$$\mathcal{S} \approx \hat{\mathcal{S}} = f(\mathbf{G}, \hat{\mathcal{X}}), \tag{3.6}$$

where $\mathcal{S} \in \mathbb{R}_{\succeq 0}^{K \times N \times C}$ is the magnitude spectra of $\mathcal{S}_{\mathcal{C}}$. $\hat{\mathcal{S}} \in \mathbb{R}_{\succeq 0}^{K \times N \times C}$ and $\hat{\mathcal{X}} \in \mathbb{R}_{\succeq 0}^{K \times N \times M}$ are the estimated magnitude spectra of the observation signals and source signals, respectively. Furthermore, $\hat{\mathcal{X}}$ is decomposed into a set of basis vectors $\mathcal{T} \in \mathbb{R}_{\succeq 0}^{K \times B \times M}$, and their corresponding activations $\mathcal{U} \in \mathbb{R}_{\succeq 0}^{B \times N \times M}$, using NMF. Each estimate of the stereo channel observation signals $\hat{s}_{knc}$ and that of the low-rank representation of individual music source signals $\hat{x}_{knm}$ are respectively modeled as follows:

$$\hat{s}_{knc} = \sum_m g_{mc} \hat{x}_{knm}, \tag{3.7}$$

$$\hat{x}_{knm} = \sum_b t_{kbm} u_{bnm}, \tag{3.8}$$

where the variables, g_{mc}, t_{kbm}, and u_{bnm} represent components of the parameter sets to be estimated while all of them are nonnegative. While this mixing process is represented as a form of NTF [21], in order to decompose the tensor-form the proposed NTF model shares gain information over frequencies and basis vectors components, and has the different gains only for the source signals, reducing model complexity.

3.3.2 Introduction to CDR

The objective function with the CDR to be minimized is defined as follows:

$$\mathcal{I}(\theta) = \mathcal{D}.(\mathcal{S}|\hat{\mathcal{S}}) + \lambda \mathcal{K}(\hat{\mathcal{X}}), \tag{3.9}$$

where $\mathcal{D}.(\mathcal{S}|\hat{\mathcal{S}})$ is an error function representing the gap between observations and estimates. λ is a regularization parameter, and $\mathcal{K}(\hat{\mathcal{X}})$ is the CDR term for estimates

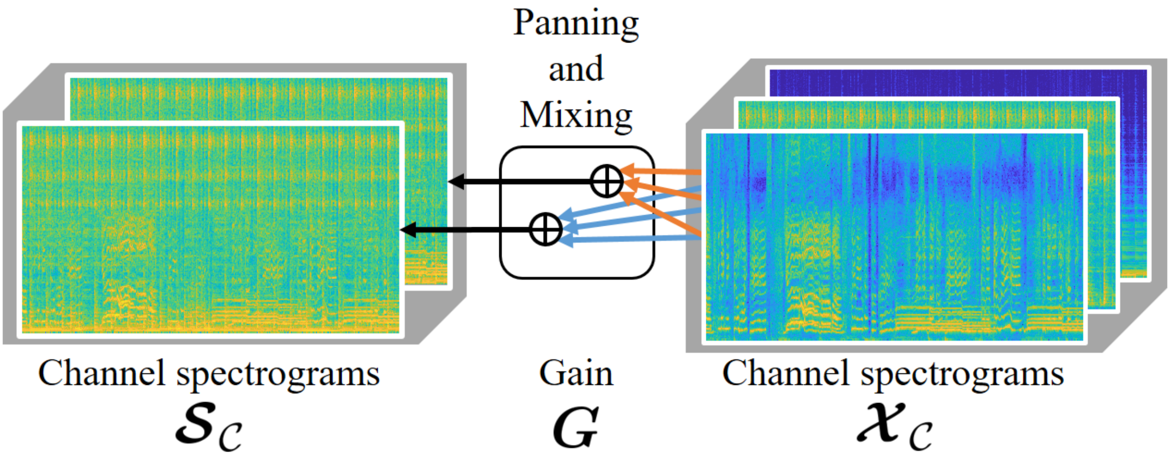


Figure 3.1: Assumed stereophonic music mixing process (3.5).

of the individual music source signals $\hat{\mathcal{X}}$, which is given by:

$$\mathcal{K}(\hat{\mathcal{X}}) = -\log \prod_{n,m} \sum_{p} w_{pm} \prod_{q} \mathcal{N}(E_{qnm}; \mu_{qpm}, \sigma^2_{qpm}), \tag{3.10}$$

where E_{qnm} is the MFCC feature of the individual estimated sources represented as:

$$E_{qnm} = \sum_{r} c_{qr} \log \sum_{k} f_{rk} \hat{x}_{knm}. \tag{3.11}$$

The regularization term insures that the spectral envelopes of the estimated individual music source signals are similar to the desired ones, which are modeled with the source-dependent GMMs (in Figure 3.2).

3.4 Parameter Estimation

We derive a convergence-guaranteed separation algorithm for minimizing (3.9) based on the MM principle.

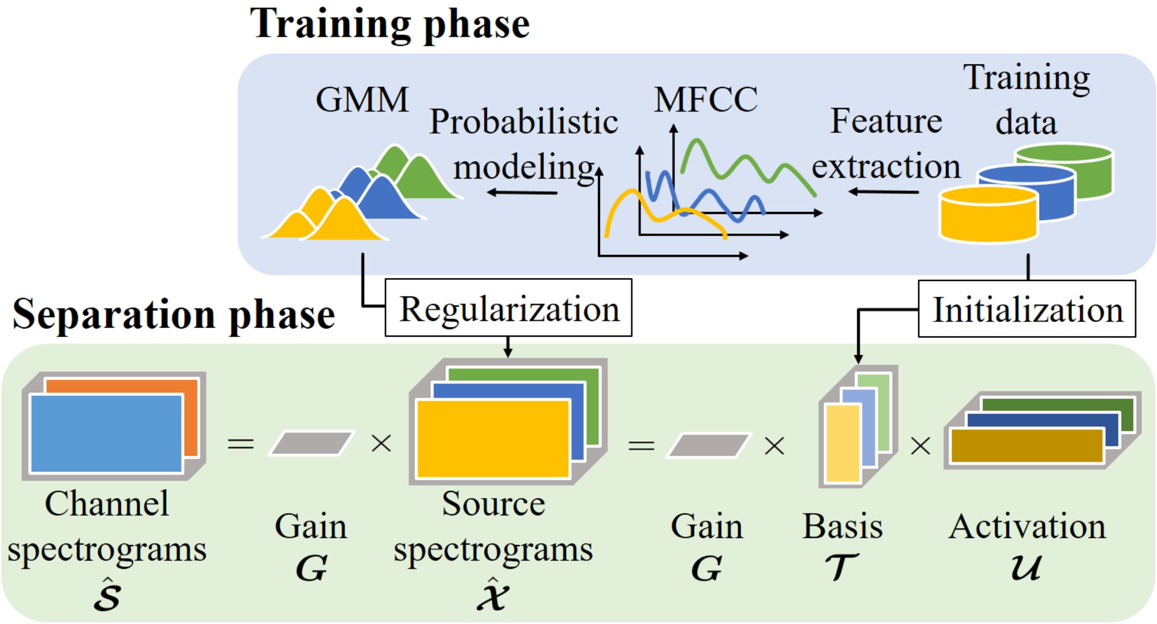


Figure 3.2: Overview of the proposed method. Training data is used to for regularization of the estimated source spectrograms as well as the initialization for the basis tensor.

3.4.1 Update Rules for G

In (3.9), the first term is related to parameter $\mathbf{G}$. For $\mathcal{D}_{\cdot}(\mathcal{S}|\hat{\mathcal{S}})$, we use the KL-divergence as the error function:

$$\mathcal{D}_{\mathrm{KL}}(y|x) = y \log \frac{y}{x} - (y - x). \tag{3.12}$$

Using Jensen's inequality, we obtain an upper bound of $\mathcal{D}_{\mathrm{KL}}(\mathcal{S}|\hat{\mathcal{S}})$ as follows:

$$\begin{aligned} &\mathcal{D}_{\mathrm{KL}}(\mathcal{S}|\hat{\mathcal{S}}) \\ &= \sum_{k,n,c} \left[s_{knc} \log \frac{s_{knc}}{\hat{s}_{knc}} - (s_{knc} - \hat{s}_{knc}) \right] \\ &\overset{\mathrm{c}}{\leq} \sum_{k,b,n,m,c} \left[g_{mc} t_{kbm} u_{bnm} - s_{knc} \alpha_{kbnmc} \log \frac{g_{mc} t_{kbm} u_{bnm}}{\alpha_{kbnmc}} \right] \end{aligned} \tag{3.13}$$

where $\overset{\mathrm{c}}{\leq}$ denotes an inequality only for the parameters to be estimated [65, 66]. We can use the right-hand side of the (3.13) as a majorizer, where $\alpha = \{\alpha_{kbnmc}\}$ is a variable satisfying $\sum_{b,m} \alpha_{kbnmc} = 1$. The equality of the (3.13) holds when:

$$\alpha_{kbnmc} = \frac{g_{mc} t_{kbm} u_{bnm}}{\hat{s}_{knc}}. \tag{3.14}$$

Then, the update rules for $\mathbf{G}$ can be obtained in the same manner as the regular NMF, as follows:

$$g_{mc} \leftarrow \frac{\sum_{k,b,n} s_{knc} \alpha_{kbnmc}}{t_{kbm} u_{bnm}}. \tag{3.15}$$

3.4.2 Update Rules for $\mathcal{T}$ and $\mathcal{U}$

In (3.9), both the first and the second terms are related to parameters $\mathcal{T}$ and $\mathcal{U}$. By applying a technique similar to the one described in [22], we obtain an upper bound of the CDR term as follows:

$$
\begin{aligned}
&\mathcal{K}(\hat{\mathcal{X}}) \\
&\overset{c}{\leq} \sum_{r,n,m} \Bigg[A_{rnm} \Bigg\{ \sum_{k,b} \frac{\phi_{rkbnm}^2}{f_{rk} t_{kbm} u_{bnm}} + p(\xi_{knm}) \varsigma_{rnm} + q(\xi_{rnm}) \Bigg\} \\
&\quad - \delta_{B_{rnm}<0} \left| B_{rnm} \right| \sum_{k,b} \psi_{rkbnm} \frac{f_{rk} t_{kbm} u_{bnm}}{\psi_{rkbnm}} + \delta_{B_{rnm} \geq 0} \left| B_{knm} \right| \left\{ \frac{\varsigma_{rnm}}{\zeta_{rnm}} + \log \zeta_{rnm} - 1 \right\} \Bigg],
\end{aligned}
\tag{3.16}
$$

where $\mathbf{A} = \{A_{rnm}\}$, $\mathbf{B} = \{B_{rnm}\}$, and $\varsigma = \{\varsigma_{rnm}\}$ are respectively defined as follows:

$$
A_{rnm} = \sum_{p,q} \frac{\beta_{pnm} c_{qr}^2}{2\sigma_{pqm}^2 \omega_{pqrnm}}, \tag{3.17}
$$

$$
B_{rnm} = -\sum_{p,q} \frac{\beta_{pnm} c_{qr} \gamma_{pqrnm}}{\sigma_{pqm}^2 \omega_{pqrnm}}, \tag{3.18}
$$

$$
\varsigma_{rnm} = \sum_{k,b} f_{rk} t_{kbm} u_{bnm}. \tag{3.19}
$$

where $p(\xi_{rnm})$ and $q(\xi_{rnm})$ are non-linear functions given by:

$$
p(\xi_{rnm}) = \frac{2 \log \xi_{rnm}}{\xi_{rnm}} + \frac{1}{\xi_{rnm}^2}, \tag{3.20}
$$

$$
q(\xi_{rnm}) = (\log \xi_{rnm})^2 - 2 \log \xi_{rnm} - \frac{2}{\xi_{rnm}}, \tag{3.21}
$$

and δ_x is an indicator function assuming the value 1 when the condition x is satisfied, and 0 otherwise. The equality in (3.16) holds when:

$$
\beta_{pnm} = \frac{w_{pm} \prod_q \mathcal{N}(E_{qnm}; \mu_{pqm}, \sigma_{pqm}^2)}{\sum_{p'} w_{p'm} \prod_{q'} \mathcal{N}(E_{q'nm}; \mu_{p'q'm}, \sigma_{p'q'm}^2)}, \tag{3.22}
$$

$$
\gamma_{pqrnm} = c_{qr} \log \varsigma_{rnm} + \omega_{pqrnm} (\mu_{pqm} - E_{qnm}), \tag{3.23}
$$

$$
\xi_{rnm} = \zeta_{rnm} = \varsigma_{rnm} = \sum_{k,b} f_{rk} t_{kbm} u_{bnm}, \tag{3.24}
$$

$$
\phi_{rkbnm} = \psi_{rkbnm} = \frac{f_{rk} t_{kbm} u_{bnm}}{\sum_{k',b'} f_{rk'} t_{k'b'm} u_{b'nm}}, \tag{3.25}
$$

where $\beta = \{\beta_{pnm}\}$ and $\gamma = \{\gamma_{pqrnm}\}$ are variables satisfying $\sum_p \beta_{pnm} = 1$ and $\sum_r \gamma_{pqrnm} = \mu_{pqm}$, respectively, and $\omega = \{\omega_{pqrnm}\}$ is an arbitrary positive constant satisfying $\sum_r \omega_{pqrnm} = 1$.

We can use the right-hand sides of the (3.13) and (3.16) as majorizers. The update rules for $\mathcal{T}$ and $\mathcal{U}$ can be derived as follows:

$$t_{kbm} \leftarrow \frac{-\mathsf{b}_{kbm} + \sqrt{\mathsf{b}_{kbm}^2 - 4\mathsf{a}_{kbm}\mathsf{c}_{kbm}}}{2\mathsf{a}_{kbm}}, \tag{3.26}$$

$$u_{bnm} \leftarrow \frac{-\mathsf{e}_{bnm} + \sqrt{\mathsf{e}_{bnm}^2 - 4\mathsf{d}_{bnm}\mathsf{f}_{bnm}}}{2\mathsf{d}_{bnm}}. \tag{3.27}$$

where a_{kbm}, b_{kbm}, c_{kbm}, d_{bnm}, e_{bnm}, and f_{bnm} are respectively given as follows:

$$\mathsf{a}_{kbm} = \sum_{n,c} g_{mc} u_{bnm} + \lambda \sum_{r,n} A_{rnm} p(\xi_{rnm}) f_{rk} u_{bnm} + \lambda \sum_{r,n} \frac{\delta_{B_{rnm} \geq 0} |B_{rnm}|}{\zeta_{rnm}} f_{rk} u_{bnm}, \tag{3.28}$$

$$\mathsf{b}_{kbm} = -\sum_{n,c} s_{knc} \alpha_{kbnmc} - \lambda \sum_{r,n} \delta_{B_{rnm} < 0} |B_{rnm}| \psi_{rkbnm}, \tag{3.29}$$

$$\mathsf{c}_{kbm} = -\lambda \sum_{r,n} A_{rnm} \frac{\phi_{rkbnm}^2}{f_{rk} u_{bnm}}, \tag{3.30}$$

$$\mathsf{d}_{bnm} = \sum_{k,c} g_{mc} t_{kbm} + \lambda \sum_{r,k} A_{rnm} p(\xi_{rnm}) f_{rk} t_{kbm} + \lambda \sum_{r,k} \frac{\delta_{B_{rnm} \geq 0} |B_{rnm}|}{\zeta_{rnm}} f_{rk} t_{kbm}, \tag{3.31}$$

$$\mathsf{e}_{bnm} = -\sum_{k,c} s_{knc} \alpha_{kbnmc} - \lambda \sum_{r,k} \delta_{B_{rnm} < 0} |B_{rnm}| \psi_{rkbnm}, \tag{3.32}$$

$$\mathsf{f}_{bnm} = -\lambda \sum_{r,k} A_{rnm} \frac{\phi_{rkbnm}^2}{f_{rk} t_{kbm}}. \tag{3.33}$$

The parameters $\mathbf{G}$ and $\mathcal{U}$ are initialized randomly, and $\mathcal{T}$ is initialized using the basis matrices trained using individual source signal data in the same manner as supervised or lightly-supervised separation in the regular NMF framework. The parameters $\mathbf{G}$

Table 3.2: Music list.

song #	Artist	Title	Duration
1	Actions	Devil's Words	3'17"
2	Actions	One Minute Smile	2'44"
3	Actions	South of The Water	3'11"

and $\mathcal{T}$ are normalized after updating (3.15) and (3.26) using the following ways.

$$g_{mc} \leftarrow \frac{g_{mc}}{\sum_c g_{mc}}, \tag{3.34}$$

$$t_{kbm} \leftarrow \frac{t_{kbm}}{\sum_k t_{kbm}}. \tag{3.35}$$

Wiener filters for each source signal can be built after parameter estimation, and then separated signals are extracted.

3.5 Experimental Evaluation

3.5.1 Experimental Settings

We conducted a music source separation experiments using real, synthesized music signals. Three songs distributed by Cambridge Music Technology [54] (Table 3.2) were used in the experiments; Songs 2 and 3 were used for training data, and Song 1 was used for development data (20–30 sec.) and evaluation data (50–65 sec.). These songs were written and performed by the same artists, and had similar source structures. Source signals of these music data were available, and four source signals; Bass (Ba), Drums (Dr), Vocals (Vo), and Guitar (Gt) were used in the experiment, and the individual source signals were also used separately for the training. All music signals were downsampled from 44.1 kHz to 16 kHz, and spectrograms were obtained with frame analysis using 32 ms window and 16 ms shift with the square-root Hanning window function. The MFCCs were extracted using 64-dimensional mel-filterbanks. The number of mixture components P and the number of MFCC coefficients Q were optimally selected based on likelihoods for development data, thus they are different for each source signal m, i.e., $P \to P_m$ and $Q \to Q_m$.

3.5.2 Separation Results

We evaluated the separation performance of the proposed method in two separation frameworks; a lightly-supervised separation framework, where all NTF parameters were estimated, and a supervised separation framework, where only the panning gain matrix and the activation tensor were updated, while the basis was set to that optimized using the training data. During the evaluation, the parameters to be estimated were first updated 200 times without the CDR, and then they were updated 200 times using the CDR. we conducted the evaluation for three different panning conditions:

1. *Three sources*: Ba, Dr, and Vo are panned to left and right channels by 17:13, 1:1, and 13:17, respectively.
2. *Four sources (line)*: Ba, Dr, Vo, and Gt are panned to left and right channels by 29:11, 23:17, 17:23, and 11:29, respectively.
3. *Four sources (set)*: Ba, Dr, Vo, and Gt are panned to left and right channels by 17:13, 1:1, 1:1, and 13:17, respectively.

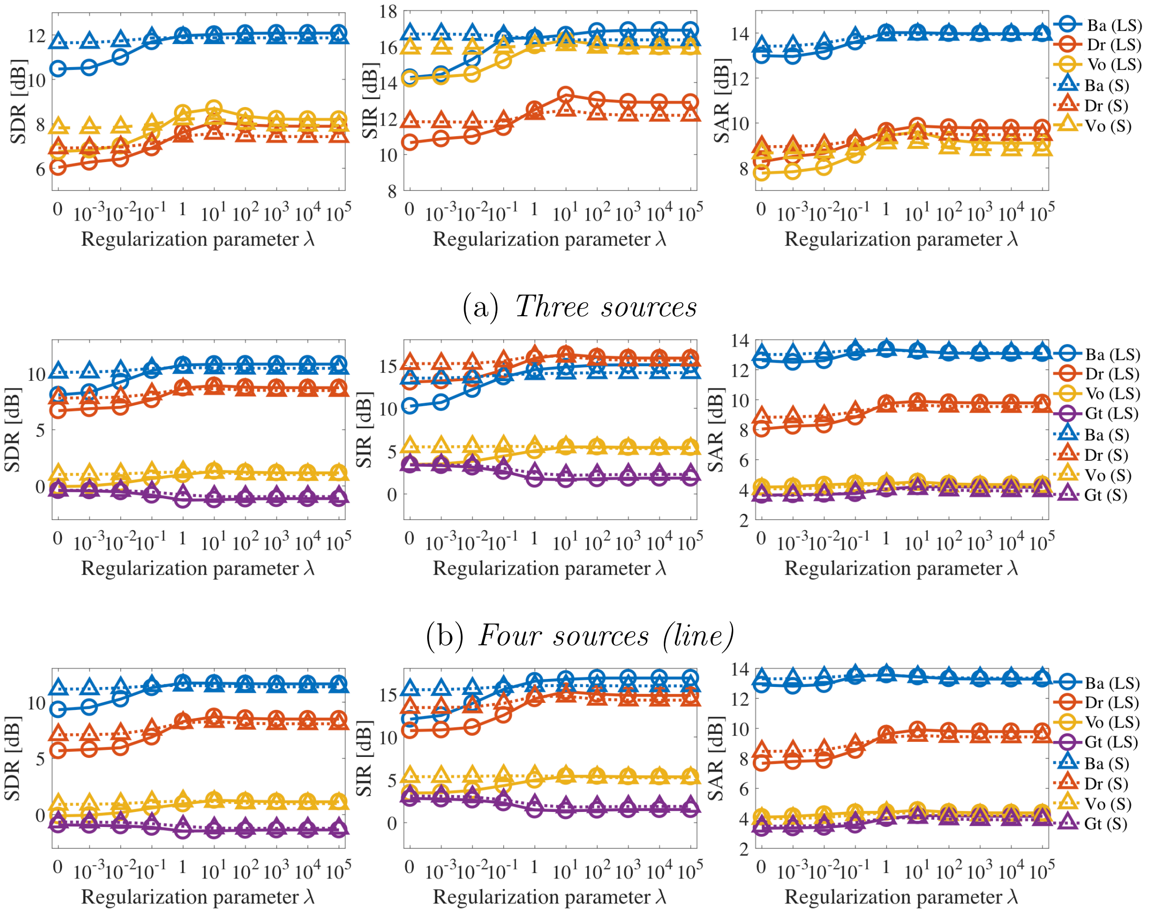
Separation performance was evaluated in each setting of the regularization parameter, i.e., $\lambda = 0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}$ and 10^{5}, where $\lambda = 0$ was equivalent to NTF-based separation without the CDR, and the parameters were updated 400 times. In order to reduce the effect of random parameter initialization on the separation performance, the separation process was conducted five times by changing an initial setting in each condition. The number of basis vectors for each source was set to 50.

As the performance measurements, Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR) of the estimated stereo channel music signals were calculated using the BSS EVAL toolbox [92]. SDR, SIR, and SAR represent the overall sound qualities of the separated signals, the level of suppression of the non-target signals in the separated signals, and the level of distortion caused by processing, respectively, with larger values representing better performance. These measurements were calculated in each channel, and then, they were averaged over two channels.

Regularization Effects for Separations

Figure 3.3 shows results of SDR, SIR, and SAR after supervised (S) and lightly-supervised (LS) separation. In each figure, the horizontal axis shows the setting of the regularization parameter, while the vertical axis represents performance. Separation performance of each source signal is shown separately in the figure. Note that results of the proposed NTF without the CDR are shown as $\lambda = 0$.

We can see that the CDR yields significant performance improvements in both lightly-supervised and supervised separation frameworks by suitably setting the regularization parameter λ to around 1 to 10^2. At such a suitable setting, we can also see that the lightly-supervised separation performance outperforms the supervised separation performance. On the other hand, if the CDR is not used ($\lambda = 0$), the lightly-supervised separation performance is significantly degraded and it becomes less effective than the supervised separation. These results suggest that 1) the supervised separation performance is limited because the basis vectors are strongly affected by acoustic



(a) *Three sources*

(b) *Four sources (line)*

(c) *Four sources (set)*

Figure 3.3: Separation performances for each signal as measured by SDR (left), SIR (center) and SAR (right), after lightly-supervised (LS) and supervised (S) separation. The CDR is not used when the regularization parameter λ is set to 0.

mismatches between the training and evaluation data, 2) updating the basis vectors is helpful in compensating those mismatches but this is difficult to be achieved in the normal NTF without the CDR, and 3) the proposed NTF-based separation with the CDR is capable of effectively updating the basis vectors, while also yielding significant improvements in separation performance. These results also show that the CDR can be helpful for not only speech enhancement but also general music source separation. This is because individual music source signals have their own specific spectral envelopes as speech signals do and they are effectively used as acoustic clues for separation. We can also see that the separation performance is strongly dependent on the individual source signals.

The CDR not only compensates for acoustic mismatches existing between the training data and the evaluation data, but also improves gain estimation. Table 3.3 shows the estimated gains to the left channel signals when using the proposed method with and without the CDR, as well as actual mixing gains (Ground truth). Gain estimated with the CDR is closer to the ground truth than without the CDR, suggesting that the CDR is able to induce the gains, improving estimation performance.

Comparison with Conventional NTF Separation

We also compared the separation performance of the proposed method with that of conventional NTF separation methods. Given the observation s_{knc}, the regular NTF [21] represent it as follows:

$$s_{knc} \approx \hat{s}_{knc} = \sum_{b'}^{B'} g_{cb'} t_{kb'} u_{nb'}. \tag{3.36}$$

While the proposed NTF framework (shown in (3.7)) uses the same gain for each individual source signal, regular NTF employs different gains for each basis vector

Table 3.3: Gain estimation results for the each signal. These are ground truth and estimated gains of individual source signals to left channel signals. Bold values represent the estimated gain values nearer to the ground truth.

(a) *Three sources*

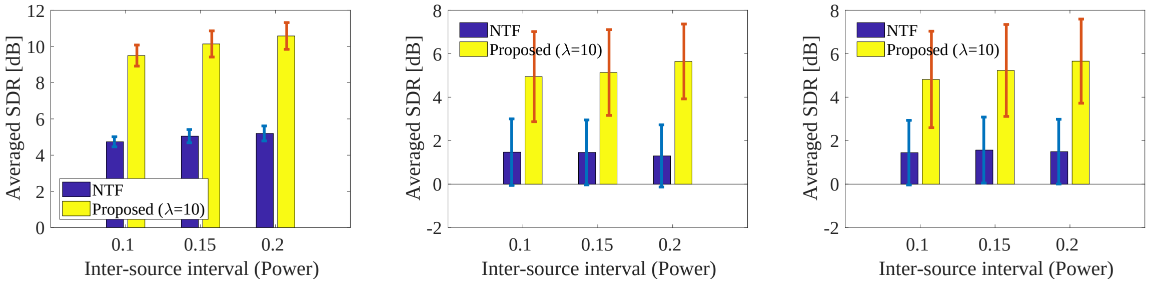
Method	Ba	Dr	Vo
Ground truth	0.350	0.500	0.650
w/o CDR	0.430	**0.503**	0.576
w/ CDR	**0.413**	**0.497**	**0.579**

(b) *Four sources (line)*

Method	Ba	Dr	Vo	Gt
Ground truth	0.300	0.425	0.575	0.700
w/o CDR	0.397	0.465	**0.577**	**0.610**
w/ CDR	**0.366**	**0.460**	**0.577**	0.609

(c) *Four sources (set)*

Method	Ba	Dr	Vo	Gt
Ground truth	0.350	0.500	0.500	0.650
w/o CDR	0.426	0.504	0.543	**0.575**
w/ CDR	**0.400**	**0.501**	**0.539**	0.572



(a) three sources (b) four sources (line) (c) four sources (set)

Figure 3.4: Comparison with the regular NTF using the lightly-supervised framework.

$\mathbf{t}_{b'} = [t_{1b'}, \cdots, t_{Kb'}]^{\mathsf{T}}$. We compared regular NTF with the proposed method within a lightly-supervised separation framework with ($\lambda = 10$). The total number of basis vectors B' in conventional NTF was set to 150 and the number of iterations was set to 400. Separation performance was investigated at various panning intervals, and panning gain values between each source were changed from 0.1 to 0.2 (prior experiments were conducted with inter-source panning intervals as 0.15.) The initial values of the basis vectors were similarly initialized by the training data.

Figure 3.4 shows a comparison of SDR results within a lightly-supervised framework when using conventional NTF and the proposed method. In each figure, horizontal axis is the panning intervals for each source signal, in which larger values represent wider mixing gain settings, while vertical axis in each figure shows averaged SDRs over estimated sources. Error bars represent the 95 % confidence intervals. These results show that 1) the proposed NTF model significantly outperforms conventional NTF, and 2) the proposed method achieves better performances consistently when enlarging the inter-source intervals.

3.6 Summary

In this chapter we have proposed a stereophonic music separation method using a lightly-supervised separation framework. The proposed method targets synthesized music signals represented as two-channel observation signals, composed of many source signals which have then been combined by distributing their gains to the observation channels and mixing them. The proposed separation method is based on Non-negative Tensor Factorization, a technique which does not require the use of phase information as acoustic clues for signal separation. Moreover, the proposed method uses Cepstral Distance Regularization, which constrains the estimated source signals to follow certain p.d.f.s distributed in a feature space. Our experimental results have demonstrated that 1) the proposed method outperforms the conventional NTF method within a lightly-supervised framework, 2) the proposed method, with appropriate regularization setting, has the potential to achieve better performance in a lightly-supervised separation framework than the conventional methods operating in a supervised separation framework, because the proposed method can compensate for mismatches between the training and evaluation data, and 3) CDR can be used to obtain superior parameter estimation results.

4 Missing Component Restoration Considering Redundancy of Time-Frequency Representation

While time-frequency masking is a powerful approach for speech enhancement in terms of signal recovery accuracy, e.g., signal-to-noise ratio, it can over-suppress and damage speech components, leading to limited performance of succeeding speech processing systems. To overcome this shortcoming, this chapter describes a method to restore missing components of time-frequency masked speech spectrograms based on direct estimation of a time domain signal. The proposed method allows us to take account of the local interdependencies of the elements of the complex spectrogram derived from the redundancy of a time-frequency representation as well as the global structure of the magnitude spectrogram. The effectiveness of the proposed method is demonstrated through experimental evaluation, using spectrograms filtered with masks to enhance noisy speech. Experimental results show that the proposed method significantly outperformed conventional methods, and has the potential to estimate both phase and magnitude spectra simultaneously and precisely.

4.1 Introduction

The presence of background noise can significantly degrade the quality of speech traveling through transmission systems and negatively affect the performance of speech recognition and speech conversion systems. The performance of these systems can be improved by suppressing the noise in observed audio signals and enhancing the target speech.

One effective approach for speech enhancement involves time-frequency masking, which extracts only the components in the time-frequency slots that are expected to be dominated by the target speech [95]. There are several ways to perform time-frequency masking. For example, by using microphone inputs we can cluster time-frequency slots according to the direction of arrival of each source [18]. For monaural recording, we can use deep neural networks to assign a source label to every time-frequency slot, or partition the spectrogram into different source regions according to the local "texture" of the spectrogram [96, 79]. While these methods allow aggressive suppression of noise components, they can also over-suppress the speech component and damage its acoustic features. As a result, the performance of speech processing systems can be limited, even if a high signal-to-noise ratio (SNR) is obtained. To overcome this limitation, this chapter deals with the problem of restoring the missing components of over-masked spectrograms.

One conventional missing component restoration approach for masked spectrograms is based on non-negative matrix factorization (NMF) [97, 37]. NMF-based methods attempt to restore missing components by assuming that the entire spectrogram can be approximated as a low-rank matrix, namely, as the product of two non-negative matrices [98]. Modeling the entire spectrogram in this way amounts to assuming that the magnitude spectrum observed at each time frame can be approximated as the sum of a limited number of spectral templates. The signal can then be reconstructed using a phase reconstruction algorithm [99]. Since spectrograms are generally redundant representations of time-domain signals, the magnitude and phase of each time-frequency slot are in fact interdependent on each other. In other words, spectrograms must satisfy a certain constraint in order to be associated with time-domain signals. [99] uses this fact as the basis for devising a phase reconstruction algorithm. This implies that we can also use this relationship as a clue to help restore the missing components of masked spectrograms. However, the performance of common phase reconstruction methods is still insufficient, and NMF-based methods also require some prior information about the target speech to be effective.

Recently, a time-domain extension of NMF called time-domain spectrogram factorization (TSF) has been proposed [24]. As the name implies, TSF performs NMF-like signal decomposition in the time domain by taking account of the intrinsically redundant structure of spectrograms. While regular NMF approximates an observed magnitude spectrogram into the sum of rank-1 spectrograms, TSF decomposes an observed time-domain signal into the sum of L signal components, such that the magnitude spectrogram of each component is as close to a rank-1 structure as possible. This chapter proposes and applies TSF to directly estimate the waveform signal such that its magnitude spectrogram can be approximated as a low-rank matrix so that missing component restoration and phase reconstruction can be performed jointly in a principled manner.

cepstral distance regularization (CDR) is a recently proposed technique used in semi-supervised NMF (SSNMF), which aims to enhance target speech in both the spectral and cepstral domains [22]. CDR does this by optimizing a combined objective function composed of an NMF-based model fitting criterion defined in the spectral domain and a Gaussian mixture model-based probability distribution defined in the mel-frequency cepstral coefficient (MFCC) domain.

We proposes a TSF-based missing component restoration method which combines the conventional methods discussed above in a novel manner. The proposed method considers; 1) cues of local dependencies of each component, which are detected using redundancy in time-frequency domain expression, and/or 2) prior information about the target speech in a feature space, in addition to cues considered by conventional NMF-based methods. The effectiveness of the proposed method is demonstrated through the experimental restoration of masked speech spectrograms which are obtained by applying ideal binary mask (IBM) filters to noisy speech. The restoration performance of the proposed TSF-based method is then compared with that of conventional NMF-based methods.

This chapter is organized as follows. NMF-based missing component restoration method is introduced and the relationship with TSF is described in Section 4.2. In Section 4.3, proposed TSF-based missing component restoration method is explained. Parameter estimation for the proposed method is derived in Section 4.4 and the experimental evaluation is reported in Section 4.5. Section 4.6 summarizes this chapter.

4.2 Related Work

4.2.1 Missing Component Restoration based on NMF

NMF can be used to approximate an observed magnitude spectrogram, interpreted as a non-negative matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{K \times M}$, as a low-rank matrix by factorizing $\mathbf{X}$ into the product of two non-negative matrices $\mathbf{H} \in \mathbb{R}_{\geq 0}^{K \times L}$ and $\mathbf{U} \in \mathbb{R}_{\geq 0}^{L \times M}$:

$$\mathbf{X} \approx \mathbf{H}\mathbf{U}. \tag{4.1}$$

This amounts to assuming that the magnitude spectrum observed at each time frame can be approximated as the sum of L basis spectra:

$$X_{k,m} \simeq \hat{X}_{k,m} = \sum_{l} H_{k,l} U_{l,m}. \tag{4.2}$$

If the magnitude spectrogram of an audio signal of interest can be assumed to have a low-rank structure, missing components in the magnitude spectrogram can be restored by fitting the NMF model (4.2) over the observable regions [98]. The time-domain signal can then be synthesized for example by using a phase reconstruction technique [99].

4.2.2 Time-domain Spectrogram Factorization (TSF)

TSF is a novel signal decomposition technique that aims to directly decompose an observed time-domain signal $\mathbf{s} \in \mathbb{R}^N$ (where N denotes the number of the samples of the entire signal) into the sum of L signal components:

$$\mathbf{s} = \sum_{l} \mathbf{s}_l, \tag{4.3}$$

such that the magnitude spectrogram of $\mathbf{s}_l$ becomes as close to a rank-1 (or low-rank) structure as possible. This idea can be formulated as an optimization problem of minimizing:

$$\mathcal{I}(\theta) = \sum_{l} \sum_{k,m} (|\psi_{k,m}^{\mathsf{H}} \mathbf{s}_l| - H_{k,l} U_{l,m})^2 + \mathcal{R}(\mathbf{U}), \tag{4.4}$$

$$\text{subject to } \sum_{l} \mathbf{s}_l = \mathbf{s}, \tag{4.5}$$

where $\mathcal{R}(\mathbf{U})$ is a sparse regularization term. $\psi_{k,m}^{\mathsf{H}}\mathbf{s}_l$ represents the time-frequency element of $\mathbf{s}_l$, i.e., a short-time Fourier transform (STFT), or Wavelet Transformation, and $\psi_{k,m} \in \mathbb{C}^N$ is a complex sinusoid windowed at time frame t_m with center frequency ω_k.

Since this method allows to directly estimate the signal components $\mathbf{s}_1, \ldots, \mathbf{s}_L$ in the time domain, the phase reconstruction procedure is implicitly involved in the spectrogram factorization process. This gives TSF its name.

4.3 Proposed Method

4.3.1 Problem Setting

Let $\mathbf{Y} \in \mathbb{C}^{K \times M}$ be an observed complex spectrogram with missing components whose components are represented as $Y_{k,m}$ and where $k \in \{1, \ldots, K\}$ and $m \in \{1, \ldots, M\}$ are indices of frequency bins, and time frames, respectively. By using Γ to denote the set of the observable time-frequency slots of $\mathbf{Y}$, here we assume that the STFT coefficients in the missing regions are zero:

$$Y_{k,m} = 0, \quad ((k,m) \notin \Gamma). \tag{4.6}$$

We would like to estimate these components so that the time-domain signal can be reconstructed.

4.3.2 Objective Function Design

We can use the TSF framework to impute missing components by considering local interdependencies of the elements of a complex spectrogram. We can also borrow the idea from the conventional NMF-based approach to estimate the magnitude part of the missing components by assuming the magnitude spectrogram to have a low-rank structure. Additionally, we use CDR to ensure that the restored spectrogram follows a pretrained distribution in the cepstral domain. Hence, we propose introducing the following objective function to be minimized:

$$\begin{aligned}\mathcal{I}(\theta) = &\sum_{(k,m)\in\Gamma} |\psi_{k,m}^{\mathsf{H}}\mathbf{s} - Y_{k,m}|^2 + \lambda_1 \sum_{k,m} \mathcal{D}_{\cdot}\left(|\psi_{k,m}^{\mathsf{H}}\mathbf{s}| \mid \hat{X}_{k,m}\right) \\ &+ \lambda_2 \sum_{(k,m)\in\Gamma} \mathcal{D}_{\cdot}\left(|Y_{k,m}| \mid \hat{X}_{k,m}\right) - \lambda_3 \mathcal{K}\left(\hat{\mathcal{X}}\right), \end{aligned} \tag{4.7}$$

where λ_1, λ_2 and λ_3 are hyperparameters that weigh the importance of the second, third and fourth terms, respectively, $\theta = \{\mathbf{s}, \mathbf{H}, \mathbf{U}\}$ is the set of parameters to be optimized, and $\mathcal{D}_{\cdot}$ is a divergence measure between non-negative arguments. Here, either squared Euclidean distance or Kullback-Leibler (KL) divergence are used. Figure 4.1 shows an overview of the proposed method, where the circled numbers correspond the individual terms of the objective function. In(4.7), the first term represents the squared error between an observed complex spectrogram and that of the estimated signal $\mathbf{s}$ over the observable regions Γ. It is important to note that $\psi_{k,m}^{\mathsf{H}}\mathbf{s}$ always satisfies the condition that all complex spectrograms must satisfy and thus the redundancy of the time-frequency representation is implicitly considered. The second term represents the error between (4.2) and the magnitude spectrogram of $\mathbf{s}$, which connects the first and the third term effects. The third term represents the error between (4.2) and the observed magnitude spectrogram. This term corresponds to the objective function of the conventional NMF-based approach. The fourth term is cepstral distance regularization term. This forces $\hat{\mathbf{X}}$ to follow the statistical distribution of target speech in the feature space domain. By optimizing the objective function, the missing components of observed spectrogram $\mathbf{Y}$ can be restored by satisfying the provided constraints.

4.4 Parameter Estimation algorithm

Here, we derive a convergence-guaranteed algorithm for minimizing (4.7) based on MM principle.

4.4.1 Update rules for $\mathbf{s}$

In (4.7), the first and second terms are related to parameter $\mathbf{s}$. When $\mathcal{D}.$ is defined as the squared Euclidean distance, as in [24], we can show:

$$\begin{aligned}
&\sum_{k,m} \mathcal{D}_{\mathrm{EU}} \left(|\psi_{k,m}^{\mathsf{H}} \mathbf{s}| \mid \hat{X}_{k,m} \right) \\
=&\sum_{k,m} \left[|\psi_{k,m}^{H} \mathbf{s}|^2 - 2|\psi_{k,m}^{H} \mathbf{s}| \hat{X}_{k,m} + \hat{X}_{k,m}^2 \right] \\
\leq&\sum_{k,m} |\psi_{k,m}^{\mathsf{H}} \mathbf{s} - \hat{X}_{k,m} a_{k,m}|^2. \qquad (4.8)
\end{aligned}$$

Here, we can use the right-hand side of this inequality as a majorizer for the second term of (4.7), where $\mathbf{a} = \{a_{k,m}\}_{k,m}$ is an auxiliary parameter. The equality holds when:

$$a_{k,m} = \frac{\psi_{k,m}^{\mathsf{H}} \mathbf{s}}{|\psi_{k,m}^{\mathsf{H}} \mathbf{s}|}. \qquad (4.9)$$

When $\mathcal{D}.$ is defined as the KL-divergence, we can show:

$$\begin{aligned}
&\sum_{k,m} \mathcal{D}_{\mathrm{KL}} \left(|\psi_{k,m}^{\mathsf{H}} \mathbf{s}| \mid \hat{X}_{k,m} \right) \\
=&\sum_{k,m} \left[|\psi_{k,m}^{\mathsf{H}} \mathbf{s}| \log \frac{|\psi_{k,m}^{\mathsf{H}} \mathbf{s}|}{\hat{X}_{k,m}} - |\psi_{k,m}^{\mathsf{H}} \mathbf{s}| + \hat{X}_{k,m} \right] \\
\leq&\sum_{k,m} \left[F_{k,m} |\psi_{k,m}^{\mathsf{H}} \mathbf{s}|^2 - 2\mathrm{Re} \left[G_{k,m}^{*} \psi_{k,m}^{\mathsf{H}} \mathbf{s} \right] + \hat{X}_{k,m} \right] + \text{const.}, \qquad (4.10)
\end{aligned}$$

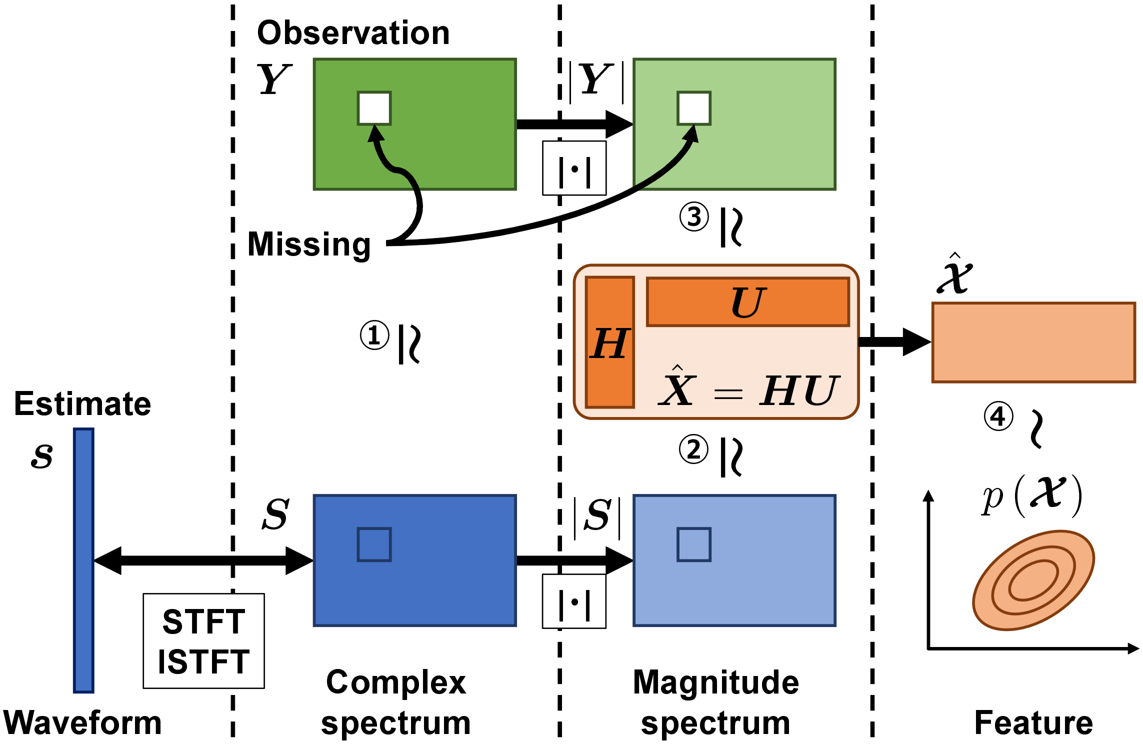


Figure 4.1: Illustration of the designed objective function, where circled numbers correspond to the terms in (4.7).

where $D_{k,m}$, $F_{k,m}$, and $G_{k,m}$ are given by:

$$D_{k,m} = \log \frac{\xi_{k,m}}{\hat{X}_{k,m}} - 2, \tag{4.11}$$

$$F_{k,m} = \begin{cases} \frac{D_{k,m}}{2b_{k,m}} + \frac{1}{\xi_{k,m}}, & (D_{k,m} \geq 0) \\ \frac{1}{\xi_{k,m}}, & (D_{k,m} < 0) \end{cases}, \tag{4.12}$$

$$G_{k,m} = \begin{cases} 0, & (D_{k,m} \geq 0) \\ -D_{k,m} a^{*}_{k,m}/2, & (D_{k,m} < 0) \end{cases}. \tag{4.13}$$

Similarly, we can use the right-hand side of this inequality as a majorizer for the case of the KL-divegence where $\mathbf{b} = \{b_{k,m}\}_{k,m}$, and $\xi = \{\xi_{k,m}\}_{k,m}$ are auxiliary parameters. The equality of (4.10) satisfies when:

$$b_{k,m} = \xi_{k,m} = |\psi^{\mathsf{H}}_{k,m}\mathbf{s}|. \tag{4.14}$$

Since both (4.8) and (4.10) are differentiable and convex, an optimal update for s minimizing (4.8) or (4.10) can be found using gradient methods. In the case of the squared Euclidean distance, parameter $\mathbf{s}$ can be efficiently updated in the following way. For the first term of (4.7), let $\tilde{Y}_{k,m}$ defined as follows:

$$\tilde{Y}_{k,m} = \begin{cases} Y_{k,m}, & ((k,m) \in \Gamma) \\ S_{k,m}, & ((k,m) \notin \Gamma) \end{cases}, \tag{4.15}$$

Since $\sum_{(k,m)\notin\Gamma} |\psi^{\mathsf{H}}_{k,m}\mathbf{s} - S_{k,m}|^2 \geq 0$, we obtain:

$$\begin{aligned} &\sum_{(k,m)\in\Gamma} |\psi^{\mathsf{H}}_{k,m}\mathbf{s} - Y_{k,m}|^2 \\ &\leq \sum_{(k,m)\in\Gamma} |\psi^{\mathsf{H}}_{k,m}\mathbf{s} - Y_{k,m}|^2 + \sum_{(k,m)\notin\Gamma} |\psi^{\mathsf{H}}_{k,m}\mathbf{s} - S_{k,m}|^2 \\ &= \sum_{k,m} |\psi^{\mathsf{H}}_{k,m}\mathbf{s} - \tilde{Y}_{k,m}|^2. \end{aligned} \tag{4.16}$$

Thus, we can also use the right-hand side of (4.16) as a majorizer where $\mathbf{S} = \{S_{k,m}\}_{k,m}$ is an additional set of auxiliary parameters. The equality of (4.15)) holds when:

$$S_{k,m} = \psi_{k,m}^{\mathsf{H}} \mathbf{s}. \tag{4.17}$$

Since this majorizer is given as a quadratic function of $\mathbf{s}$, obtain an update rule for s analytically as follows:

$$\begin{aligned}\mathbf{s} =& \frac{1}{1+\lambda_1} \left(\sum_{k,m} \mathrm{Re}[\psi_{k,m} \psi_{k,m}^{\mathsf{H}}] \right)^{-1} \\ & \times \left(\sum_{k,m} \mathrm{Re}[\psi_{k,m} (\tilde{Y}_{k,m} + \lambda_1 \hat{X}_{k,m} a_{k,m})] \right). \end{aligned} \tag{4.18}$$

Although (4.18) contains inverse matrix computation, this can be avoided by selecting $\psi_{k,m}$, so that $\sum_{k,m} \psi_{k,m} \psi_{k,m}^{\mathsf{H}}$ becomes a circulant matrix. It can be diagonalized using discrete Fourier transform matrix $\mathbf{F}$ as follows: $\sum_{k,m} \mathrm{Re}[\psi_{k,m} \psi_{k,m}^{\mathsf{H}}] = \mathbf{F} \mathbf{V} \mathbf{F}^{\mathsf{H}}$. The inverse matrix can now be calculated efficiently. For example, when $\psi_{k,m}$ represents an STFT with a square-root Hanning window, diagonal matrix $\mathbf{V}$ becomes an identity matrix.

When $\mathcal{D}.$ is defined as the a KL-divergence, the inverse matrix computation is unavoidable because the matrix to be inverted does not become a circulant matrix. Instead of trying to obtain an update rule with an analytical form, here this work chooses to update $\mathbf{s}$ be obtained using a gradient method where the gradient is given in the form:

$$\begin{aligned}\nabla_{\mathbf{s}} \mathcal{I}(\theta) =& 2 \sum_{k,m} \mathrm{Re}[\psi_{k,m} \{ (r_{k,m} + \lambda_1 F_{k,m}) \psi_{k,m}^{\mathsf{H}} \mathbf{s} \\ & - (r_{k,m} Y_{k,m} + \lambda_1 G_{k,m}) \}]. \end{aligned} \tag{4.19}$$

Here, $r_{k,m}$ is binary variable defined as:

$$r_{k,m} = \begin{cases} 1, & ((k,m) \in \Gamma) \\ 0, & ((k,m) \notin \Gamma) \end{cases}. \tag{4.20}$$

Note that terms with the form $\sum_{k,m} \mathrm{Re}[\psi_{k,m} \cdot]$ can be computed efficiently using the fast Fourier transform (FFT).

4.4.2 Update rules for H and U

In (4.7), the second, third and fourth terms are related to parameters $\mathbf{H}$, and $\mathbf{U}$. When $\mathcal{D}.$ is defined as the squared Euclidean distance, the update rules for $\mathbf{H}$, and $\mathbf{U}$ can be obtained in the same manner as the regular NMF, as follows:

$$H_{k,l} = \frac{\sum_m (\lambda_1 |\psi_{k,m}^{\mathsf{H}} \mathsf{s}| + \lambda_2 r_{k,m} |Y_{k,m}|) U_{l,m}}{\sum_m (\lambda_1 + \lambda_2 r_{k,m}) \frac{U_{l,m}^2}{\beta_{k,l,m}}}, \tag{4.21}$$

$$U_{l,m} = \frac{\sum_k (\lambda_1 |\psi_{k,m}^{\mathsf{H}} \mathsf{s}| + \lambda_2 r_{k,m} |Y_{k,m}|) H_{k,l}}{\sum_k (\lambda_1 + \lambda_2 r_{k,m}) \frac{H_{k,l}^2}{\beta_{k,l,m}}}, \tag{4.22}$$

where $\beta = \{\beta_{k,l,m}\}_{k,l,m}$ satisfies:

$$\beta_{k,l,m} = \frac{H_{k,l} U_{l,m}}{\hat{X}_{k,m}}. \tag{4.23}$$

When $\mathcal{D}.$ is defined as the KL-divergence, the update rules for $\mathbf{H}$ and $\mathbf{U}$ can be derived as in [22], as follows:

$$H_{k,l} = \frac{-\mathsf{b}_{k,l} + \sqrt{\mathsf{b}_{k,l}^2 - 4\mathsf{a}_{k,l}\mathsf{c}_{k,l}}}{2\mathsf{a}_{k,l}}, \tag{4.24}$$

$$U_{l,m} = \frac{-\mathsf{e}_{l,m} + \sqrt{\mathsf{e}_{l,m}^2 - 4\mathsf{d}_{l,m}\mathsf{f}_{l,m}}}{2\mathsf{d}_{l,m}}, \tag{4.25}$$

where $\mathsf{a}_{k,l}$, $\mathsf{b}_{k,l}$, $\mathsf{c}_{k,l}$, $\mathsf{d}_{l,m}$, $\mathsf{e}_{l,m}$, and $\mathsf{f}_{l,m}$ are defined as follows:

$$\mathsf{a}_{k,l} = \sum_{m} (\lambda_1 + \lambda_2 r_{k,m}) U_{l,m} + \lambda_3 \sum_{r,m} \left(A_{r,m} p(\zeta_{r,m}) + \frac{\delta_{B_{r,m} \geq 0} |B_{r,m}|}{\phi_{r,m}} \right) f_{r,k} U_{l,m}, \tag{4.26}$$

$$\mathsf{b}_{k,l} = -\sum_{m} (\lambda_1 |\psi_{k,m}^{\mathsf{H}} \mathbf{s}| + \lambda_2 r_{k,m} |Y_{k,m}|) \beta_{k,l,m} - \lambda_3 \sum_{r,m} \delta_{B_{r,m} < 0} |B_{r,m}| v_{r,k,l,m}, \tag{4.27}$$

$$\mathsf{c}_{k,l} = -\lambda_3 \sum_{r,m} A_{r,m} \frac{\rho_{r,k,l,m}^2}{f_{r,k} U_{l,m}}, \tag{4.28}$$

$$\mathsf{d}_{l,m} = \sum_{k} (\lambda_1 + \lambda_2 r_{k,m}) H_{k,l} l + \lambda_3 \sum_{r,k} \left(A_{r,m} p(\zeta_{r,m}) + \frac{\delta_{B_{r,m} \geq 0} |B_{r,m}|}{\phi_{r,m}} \right) f_{r,k} H_{k,l}, \tag{4.29}$$

$$\mathsf{e}_{l,m} = -\sum_{k} (\lambda_1 |\psi_{k,m}^{\mathsf{H}} \mathbf{s}| + \lambda_2 r_{k,m} |Y_{k,m}|) \beta_{k,l,m} - \lambda_3 \sum_{r,k} \delta_{B_{r,m} < 0} |B_{r,m}| v_{r,k,l,m}, \tag{4.30}$$

$$\mathsf{f}_{l,m} = -\lambda_3 \sum_{r,k} A_{r,m} \frac{\rho_{r,k,l,m}^2}{f_{r,k} H_{k,l}}. \tag{4.31}$$

δ_x is an indicator function which takes the value of one when condition x is satisfied, otherwise its value is zero. Note that $L_{r,m}$, $A_{r,m}$, $B_{r,m}$, and $p(\cdot)$ are defined as follows:

$$L_{r,m} = \sum_{k} f_{r,k} \hat{X}_{k,m}, \tag{4.32}$$

$$A_{r,m} = \sum_{p,q} \frac{\eta_{p,m} c_{q,r}^2}{2\sigma_{p,q}^2 \omega_{p,q,r,m}}, \tag{4.33}$$

$$B_{r,m} = -\sum_{p,q} \frac{\eta_{p,m} c_{q,r} \varphi_{p,q,r,m}}{\sigma_{p,q}^2 \omega_{p,q,r,m}}, \tag{4.34}$$

$$p(\zeta_{r,m}) = \frac{2 \log \zeta_{r,m}}{\zeta_{r,m}} + \frac{1}{\zeta_{r,m}^2}. \tag{4.35}$$

$\eta = \{\eta_{p,m}\}_{p,m}$, $\varphi = \{\varphi_{p,q,r,m}\}_{p,q,r,m}$, $\rho = \{\rho_{r,k,l,m}\}_{r,k,l,m}$, $\mathbf{v} = \{v_{r,k,l,m}\}_{r,k,l,m}$, $\zeta =$

$\{\zeta_{r,m}\}_{r,m}$, and $\phi = \{\phi_{r,m}\}_{r,m}$ are all auxiliary parameters satisfying following relations:

$$\eta_{p,m} = \frac{w_p \prod_q \mathcal{N}(\mathcal{X}_{q,m}; \mu_{p,q}, \sigma_{p,q}^2)}{\sum_{p'} w_{p'} \prod_{q'} \mathcal{N}(\mathcal{X}_{q',m}; \mu_{p',q'}, \sigma_{p',q'}^2)}, \tag{4.36}$$

$$\varphi_{p,q,r,m} = c_{q,r} \log L_{r,m} + \omega_{p,q,r,m}(\mu_{p,q} - \mathcal{X}_{q,m}), \tag{4.37}$$

$$\rho_{r,k,l,m} = v_{r,k,l,m} = \frac{f_{r,k} H_{k,l} U_{l,m}}{\sum_{k',l'} f_{r,k'} H_{k',l'} U_{l',m}}, \tag{4.38}$$

$$\zeta_{r,m} = \phi_{r,m} = L_{r,m}, \tag{4.39}$$

and $\omega = \{\omega_{p,q,r,m}\}_{p,q,r,m}$ is an arbitrary positive constant parameter satisfying $\sum_r \omega_{p,q,r,m} = 1$.

4.5 Experimental Evaluation

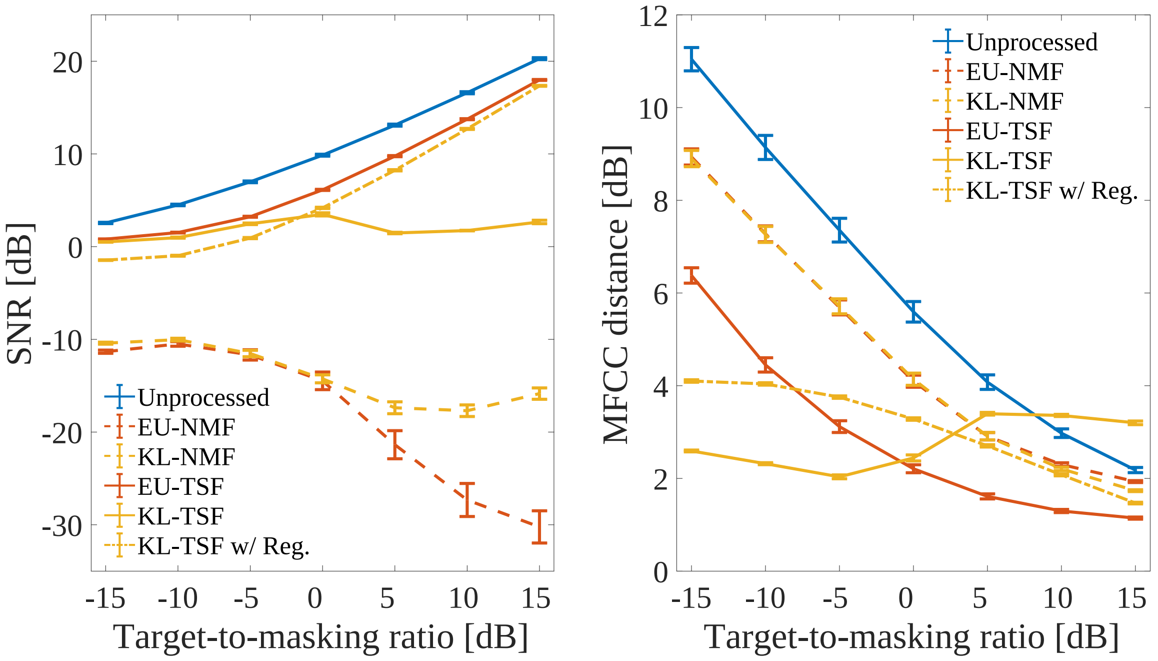
4.5.1 Experimental Settings

The performance of the proposed methods was evaluated through experiments with speech spectrograms which had been masked using IBMs for noise elimination. The masked spectrograms were prepared using IBMs constructed of clean speech and noise data, which can be represented as:

$$M_{\mathrm{IBM}} = \begin{cases} 1, & \left(10 \log_{10} \frac{|S_{k,m}^{(\mathrm{C})}|^2}{|S_{k,m}^{(\mathrm{N})}|^2} > \epsilon\right) \\ 0, & (\text{otherwise}) \end{cases}, \tag{4.40}$$

where $S_{k,m}^{(\mathrm{C})}$ and $S_{k,m}^{(\mathrm{N})}$ are complex spectrograms of clean speech and noise, respectively, and ϵ is the threshold determining whether each component activates or not. As clean data, 200 utterances of 20 speakers from ATR 503 database, including males and females were used [100]. Babble noise was added to each clean speech sample at various SNR or threshold settings in (4.40) while building the IBMs for noisy speech, and then the spectrograms of the noisy speech were masked. Three datasets were built for the experiments; *Target-to-masking ratio dataset (TMR dataset)* in which noisy speech were made under varying SNR conditions, while the corresponding IBMs were constructed at a fixed threshold parameter (0 dB), *Target-to-masking threshold dataset (TMT dataset)* in which noisy speech were made at a fixed SNR setting (0 dB), but the corresponding IBMs were constructed at varying thresholds. and a *Over-masking dataset* in which noisy speech were made at a fixed SNR setting (0 dB), and the corresponding IBMs were constructed at a fixed threshold parameter (0 dB), however, the existing components through IBM filtering were moreover erased by making them zeros randomly and compulsorily, and it was considered in more practical conditions.

Three TSF-based methods were investigated as proposed methods: a TSF using the squared Euclidean distance (EU-TSF), and the TSFs using KL-divergence with or without cepstral distance regularization (KL-TSF w/ Reg., KL-TSF w/o Reg). Two NMF-based methods using the squared Euclidean distance (EU-NMF) and the KL-divergence (KL-NMF) were used to represent conventional methods. Each speech signal was sampled at 16 kHz, and the spectrograms were obtained through frame analysis using 32 ms and 16 ms shifts with square-root Hanning windows. The total number of basis spectra was set to 30, and the total number of iterations for parameter updating was 200. Weight parameters λ_1, λ_2, λ_3 were adjusted during the first half of the iterations so that each term of the objective function in (4.7) had the same magnitude, and these weight parameters were then fixed during the second half of the iterations. For cepstral distance regularization, 0-to-13th MFCCs with 20-dimensional mel-filterbanks were extracted from 100 other utterances of the individual speakers and were for GMM training. The mixture component of the GMM was set to 30. For KL-TSF, the Adadelta technique was used for gradient descent for $\mathbf{s}$ [101]. A



(a) SNR

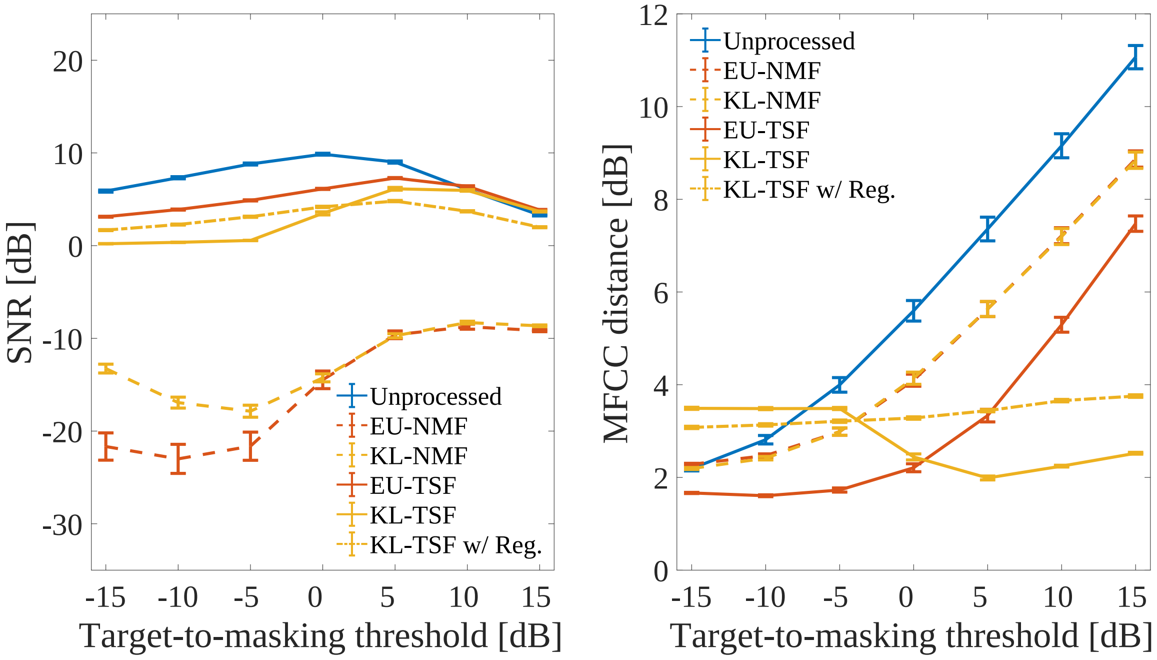
(b) MFCC distance

Figure 4.2: Results for *TMR dataset*.

phase spectrogram for the NMF-based methods was reconstructed using the Griffin-Lim algorithm with 100 iterations [99]. As measurements of performance, SNRs and the MFCC distances between the restored speech and the corresponding clean speech were used. In addition, restoration of each masked spectrogram was repeated 3 times owing to the weakening effect of the initial values, and measurements were averaged over all of the iterations and speech.

4.5.2 Experimental Results

Figures 4.2–4.4 show the SNR and the MFCC distance results for *TMR dataset*, *TMT dataset*, and *Over-masking dataset*. In Figure 4.2, the horizontal axis shows SNR



(a) SNR (b) MFCC distance

Figure 4.3: Results for *TMT dataset*.

settings, while in Figure 4.3 it represents the threshold settings of the IBMs. In Figure 4.4, the horizontal axis shows missing rate for existing components of IBMs. The vertical axes in these figures represent performance. Error bars in the figures represent 95 % confidence intervals. Unprocessed results, whose waveform signals were obtained by reproducing the masked spectrograms straightforwardly without any reconstruction methods, are also shown in each figure.

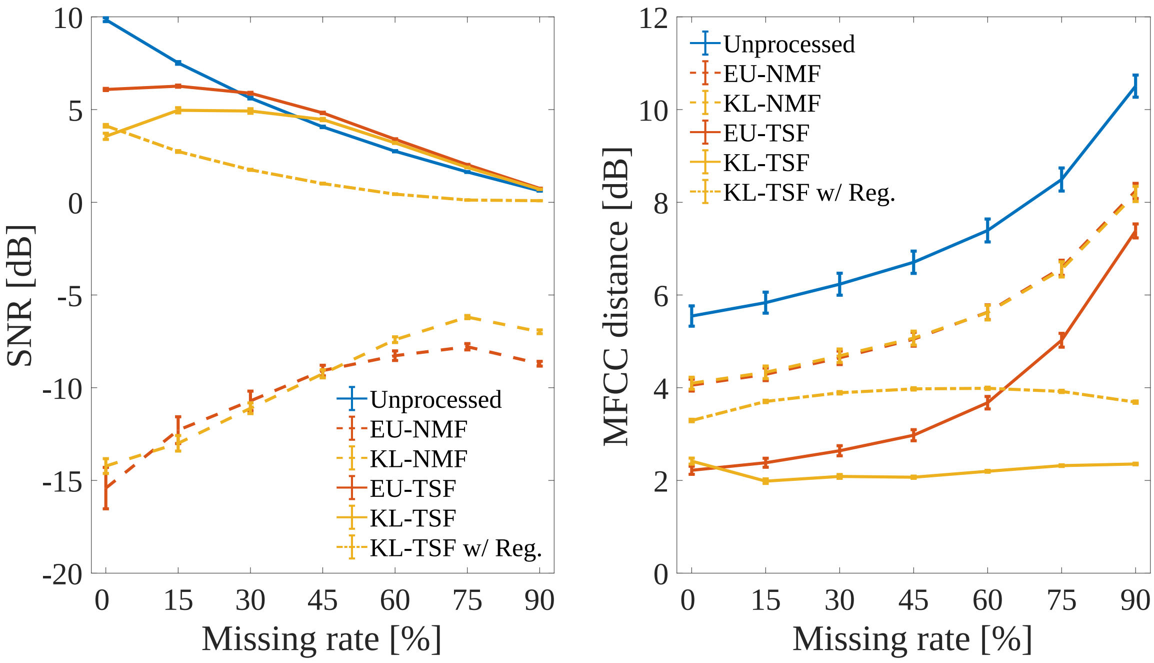
These results show that the proposed TSF-based methods outperformed conventional NMF-based methods. Especially, we can see that the SNRs of TSF-based methods follow similar tendencies to that of the unprocessed result unlike NMF-based methods. This is because that the NMF-based methods estimate not missing correct phase information, but estimate consistent phase information for the reconstructed spectrogram, which leads insufficient performance.

Moreover, the TSF-based methods maintained high SNRs similar to Unprocessed while greatly improving the MFCC distances. The proposed method using the squared Euclidean distance (EU-TSF) delivered especially stable results. In the over-suppress conditions (Figure 4.4), we can see that the proposed methods using either the squared Euclidean distance (EU-TSF) or the KL-divergence (KL-TSF) exceed unprocessed results both in the SNRs and the MFCC distances. This suggests that the proposed methods are potential to restore the missing components well by applying over-masked spectrograms, and especially KL-TSF could achieve quite robust performance for the spectrograms.

These experiments also show that cepstral distance regularization does not consistently improve restoration performance. This is because the intensity of regularization can be unsuitable and can actually degrade performance. The balancing of error functions and regularization terms has not been sufficiently researched and remains a challenging problem.

4.6 Summary

This chapter described a novel missing component restoration method for masked speech spectrograms based on a TSF signal decomposition model. The proposed method attempts to utilize as many acoustical cues as possible, e.g., cues observed in spectrograms as well as cues from spectrograms of target speech in a feature space, and, if possible, to directly estimate waveform signals. The experimental results showed that the proposed TSF-based restoration significantly outperform conventional NMF-based methods, and has potential to estimate both magnitude and phase spectra simulta-



(a) SNR

(b) MFCC distance

Figure 4.4: Results for *Over-masking dataset*.

neously and precisely. These results also demonstrated that a part of the TSF-based restoration methods has quite robust performance for over-suppressing occurring in time-frequency masking.

5 Multichannel Source Separation Based on a Deep Generative Model

This chapter deals with a multichannel audio source separation problem under underdetermined conditions. MNMF is a powerful method for underdetermined audio source separation, which adopts the NMF concept to model and estimate the power spectrograms of the sound sources in a mixture signal. This concept is also used in independent low-rank matrix analysis (ILRMA), a special class of the MNMF formulated under determined conditions. While these methods work reasonably well for particular types of sound sources, one limitation is that they can fail to work for sources with spectrograms that do not comply with the NMF model. To address this limitation, an extension of ILRMA called the multichannel variational autoencoder (MVAE) method was recently proposed, where a conditional VAE (CVAE) is used instead of the NMF model for expressing source power spectrograms. This approach has performed impressively in determined source separation tasks thanks to the representation power of deep neural networks. While the original MVAE method was formulated under determined mixing conditions, we propose a generalized version of it by combining the ideas of MNMF and MVAE so that it can also deal with underdetermined cases. We call this method the generalized MVAE (GMVAE) method. In underdetermined source separation and speech enhancement experiments, the proposed method performed better than baseline methods.

5.1 Introduction

Blind source separation (BSS) refers to the problem of separating out underlying source signals present in observed mixture signals received by a microphone array. A frequency-domain method is typically used to tackle BSS problems for convolutive mixtures by using various models for source signals and/or array responses. For example, an extension of ICA [7] called IVA [9, 10] makes it possible to jointly perform frequency-wise source separation and permutation alignment by assuming that the magnitudes of the frequency components originating from the same source are likely to vary coherently over time.

Other methods involve multichannel extensions of NMF [14, 11, 15, 13, 102, 103]. NMF is a dimension reduction method for matrices consisting of only non-negative entries. In audio signal processing, NMF was originally applied for music transcription and monaural source separation tasks [37, 43], where the power spectrogram (or the magnitude spectrogram) of a mixture signal is regarded as a non-negative matrix to be approximated as the product of two non-negative matrices. This can be viewed as approximating the power spectrum (or the magnitude spectrum) of a mixture signal observed at each time frame by the sum of a fixed number of basis spectra scaled by time-varying magnitudes.

MNMF is a method that extends the NMF so that it can additionally use spatial information for source separation. It can also be seen as a frequency-domain BSS method that uses spectral templates as clues for jointly performing frequency-wise source separation and permutation alignment. MNMF was originally formulated as a method [14] for handling underdetermined as well as determined scenarios in which sources can outnumber microphones. A determined version of MNMF, focused on solving BSS problems in determined settings, was subsequently proposed [11]. While the determined version of MNMF is applicable only to determined cases, it provides a significantly faster algorithm than the general version. This determined MNMF framework was later called independent low-Rank matrix analysis (ILRMA) [104]. It is worthwhile to note that the optimization algorithms for MNMF and ILRMA are guaranteed to converge to a stationary point, and work reasonably well for some types of sound sources. However, they can fail to work when encountering sound sources with spectrograms that do not follow the NMF model, resulting in performance limitations.

To address these limitations, new methods using variational autoencoders (VAEs) [28] have been proposed as alternatives to NMF-based source modeling [26, 27, 63, 64, 86, 87, 88, 89, 90, 91]. A VAE is a type of generative neural network capable of modeling high-dimensional data such as images. The idea of these methods is to use a VAE to model the spectra of source signals. Some of these methods [88, 89, 90] were designed to deal with speech enhancement tasks by modeling the spectrogram of a particular source to be enhanced using a regular VAE and expressing the spectrograms of the other sources using the NMF model. This allows these methods to handle semi-supervised scenarios in which interference sources are unseen in the training set. We hereafter refer to this type of method as “VAE-NMF”. Another VAE-based method worth noting is the multichannel VAE (MVAE) method [26, 27, 63, 64]. This method is an extension of ILRMA with the difference being that a conditional VAE (CVAE) [29] instead of the NMF model is used as a generative model of source spectrograms. By training the CVAE using the spectrograms of class-labeled speech samples, the resulting decoder can be used as a generative model of the speech spectrograms of multiple speakers

Table 5.1: Categorization of proposed and conventional methods.

Method	Mixing condition	Source model
ILRMA [11, 13, 103]	Determined	NMF
MNMF [14, 15, 102]	Underdetermined	NMF
MVAE [26, 27]	Determined	VAE
GMVAE (Proposed)	Underdetermined	VAE

where its inputs are interpreted as the model parameters to be optimized. Thanks to the ability of a VAE to accurately represent spectrograms, the MVAE method consistently performed better than ILRMA in determined source separation tasks.

While the original MVAE method was formulated under determined mixing conditions, we propose a generalized version of the original MVAE method by combining the ideas of MNMF and the MVAE method so that it can also deal with underdetermined cases. We call this method the generalized MVAE (GMVAE) method to distinguish it from the MVAE method (Table 5.1). The remainder of this chapter is organized as follows. In Section 5.2, we review ILRMA and the MVAE method and show that the relationship between MNMF and the GMVAE method corresponds to that between ILRMA and the MVAE method. In Section 5.3, we discuss the development of a convergence-guaranteed parameter optimization algorithm for the GMVAE method by combining the ideas for the parameter optimization processes introduced in MNMF and the MVAE method. In Section 5.4, we experimentally show the superiority of the GMVAE method over MNMF in underdetermined source separation tasks and over VAE-NMF in semi-supervised speech enhancement tasks.

5.2 Multichannel Variational Autoencoder (MVAE) Method

This chapter follows and uses the same notation descibed in Chapter 2. ILRMA is a method to solve determined source separation problems, which can be treated as a special class of MNMF. Unlike MNMF, which uses the mixing system shown in (2.6), ILRMA uses the following separation system:

$$\mathbf{s}(f,n) = \mathbf{W}^{\mathsf{H}}(f)\mathbf{x}(f,n), \tag{5.1}$$

assuming the mixing matrix is invertible. The inverse matrix $\mathbf{W}^{\mathsf{H}}(f) = \left[\mathbf{w}_1^{\mathsf{H}}(f), \ldots, \mathbf{w}_J^{\mathsf{H}}(f)\right]^{\mathsf{H}} \in \mathbb{C}^{J\times I}$ is called the separation matrix.

As with MNMF, the MM-based update equations for $\mathcal{H}_1$ and $\mathcal{U}_1$ or for $\mathcal{B}$, $\mathcal{H}_2$ and $\mathcal{U}_2$ are obtained as closed-form expressions. The separation matrix $\mathbf{W}^{\mathsf{H}}(f)$ can be updated using a fast update rule called iterative projection (IP) [12], originally developed for IVA.

One limitation of the MNMF framework including ILRMA is that it can fail to work for sources with spectrograms that are difficult to express using the NMF model given by (2.10) or (2.11). To overcome the limitation, the MVAE method has been proposed [26, 27]. The MVAE method is an improved variant of ILRMA that replaces (2.10) with a CVAE. The MVAE method models the generative model of the complex spectrogram of a particular sound source using a CVAE with an auxiliary input, indicating the classes of a source, which is represented as a one-hot vector.

The optimization algorithm of the MVAE method consists of updating the separation matrices using IP, the global scale using the MM algorithm and the inputs to the pretrained decoder using backpropagation. The advantage of using the MVAE method is that it can leverage the strong representational power of a VAE for modeling the

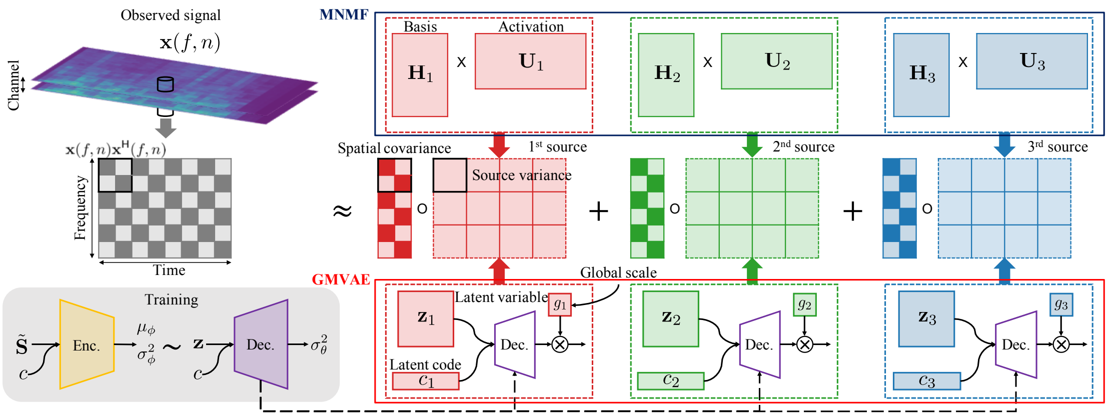


Figure 5.1: Illustration of the modeling concepts of MNMF and GMVAE method. Network parameters to be optimized at training time and parameters of the NMF and CVAE source models to be optimized at separation (inference) time are in colored blocks.

power spectrogram of sources.

5.3 Generalized MVAE (GMVAE) method

5.3.1 Overview

Figure 5.1 illustrates the modeling concepts of MNMF and the GMVAE method. These methods share the same log-likelihood (2.8) to maximize, which can be interpreted as the similarity between the outer product of each observed signal vector $\mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n)$ and the sum of full-rank spatial covariances scaled by source variances (2.12). As this figure shows, while MNMF represents source spectrograms using the NMF model, the GMVAE method represents them using a trained CVAE decoder network. Note that with the GMVAE method, we treat the spatial covariance $\mathbf{R}_j(f)$ in the same manner as MNMF. At separation (inference) time, the network parameters are fixed at the pretrained values for all the assumed sources and decoder inputs, namely the latent variable $\mathbf{z}_j$, latent code c_j, and global scale g_j become the parameters to be estimated.

5.3.2 CVAE Pretraining

CVAE consists of an encoder network and decoder network, which we train using class-labeled training examples prior to separation. Given a source spectrogram $\tilde{\mathbf{S}}$ with the one-hot encoded class label c, the encoder distribution $q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c)$ is expressed as a Gaussian distribution:

$$q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c) = \prod\nolimits_d \mathcal{N}(z(d)|\mu_\phi(d;\tilde{\mathbf{S}}, c), \sigma_\phi^2(d;\tilde{\mathbf{S}}, c)), \tag{5.2}$$

where $\mathbf{z}$ denotes a latent variable, and $z(d)$, $\mu_\phi(d;\tilde{\mathbf{S}}, c)$, and $\sigma_\phi^2(d;\tilde{\mathbf{S}}, c)$ represent the d–th elements of $\mathbf{z}$, $\mu_\phi(\tilde{\mathbf{S}}, c)$, and $\sigma_\phi^2(\tilde{\mathbf{S}}, c)$, respectively. The decoder distribution $p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g)$ is expressed as a zero-mean complex Gaussian distribution, i.e., the LGM:

$$p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g) = \prod\nolimits_{f,n} \mathcal{N}_{\mathbb{C}}(s(f,n)|0, v(f,n)), \tag{5.3}$$

$$v(f,n) = g\sigma_\theta^2(f,n;\mathbf{z}, c), \tag{5.4}$$

where $\sigma_\theta^2(f,n;\mathbf{z}, c)$ represents the (f,n)–th element of the decoder output $\sigma_\theta^2(\mathbf{z}, c)$ and g is the global scale of the generated spectrogram. During CVAE training, both the encoder and decoder network parameters ϕ and θ are trained using the following objective function:

$$\mathcal{J}(\phi, \theta; \tilde{\mathbf{S}}, c) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c)}[\log p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c)] - \mathrm{KL}[q_\theta(\mathbf{z}|\tilde{\mathbf{S}}, c)||p(\mathbf{z})], \tag{5.5}$$

where $p(\mathbf{z})$ is a standard Gaussian distribution and KL$[\cdot||\cdot]$ is the Kullback-Leibler divergence.

The trained decoder distribution $p_\theta(\tilde{\mathbf{S}}|\mathbf{z},c,g)$ can be used as a generative model capable of generating spectrograms of all the sources involved in the training examples.

5.3.3 Parameter Estimation

Since the decoder distribution is designed to be of the same form as the LGM, using $p_\theta(\tilde{\mathbf{S}}_j|\mathbf{z}_j,c_j,g_j)$ leads to the same log-likelihood as (2.8). Thus, we can derive an iterative algorithm for estimating $\mathcal{Z}=\{\mathbf{z}_j\}_j$, $\mathcal{C}=\{c_j\}_j$, $\mathcal{G}=\{g_j\}_j$ and $\mathcal{R}$ in the same manner as the derivation of an MM algorithm for MNMF.

As shown in a previous study [105], we can build a majorizer $\mathcal{L}^+$ for the negative log-likelihood function $\mathcal{L}=-\log p(\mathcal{X}|\mathcal{A},\mathcal{V})$ using the right side of the following inequality:

$$\begin{aligned}\mathcal{L}&=-\log p(\mathcal{X}|\mathcal{A},\mathcal{V})\\&\overset{c}{\leq}\sum\nolimits_j\sum\nolimits_{f,n}\Bigg[\frac{\mathrm{tr}(\mathbf{X}(f,n)\mathbf{P}_j(f,n)\mathbf{R}_j^{-1}(f)\mathbf{P}_j(f,n))}{g_j\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)}\\&\qquad+g_j\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\mathbf{Q}^{-1}(f,n)\mathbf{R}_j(f))\Bigg],\end{aligned}\tag{5.6}$$

where $\overset{c}{\leq}$ denotes the inequality that holds when constant terms are ignored. The equality holds when the auxiliary variables $\mathcal{P}=\{\mathbf{P}_j(f,n)\}_{j,f,n}$ and $\mathcal{Q}=\{\mathbf{Q}(f,n)\}_{f,n}$ are given by

$$\mathbf{P}_j(f,n)=g_j\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathbf{R}_j(f)\times\left(\sum\nolimits_j g_j\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathbf{R}_j(f)\right)^{-1},\tag{5.7}$$

$$\mathbf{Q}(f,n)=\hat{\mathbf{X}}(f,n).\tag{5.8}$$

An iterative algorithm that consists of minimizing this majorizer with respect to $\mathcal{Z}$, $\mathcal{C}$, $\mathcal{G}$, and $\mathcal{R}$ and updating $\mathcal{P}$ and $\mathcal{Q}$ using (5.7) and (5.8) is guaranteed to not increase the negative log-likelihood $\mathcal{L}$. The optimal update of $\mathcal{R}$ is obtained by (2.16). Since the majorizer is split into source-wise terms, $\mathcal{Z}$ and $\mathcal{C}$ can be updated in parallel using backpropagation. Since the sum-to-one constraints for c_j must be taken into account, this can be easily implemented by inserting an appropriately designed softmax layer that outputs c_j:

$$c_j = \mathrm{softmax}(e_j), \tag{5.9}$$

and treating e_j as the parameter to be estimated instead. The optimal update of $\mathcal{G}$ is obtained as follows:

$$g_j \leftarrow g_j \sqrt{\frac{\sum_{f,n} \sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}{\sum_{f,n} \sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}}.$$

5.3.4 Regularization of z and c

In CVAE pretraining, the encoder is trained so that the distribution of the latent variable $\mathbf{z}$ becomes close to a standard Gaussian distribution. Thus, to let the trained decoder produce spectrograms that resemble those seen in the training data, $\mathbf{z}$ must not deviate from the assumed distribution. To prevent $\mathbf{z}$ from deviating from a standard Gaussian distribution, we consider introducing regularization for $\mathbf{z}_j$ given by

$$\mathcal{L}_{\mathcal{Z}} = -\sum_j \log p(\mathbf{z}_j), \tag{5.10}$$

where $p(\mathbf{z}_j) = \mathcal{N}(\mathbf{z}_j; \mathbf{0}, \mathbf{I})$.

For the optimization of the latent code c, the resulting $c_1, \ldots, c_J$ must be disjoint since the class of each source is usually different. To promote the orthogonality between $c_1, \ldots, c_J$, we use the following regularization term:

$$\mathcal{L}_{\mathcal{C}} = \|\mathbf{C}\mathbf{C}^{\mathsf{T}} - \mathbf{I}\|_1, \tag{5.11}$$

where $\mathbf{C} \in [0,1]^{J\times L}$ is a matrix composed of J latent codes (L-dimensional vectors) and $\mathbf{I} \in \mathbb{R}^{J\times J}$ is an identity matrix. This regularization term plays the role of encouraging each latent code c_j to become a different one-hot vector.

Algorithm 1 Fully informed GMVAE

```
Train $\phi$ and $\theta$ with (5.5)
for each $j$ do
    Fix $c_j$ at a specific one-hot vector
end for
Initialize $\mathcal{Z}$, $\mathcal{G}$, and $\mathcal{R}$
repeat
    Update $\mathcal{Z}$ with (5.6) using backpropagation
    Update $\mathcal{G}$ using (5.10)
    Update $\mathcal{R}$ using (2.16)
until converge
```

Thus, the objective function for $\mathcal{Z}$ and $\mathcal{C}$ is given as

$$\mathcal{I} = \mathcal{L}^{+} + \lambda_{\mathcal{Z}}\mathcal{L}_{\mathcal{Z}} + \lambda_{\mathcal{C}}\mathcal{L}_{\mathcal{C}}, \tag{5.12}$$

where $\lambda_{\mathcal{Z}} \geq 0$ and $\lambda_{\mathcal{C}} \geq 0$ are weight parameters.

5.3.5 Advantages Over Conventional Work

The GMVAE method has several important advantages. First, it provides the flexibility of allowing it to adapt to different scenarios. A typical case is that in which we know which sources are present in a mixture. In this case, we can simply fix c_j at the corresponding one-hot vector and run the iteration (Algorithm 1). Another case is that in which we are given no information about the sources. It may appear that the GMVAE method works only in supervised and informed scenarios where audio samples of all the sources in a test mixture are included in the training set. However, thanks to the CVAE-based source modeling, if the training set contains a wide enough

Algorithm 2 Uninformed GMVAE

```
Train $\phi$ and $\theta$ with (5.5)
for each $j$ do
    Initialize $c_j$ at a uniform distribution
end for
Initialize $\mathcal{Z}$, $\mathcal{G}$, and $\mathcal{R}$
repeat
    Update $\mathcal{Z}$ and $\mathcal{C}$ with (5.6) using backpropagation
    Update $\mathcal{G}$ using (5.10)
    Update $\mathcal{R}$ using (2.16)
until converge
```

variety of sources, the GMVAE method can work in nearly blind settings where there is no information about which of the sources are present in a test mixture and can even handle sources that are unseen in the training set. For such cases, one simple way would be to treat c_j as a free parameter, initialized for example at a uniform distribution, i.e., $[1/L, \dots, 1/L]$, and run the iteration until convergence (Algorithm 2). For semi-supervised speech enhancement scenarios where only the source to be enhanced is known, we can simply specify (instead of having it estimate) one of the latent codes (Algorithm 3).

Second, the CVAE modeling can potentially have a certain effect in avoiding local optima problems in supervised and semi-supervised scenarios. One possible situation in these scenarios that can lead to poor local optima is when the source index pre-assigned to each $v_j(f,n)$ is different from the source to which the estimate of $\mathbf{R}_j(f)$ corresponds most closely. Once this kind of mismatch occurs, it usually becomes difficult to avoid getting stuck in incorrect local optima. This is one of telling examples of the problem

Algorithm 3 Partially informed GMVAE

```
Train $\phi$ and $\theta$ with (5.5)
Initialize $c^{\mathrm{Target}}$ at a specific one-hot vector
Initialize $c^{\mathrm{Non-target}}$ at a uniform distribution
Initialize $\mathcal{Z}$, $\mathcal{G}$, and $\mathcal{R}$
repeat
    Update $\mathcal{Z}$ and $c^{\mathrm{Non-target}}$ with (5.6) using backpropagation
    Update $\mathcal{G}$ using (5.10)
    Update $\mathcal{R}$ using (2.16)
until converge
```

that is very likely to occur when the source index is pre-specified for each j. It should be noted that supervised MNMF and VAE-NMF fall into this type of method. With the GMVAE method, however, we can take a soft-decision approach by treating c_j as a free parameter (instead of specifying it), initialized as a uniform distribution, and let the algorithm find the best c_j so that the distribution of the source to which $\mathbf{R}_j(f)$ is likely to correspond can be estimated along with $\mathbf{R}_j(f)$. We can then determine the index $\hat{j}$ that corresponds to the source of interest from inspection of $c_1, \dots, c_J$ and forcing $c_{\hat{j}}$ to the corresponding one-hot vector during the iteration (Algorithm 4).

5.4 Experimental Evaluation

5.4.1 Experimental Settings

We conducted three experiments to evaluate the GMVAE method. The first two are speaker-closed and speaker-open underdetermined source separation experiments where the task is to separate out three sources from their mixtures captured by two

Algorithm 4 GMVAE with one-hot enforcement

```
Train $\phi$ and $\theta$ with (5.5)
for each $j$ do
    Initialize $c_j$ with a specific one-hot vector
end for
Initialize $\mathcal{Z}$, $\mathcal{G}$, and $\mathcal{R}$
repeat
    Update $\mathcal{Z}$ and $\mathcal{C}$ with (5.6) using backpropagation
    Update $\mathcal{G}$ using (5.10)
    Update $\mathcal{R}$ using (2.16)
until converge
Determine $c_j$ which is the most similar to the target as $c^{\mathrm{Target}}$
Determine $c_{j'} (j \neq j')$ as $c^{\mathrm{Non-target}}$
Update $c^{\mathrm{Target}}$ with a specific one-hot vector
repeat
    Update $\mathcal{Z}$ and $c^{\mathrm{Non-target}}$ with (5.6) using backpropagation
    Update $\mathcal{G}$ using (5.10)
    Update $\mathcal{R}$ using (2.16)
until converge
```

microphones. The other is a semi-supervised speech enhancement experiment where the task is to extract a known source from noisy observations contaminated by unknown sources. As the experimental data, we used audio samples from the Voice Conversion Challenge (VCC) 2018 dataset [106], which contains recordings of 6 female and 6 male U.S. English speakers. The average duration of each utterance is 3.5 seconds, and the dataset includes 81 utterances of individual speakers for training and 35 utterances for evaluation. For these experiments, we used the utterances of four female and four male speakers, 'SF1', 'SF2', 'SF3', 'SF4', 'SM1', 'SM2', 'SM3', and 'SM4'. For training, we used 100 utterances of 'SF1', 'SF2', 'SM1', and 'SM2'. Another 10 utterances of 'SF1', 'SF2', 'SM1', and 'SM2' were used for evaluation under speaker-closed conditions in the source separation task and treated as the target sources in the speech enhancement task. Similarly, 10 utterances of 'SF3', 'SF4', 'SM3', and 'SM4' were used for evaluation under speaker-open conditions in the source separation task and treated as the interference sources in the speech enhancement task.

Figure 5.2 shows the configuration of the room used for the experiments. Reverberation time T_{60} was set to 78 and 351 ms. In the source separation task, we created test data using all possible combinations of three speakers for both the speaker-closed and speaker-open conditions. For each set of speakers, 10 speech mixtures were generated by randomly choosing the utterances and randomly allocating them at locations indicated in Figure 5.2. In the speech enhancement task, 40 speech mixtures were generated by randomly choosing the utterances of the target and interference speakers where target and interference sources are located at the Src. 1 and Src. 2, respectively.

We tested several different versions of the proposed and baseline methods for comparison. We use the terms "fully supervised/semi-supervised/unsupervised" and "fully informed/partially informed/uninformed" to properly categorize each version of the

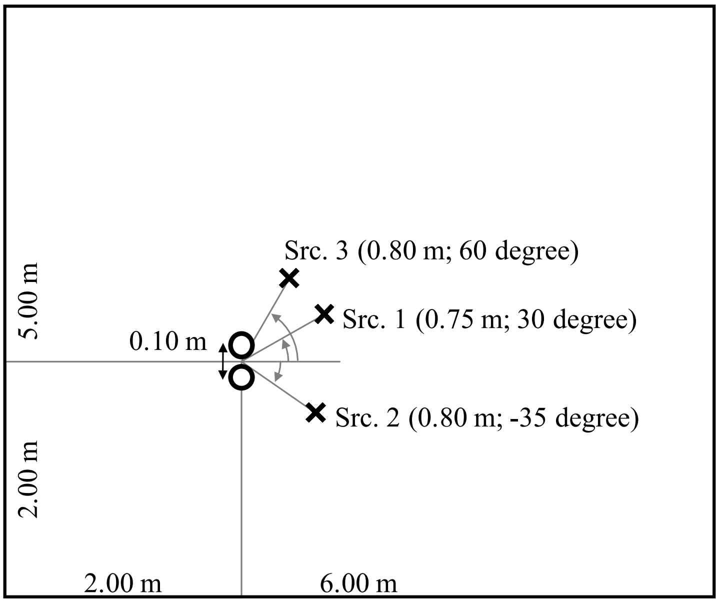

Src. 3 (0.80 m; 60 degree)
Src. 1 (0.75 m; 30 degree)
Src. 2 (0.80 m; -35 degree)
0.10 m
5.00 m
2.00 m
2.00 m
6.00 m

Figure 5.2: Configuration of room used for our experiments, where ○ and × are locations of microphones and sound sources, respectively.

methods. Fully supervised, semi-supervised, and unsupervised refer to whether a method requires training examples and fully informed, partially informed, and uninformed refer to how much information about which sources are present in a test mixture is given to a method. All versions of the GMVAE method are fully supervised since they all require training examples to train the CVAE. Thus, we omit "fully supervised" when referring to this method. At separation time, the GMVAE method can be implemented in either fully informed, uninformed, or partially informed manners. Hence, we refer to these versions as fully informed GMVAE, uninformed GMVAE, and partially informed GMVAE. MNMF can perform in either unsupervised, semi-supervised, or fully supervised manners. We implemented unsupervised uninformed, fully supervised uninformed, and fully supervised fully informed MNMFs for comparison. VAE-NMF falls into the semi-supervised partially informed category. Categorization of each version is summarized in Table 5.2.

All the speech signals were resampled at 16 kHz. We tested two different STFT configurations, i.e., a 128-ms window length with a 64-ms shift length and a 256-ms window length with a 128-ms shift length. The numbers of basis spectra for these baseline versions were set to 10 per speaker, as in a previous study [14]. The spectral dictionaries used for the fully/semi-supervised MNMF versions were trained for each speaker using the same dataset used for the CVAE training and obtained using an Itakura-Saito NMF (IS-NMF) [43] with 1000 iterations. For a fair comparison, MNMF was run for 200 iterations for the initialization of each method. All the versions, including the baseline ones, were then run for 100 iterations. For the speech enhancement task, we implemented Algorithm 4, which consists of updating $c_1, \ldots, c_J$ freely during the first 50 iterations, then searching for the index $\hat{j}$ that corresponds to the target speaker, and finally running the last 50 iterations while fixing $c_{\hat{j}}$ at the corresponding

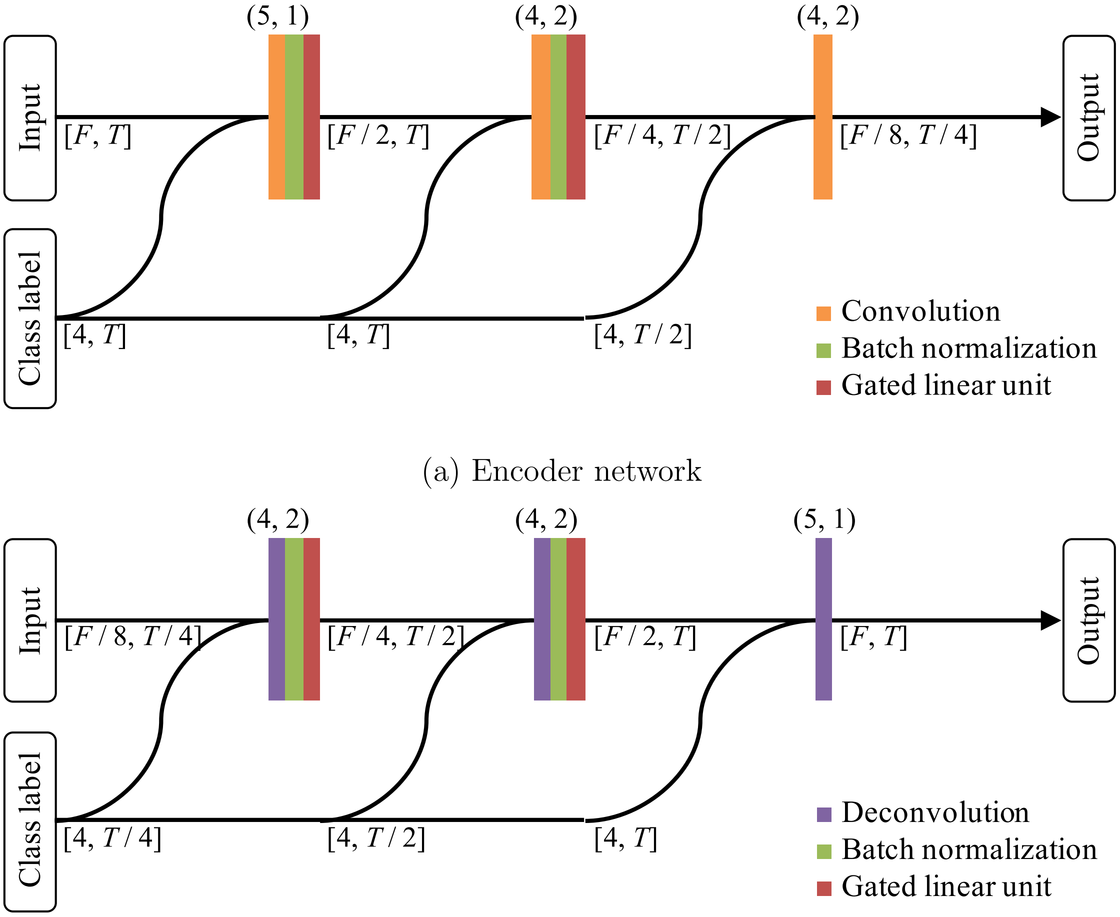
Table 5.2: Methods for comparison.

(a) Source separation task

Notation	Method	Initialization
Baseline1	Unsupervised uninformed MNMF [15] (Equation (2.10))	–
Baseline2	Unsupervised uninformed MNMF [15] (Equation (2.11))	–
Baseline3	Fully-supervised uninformed MNMF [15]	–
Baseline4	Fully-supervised fully-informed MNMF [15]	–
Proposed1	Uninformed GMVAE (Algorithm 2)	Baseline1
Proposed2	Uninformed GMVAE (Algorithm 2)	Baseline2
Proposed3	Uninformed GMVAE (Algorithm 2)	Baseline3
Proposed4	Fully informed GMVAE (Algorithm 1)	Baseline4

(b) Speech enhancement task

Notation	Method	Initialization
Baseline5	Semi-supervised partially-informed MNMF [15]	–
Baseline6	VAE-NMF [89]	Baseline5
Proposed5	Partially informed GMVAE (Algorithm 3)	Baseline5
Proposed6	GMVAE with one-hot enforcement (Algorithm 4)	Baseline5



(a) Encoder network

(b) Decoder network

Figure 5.3: Network configurations of (a) encoder and (b) decoder, where [c, t] denotes input channel and input length. Both convolution and deconvolution represent 1-dimensional operation. (k, s) represent kernel size and stride size along frame, respectively.

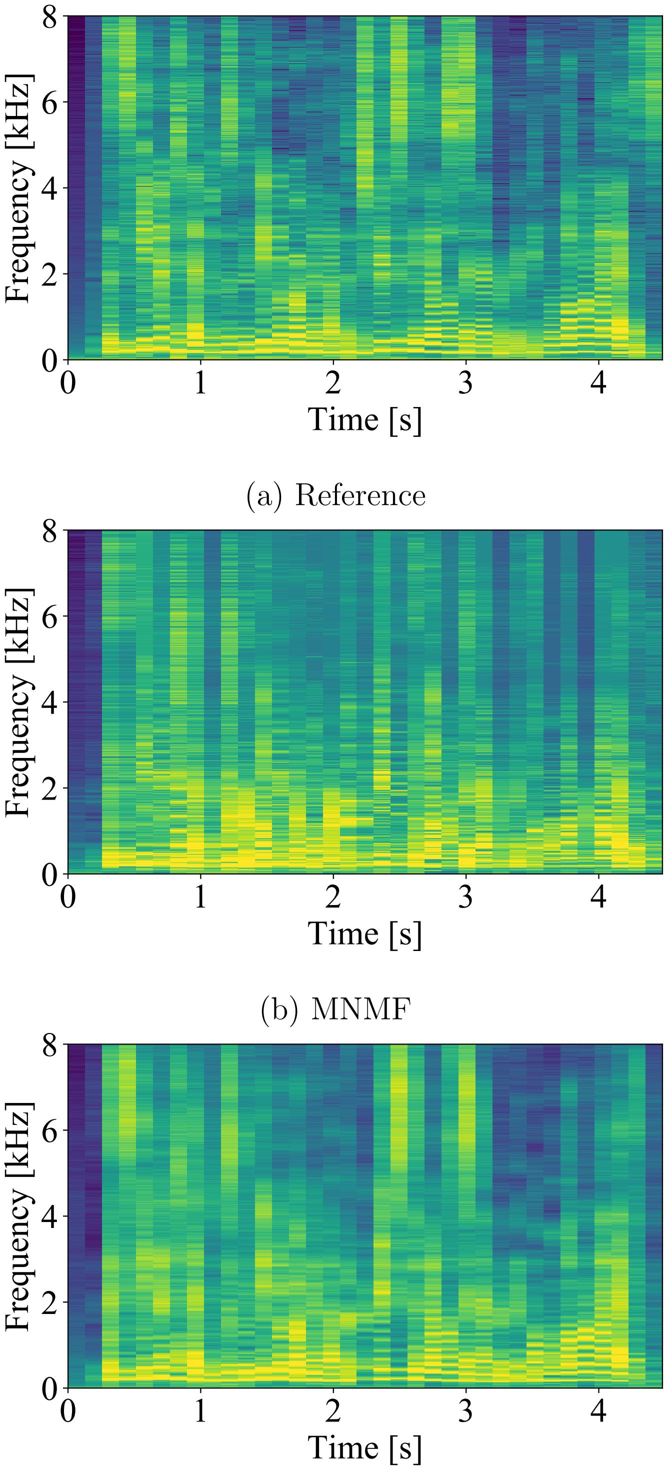
one-hot vector. We refer to this algorithm as "GMAVE with one-hot enforcement". The encoder and decoder networks of the CVAE are shown in Figure 5.3. At training time, the batch size and length were set to 9 and 128, respectively. The Adam algorithm [107] with a learning rate of 0.0001 was used for the CVAE pretraining. The number of training epochs was set to 1000. The VAE used with VAE-NMF was trained for each speaker using the same training dataset and training configuration, where the same network architectures as the CVAE except for the conditioning part were used. At separation time, the Adam algorithm with a learning rate of 0.01 was used for updating $\mathcal{Z}$ and $\mathcal{C}$. The number of training epochs per iteration was set to 10.

As the evaluation metrics, we used the averages of the signal-to-distortion ratio (SDR), source image-to-spatial distortion ratio (ISR), signal-to-inference ratio (SIR), and signal-to-artifact ratio (SAR) [92] between the reference signals and separated signals. Note that, in the speech enhancement task, separation performances of both the target source and interference source were evaluated and permutation of estimated sources was not considered in the evaluation.

5.4.2 Experimental Results

Figure 5.4 (b) and (c) show examples of the NMF- and CVAE-based source models fitted to the speech spectrogram shown in Figure 5.4 (a). As these examples show, the CVAE source model was able to express harmonic structures and higher-frequency components better than the NMF model.

We next show the performances in the source separation task. A comparison of the separation performance of each version under speaker-closed conditions is shown in Figure 5.5, where error bars show the 95 % confidence intervals. When comparing the performance of the uninformed versions (Baseline1 to Baseline3 and Proposed1



(a) Reference

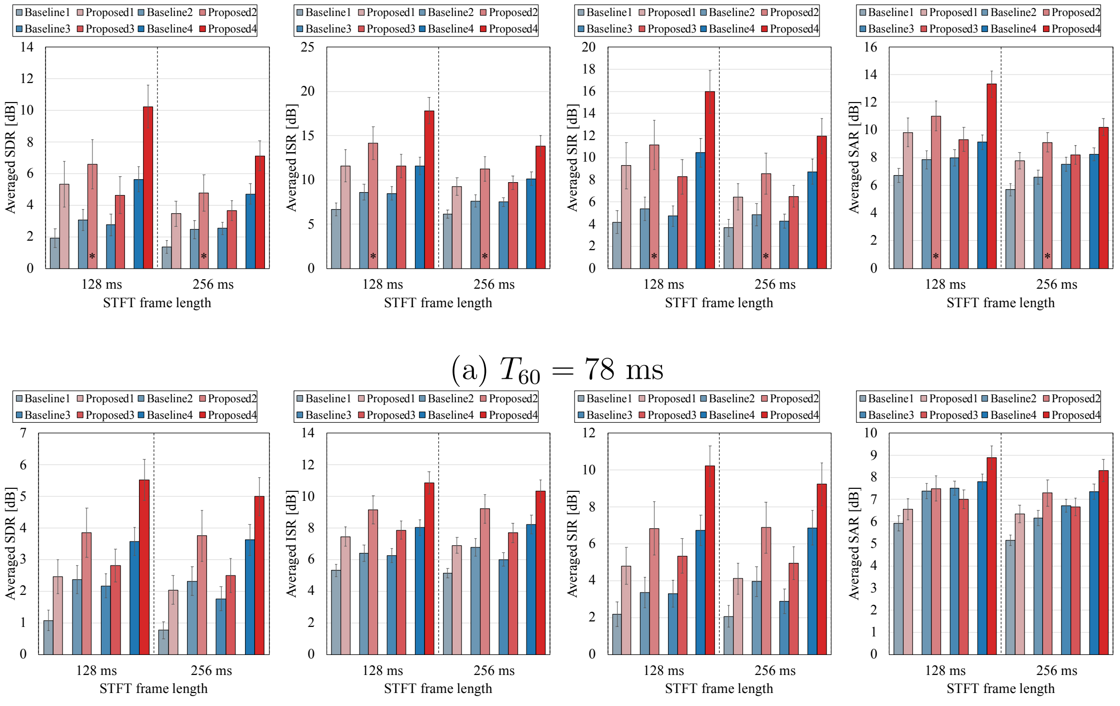
(b) MNMF

(c) GMVAE

Figure 5.4: Spectrograms of (a) reference source and estimated sources by using (b) MNMF and (c) GMVAE.

to Proposed3) at $T_{60} = 78$ ms, the proposed versions outperformed the baseline ones for both STFT configurations. The comparison of Baseline3 and Proposed3 directly reflects the difference in ability between the NMF- and CVAE-based source models. The results thus indicate the superiority of the CVAE source model over the NMF counterpart. The comparison between Proposed1, 2 &, 3 indicates that initialization can affect separation performances. It indicates that using Baseline2 for initialization worked better than using Baseline1 & 3. Focusing on the comparison of the fully informed versions (Baseline4 and Proposed4), Proposed4 significantly outperformed Baseline4 and achieved the best performance. This indicates that the prior information for the sources in a target mixture can contribute to improving performance. Although the performances of all of the versions degraded for the longer reverberant condition ($T_{60} = 351$ ms), the proposed versions still performed better than the baseline ones. The comparisons between the performances obtained with the two STFT configurations showed that using a 128-ms frame length worked better, especially for a shorter reverberant condition.

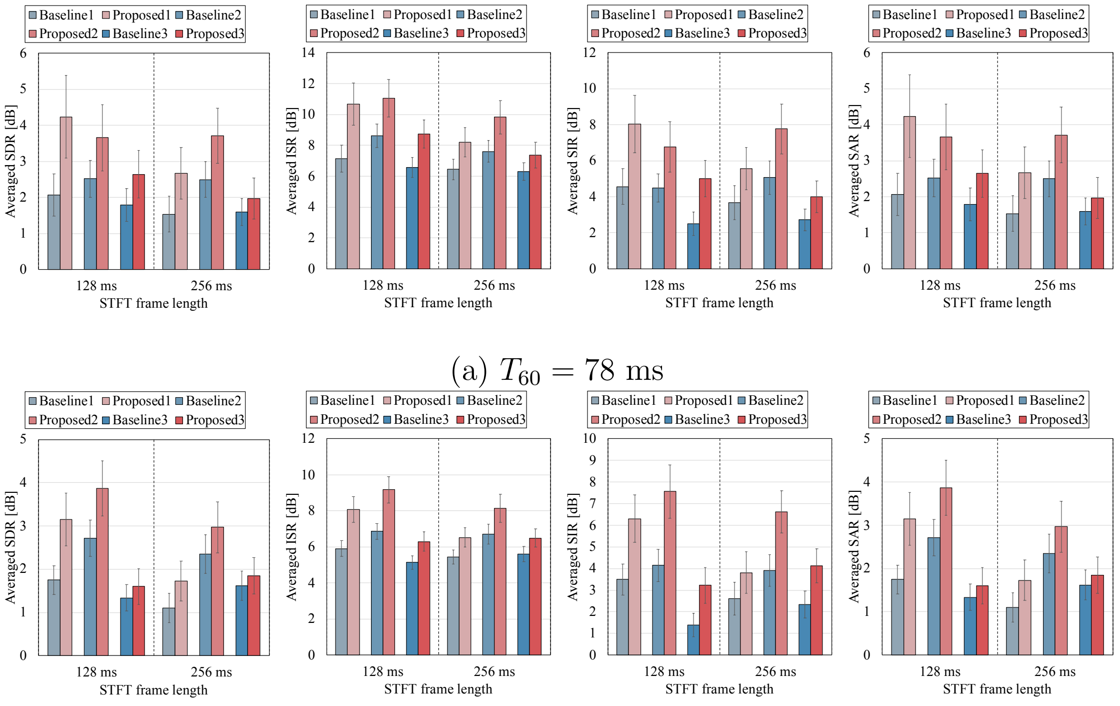
A comparison of the separation performance of each method under speaker-open conditions is shown in Figure 5.6, where the fully informed versions (Baseline4 and Proposed4) are omitted since all the sources in the mixture are unseen in the training data. We can confirm from the comparisons between Baseline1 and Proposed1, Baseline2 and Proposed2, and Baseline3 and Proposed3 that the proposed versions consistently performed better than the baseline ones, especially in terms of the SIR metric. This may imply the ability of the GMVAE method to estimate the spectrogram of each source accurately, leading to an accurate estimation of its spatial covariance. Another interesting finding from these results is that the GMVAE method can perform reasonably well under speaker-open conditions even though it is a method that requires



(a) $T_{60} = 78$ ms

(b) $T_{60} = 351$ ms

Figure 5.5: Source separation performances under speaker-closed conditions.



(a) $T_{60} = 78$ ms

(b) $T_{60} = 351$ ms

Figure 5.6: Source separation performances under speaker-open conditions.

supervisions. We also confirmed that unlike under the speaker-closed conditions, using a 128-ms STFT frame length was more robust against varying reverberation conditions than using longer frame lengths.

Table 5.3 shows an ablation study on Proposed2, where the best performances are denoted in bold font and the last columns correspond to the separation performances denoted as $*$ in Figure 5.5. These results indicate that each regularization technique improved the separation performance, and Proposed2 using both regularizations achieved the best performance. These results also indicate that the regularizations were effective, especially when the STFT frame length was 128 ms. Figure 5.7 shows examples of the estimated $\mathcal{Z}$ and $\mathcal{C}$ without and with the regularizations, where the histograms represent $\mathcal{Z}$ at initialization step and separation step. Estimated $\mathcal{C}$ is also shown in the figure. We can confirm that the regularization for $\mathcal{Z}$ prevented $\mathcal{Z}$ from deviating from a standard Gaussian distribution, and the regularization for $\mathcal{C}$ promoted the orthogonality of $\mathcal{C}$.

We finally show the performances of the speech enhancement task. A comparison of the enhancement performances of Baseline5, Baseline6, Proposed5, and Proposed6 is shown in Figure 5.8. Comparisons among Baseline5, Baseline6, and Proposed5 revealed that Proposed5 outperformed the baseline versions and performed better than VAE-NMF. Moreover, Proposed6 performed better than the other versions particularly under the small reverberant condition. This shows a certain effect of the one-hot enforcement process adopted in Algorithm 4.

5.5 Summary

We proposed the GMVAE method, a generalized version of the MVAE method that can also deal with underdetermined cases. We developed a convergence-guaranteed

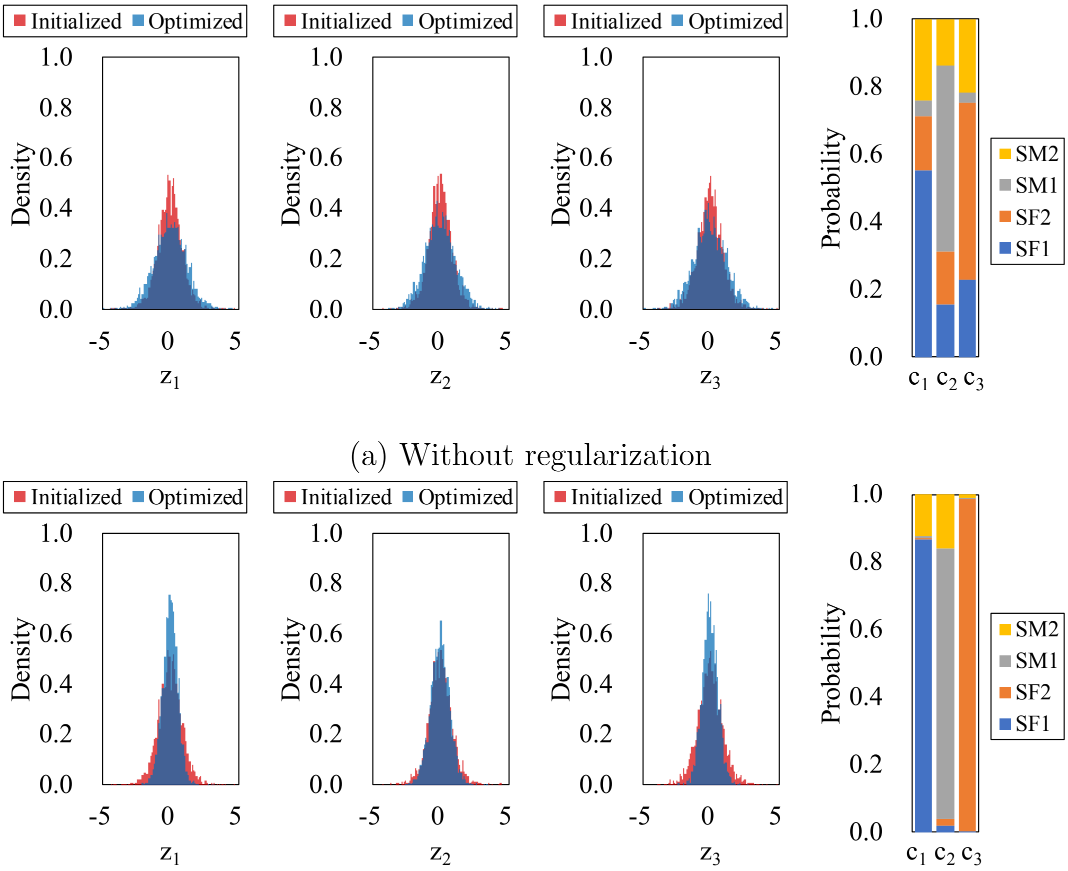
Table 5.3: Ablation study on Proposed2 under speaker-closed conditions at $T_{60} = 78$ [ms].

(a) 128-ms STFT frame length

$\mathcal{L}_{\mathcal{Z}}$	$\mathcal{L}_{\mathcal{C}}$	Avg. SDR	Avg. ISR	Avg. SIR	Avg. SAR
✗	✗	5.93	13.45	10.34	10.45
✗	✓	6.13	13.69	10.50	10.62
✓	✗	6.43	14.02	10.99	10.78
✓	✓	**6.59**	**14.16**	**11.14**	**11.02**

(b) 256-ms STFT frame length

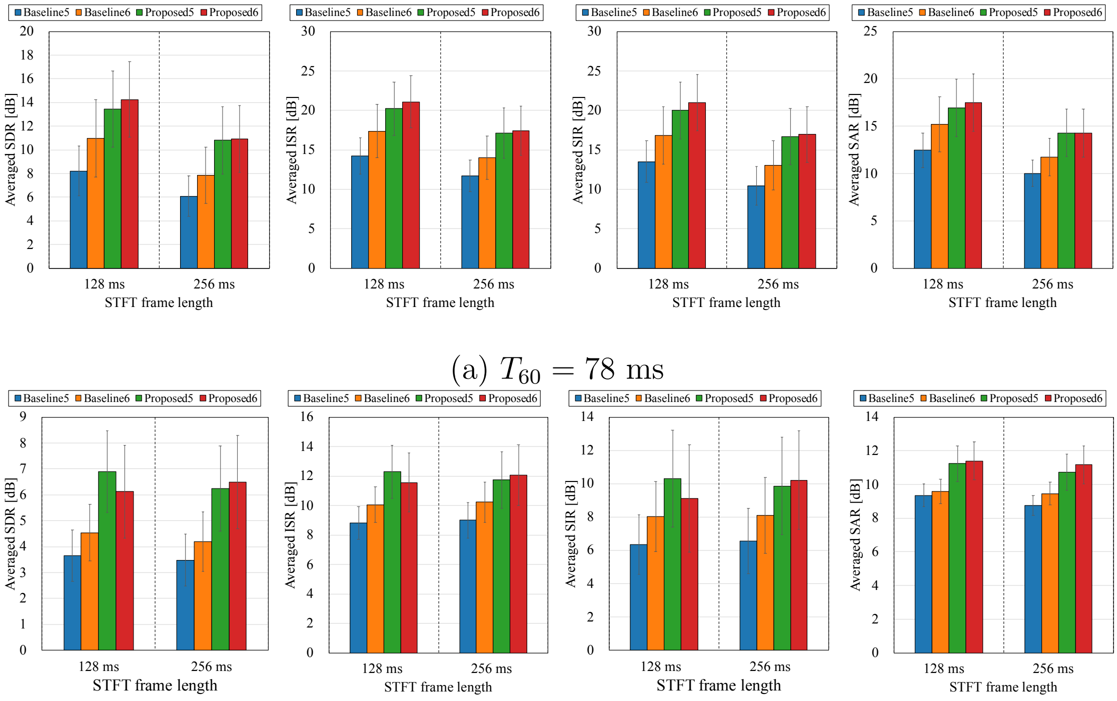
$\mathcal{L}_{\mathcal{Z}}$	$\mathcal{L}_{\mathcal{C}}$	Avg. SDR	Avg. ISR	Avg. SIR	Avg. SAR
✗	✗	4.46	10.80	8.12	8.73
✗	✓	4.56	10.92	8.24	8.84
✓	✗	4.63	11.08	8.41	8.95
✓	✓	**4.78**	**11.24**	**8.56**	**9.11**



(a) Without regularization

(b) With regularization

Figure 5.7: Regularization effects on $\mathcal{Z}$ and $\mathcal{C}$.



(a) $T_{60} = 78$ ms

(b) $T_{60} = 351$ ms

Figure 5.8: Speech enhancement performances.

parameter optimization algorithm for the GMVAE method by combining the ideas for the parameter optimization processes introduced in MNMF and the MVAE method. We further introduced two regularization techniques for avoiding undesirable solutions and presented several algorithms designed for fully informed, partially informed, and uninformed source separation and speech enhancement tasks. Our experimental results revealed that the proposed GMVAE method outperformed MNMF in source separation tasks and VAE-NMF in speech enhancement tasks, demonstrating the advantage of the CVAE source model. The results also indicate that the GMVAE method can perform reasonably well even under speaker-open conditions.

6 Relationship with Real-World Data Circulation

Real-world data circulation (RWDC) is the multidisciplinary study of the flows of acquisition, analysis and implementation data, and the circulation of these flows, which is a key to the successful development of commercial services and products. This chapter reviews the analysis and use of these flows in RWDC, and describes how source separation problems and RWDC are related. We also discuss how source separation can contribute to creating new value.

6.1 Introduction

The development of new commercial services and products begins with the collection of real-world data about the needs and desires of potential customers. This data is analyzed and the output is used to refine these new services and products. Once launched, feedback from customers about the new applications is also collected, thus there should be certain flows of data circulating during each phase of development. RWDC is a multidisciplinary study to consider flows of acquisition, analysis and implementation data, as well as the circulation of these flows.

An overview of the concept of RWDC is shown in Figure 6.1. In the data acquisition phase, various phenomena in real world are acquired in the form of digital data through observations. In the data analysis phase, information technologies such as pattern

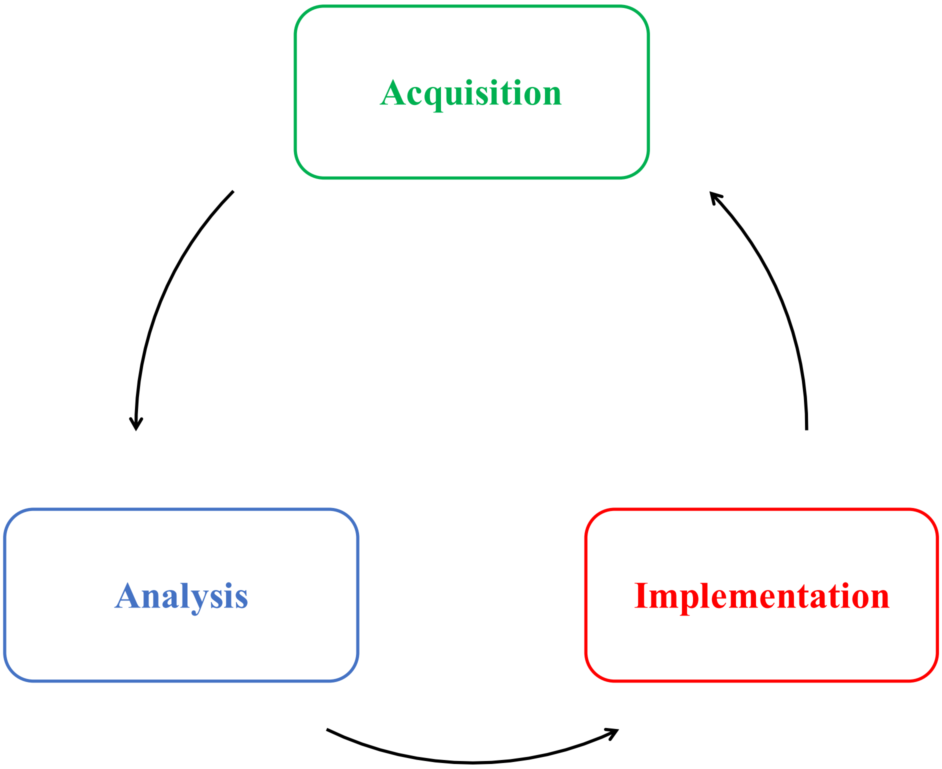


Figure 6.1: Overview of Real-World Data Circulation. The acquisition, analysis, and implementation boxes represent the phases of the RWDC process. Various real-world phenomena, such as digital data and observations, are acquired. This data is then analyzed using information technology. Changes are then implemented, resulting in new or improved products and services.

recognition and machine learning techniques are used to reveal the characteristic or structure of the data. In the data implementation phase, the output results of the data analysis are used to create new services and products. Thus, iteratively repeating these flows, new services and products can be developed which reflect users’ demands, creating new social value. From a value creation point of view, two connections between RWDC and the source separation techniques proposed in this dissertation are discussed in this chapter. One is data circulation within the source separation problem, and the other is source separation during data circulation. The rest of this chapter is organized as follows. In Section 6.2, we discuss what kind of data circulation can be expected in source separation problems. Section 6.3 describes, in terms of data circulation, how source separation can contribute to the creation of new value. This chapter is then summarized in Section 6.4.

6.2 Data Circulation in Source Separation

Source separation is a method of dividing a mixture of audio signals recorded by microphones into the individual source signals of its components, a process which is closely related to data circulation. The method of supervised source separation addressed in this dissertation is an improved method that uses training data to obtain as much prior information about the source signals in a mixed signal as possible into account, which results in better separation performance. This use of prior knowledge for signal separation can be viewed as a type of analysis of the input mixture signals. The outputs of source separation are the estimated source signals, which are generally used in subsequent systems. The goal is to transmit clean signal data to the following systems, and this flow of processed data represents the implementation phase of a data circulation system. Moreover, if the estimated signals are sufficiently separated,

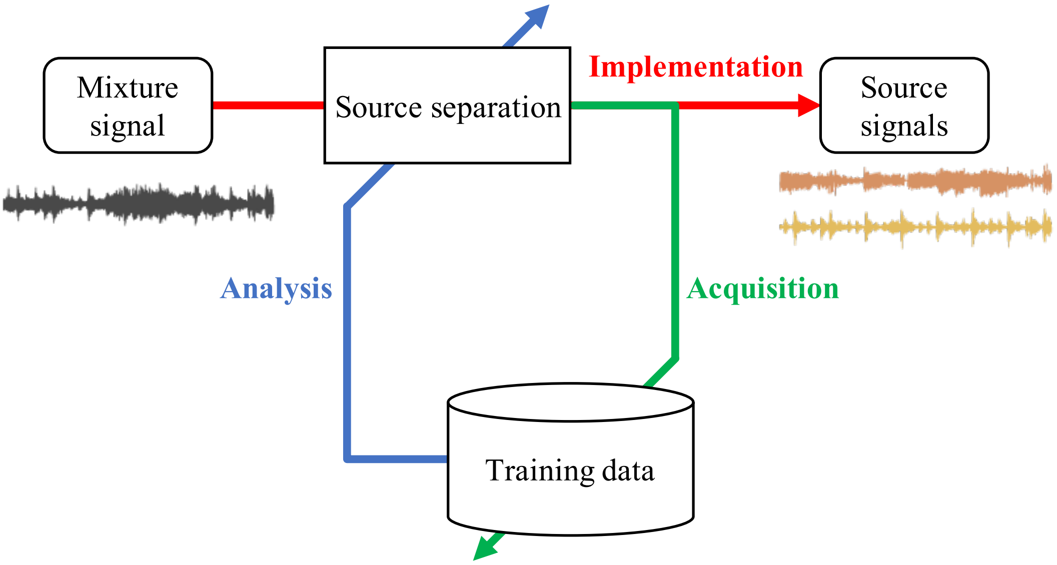


Figure 6.2: Data circulation in (supervised) source separation.

these output signals can be used as additional training data. Consequently, due to this increase in training data, the source separation process is expected to be further improved. Thus, as depicted in Figure 6.2, the source separation process itself is an example of data circulation.

6.3 Value Creation through Source Separation

As described in Chapter 1, source separation has been studied and developed in order to replicate complex human hearing functions. As a result, source separation has great potential for replacing or augmenting human hearing. Figure 6.3 illustrates a cyclic flow of human hearing activity being improved through the use of source separation. The output separated signals can be used in various kinds of hearing enhancement applications, allowing people to acquire additional abilities, i.e., functional

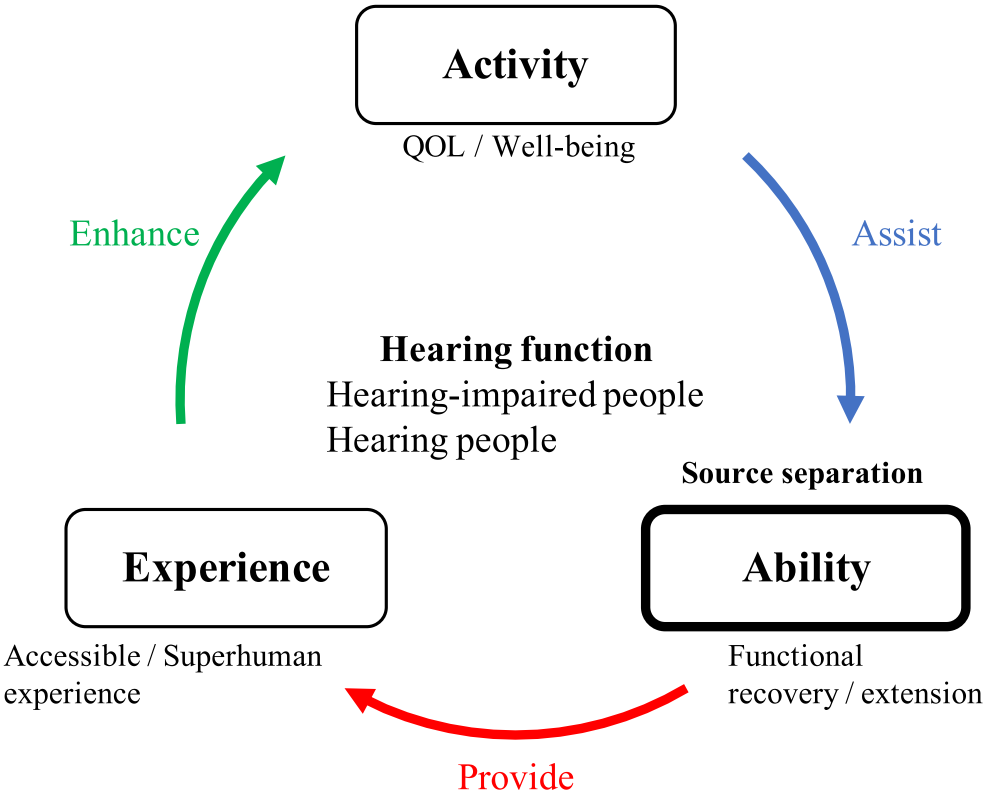


Figure 6.3: Value creation through source separation. Separation of audio signal data can improve the hearing abilities of the public.

recovery for hearing-impaired persons or functional extension for people with normal hearing. Thanks to their improved abilities, users of these applications can enjoy new experiences, e.g., improved hearing allows people with impaired hearing to function more normally and while enhanced hearing for unimpaired users provides superhuman hearing experiences. Ultimately, these experiences enhance the hearing abilities and functioning of each user, which provides the public with improved QOL and a higher level of wellness.

6.4 Summary

In this chapter I have explained the concept of RWDC and reviewed the components of RWDC. Data circulation during source separation, and source separation as a component of RWDC processes for providing improved products and services, i.e., how source separation can contribute to value creation, were also discussed.

7 Conclusion

7.1 Summary of this Dissertation

Source separation, which is the task of separating individual source signals out of a mixture of signals, has been used to develop various useful applications in combination with other technologies, and has the potential to continue to contribute to new advances. Supervised learning, which considers prior knowledge about the sources in a mixture, is a useful way to improve the performances of underdetermined source separation methods, although it is not helpful if there is an insufficient amount of prior information available to be considered. Moreover, depending on how the source signals are extracted, errors in the estimated signals can degrade the performance of subsequent systems.

This dissertation has addressed two problems encountered when using supervised approaches for underdetermined source separation. One is how to improve the use of prior information for the processing of stereophonic music. The other is how to improve source models to achieve better separation performance.

In Chapter 2, source separation methods were categorized according to several variables, such as how many sources and microphones are present, the methods that are used, and whether or not training data is available. Fundamental components in separation algorithms were then described. Supervised source separation methods were also described and various methods were reviewed.

In Chapter 3, a method of improving source separation performance when processing stereophonic music signals, which consist of various synthesized source signals, was proposed, using a supervised source separation method. To model unrealistic mixing of music signals, a non-negative matrix factorization (NMF)-based method employing the amplitude spectrograms of a mixture was introduced. Furthermore, a cepstral distance regularization (CDR) method was incorporated into the proposed model to regularize the timbre information of the sources. Experimental results revealed that, when compared with a conventional supervised method, the proposed method yielded significant improvements, providing better estimation of the synthesizing parameters.

In Chapter 4, a new method of restoring the missing components of separated signals extracted from time-frequency masking, to prevent negative effects on subsequent systems, was proposed, which is based on time-domain spectrogram factorization (TSF). The proposed method allows us to take into account the local interdependencies of the elements of a complex spectrogram derived from the redundancy of a time-frequency representation, in addition to the global structure of the magnitude spectrogram considered in conventional methods. Experimental results demonstrated that the proposed method significantly outperformed conventional methods, and has the potential to estimate both phase and magnitude spectra simultaneously and precisely.

In Chapter 5, a new supervised source separation method was proposed in order to overcome the fundamentally unsatisfactory performance of supervised MNMF methods when used for underdetermined source separation. The proposed method is a generalized version of the multi-channel variational autoencoder (MVAE), in which a conditional VAE (CVAE) is used instead of the NMF model used for expressing source power spectrograms in ILRMA. The proposed method, which is referred to as generalized MVAE (GMVAE), can successfully perform underdetermined source separation.

Through experimental evaluations using source separation and speech enhancement tasks, the proposed method demonstrated better performance than baseline

In Chapter 6, real-world data circulation (RWDC), which is a key process used to develop commercial services and products, is described. Each of of the components of RWDC were reviewed, and examples of how data circulation and source separation are related were provided. How source separation can contribute to creating new value through RWDC was also explained.

7.2 Future Work

Although the methods proposed in this dissertation for addressing each of the challenges described demonstrated improvements in performance, several challenges remain to be addressed.

7.2.1 Investigation of the Performances for Real-World Data

Thanks to open dataset of impulse responses [108, 109, 110, 111, 112] and the room impulse response simulator [113], evaluation of the proposed methods under simulated conditions were easily performed. However, looking toward the development of practical applications, it will be necessary to evaluate these separation methods using real-world data.

For the task of music source separation, to the best of my knowledge, few datasets provide the original source signals recorded with microphones [114]. Moreover, the goal of source separation generally targets source signals after they have been subjected to various forms of processing. To clarify the performance of existing source separation methods, building a dataset which contains such original music source signals anre evaluation using such a dataset is needed.

7.2.2 Comparison with Generative and Discriminative Approaches

With the development of a number of machine learning techniques, there are various ways to handle source separation problems. As described in Chapter 2, source separation methods can be divided into two approaches, i.e., generative and discriminative approaches. While various separation methods are evaluated within each approach, there are few literatures to evaluate separation methods with different approaches [115]. Although there exists difficulties to compare different approaches depending on assumptions how much amount of data is available, it should be compared and clarified.

7.2.3 Acceleration of Separation Algorithms

Source separation methods based on generative approach employ the algorithm iteratively updating source model parameters and spatial covariance parameters. Some of such separation algorithms ensure convergence to a stationary point, however, it requires computational time to perform sufficient performances especially for underdetermined source separation. As some acceleration methods have been proposed recently [74], toward developing practical applications, the development of accelerated algorithms should be required.

7.2.4 Source Separation for Specific Purposes

Source separation methods have been studied and developed so that the separated signals are more distinguishable. However, when considering incorporation with other technologies such as automatic speech recognition (ASR) and voice conversion (VC), source separation methods can affect on subsequent systems as addressed in Chapter 4. This can be occurred if the objectives of source separation do not match with those of subsequent systems. Thus, source separation methods for specific purposes should be developed. End-to-end processing is one of ways to alleviate these gaps. As several studies have recently proposed end-to-end approaches in ASR, it would be worthwhile developing source separation methods with specific purposes.

Acknowledgements

I would like to express the deepest appreciation to my thesis advisor, Professor Kazuya Takeda, Nagoya University, for his thoughtful guidance and continuous encouragement throughout my bachelor's, master's and doctor's courses.

I would like to express my sincere gratitude to Professor Tomoki Toda, Nagoya University, for his fruitful comments and constant support throughout my master's and doctor's courses. He has provided me the basics of research and various opportunities to collaborate with other researchers and students. I have been happy to work on my research with him. Without his support, this thesis would not have been materialized.

I would also like to express my heartfelt appreciation to Dr. Hirokazu Kameoka, Distinguished Researcher at NTT Communication Science Laboratories, for insightful comments and continual collaboration throughout my master's and doctor's courses. The core of this thesis originated from his ingenious and pioneering works in source separation, which greatly helped me to develop new methods. He has offered me to work for NTT Communication Science laboratories as intern. I have enjoyed many opportunities to discuss and research with him and I have been able to improve my research skills. This thesis would not have been accomplished without his support.

I would like to express my gratitude to Professor Hiromi Nakaiwa, Nagoya University, Associate Professor Hiroaki Kudo, Nagoya University, and Associate Professor Eijiro Takeuchi, Nagoya University, for their invaluable comments to the thesis.

I would sincerely like to thank Dr. Jonathan Le Roux, Senior Principal Research Scientist at Mitsubishi Electric Research Laboratories (MERL), for giving me to work for MERL as guest and providing productive and continued discussion about source separation.

I would like to offer my special thanks to Professor Takanori Nishino, Meijo University, Associate Professor Yoshio Ishiguro, Nagoya University, Associate Professor Chiyomi Miyajima, Daido University, Associate Professor Kazuyoshi Suzuki, Yamagata University, Senior Lecturer Akira Tamamori, Aichi Institute of Technology, Assistant Professor Alexander Carballo Segura, Nagoya University, Assistant Professor Kento Ohtani, Nagoya University, Assistant Professor Keisuke Fujii, Nagoya University, and Dr. Reo Iijima for their supportive comments and suggestions.

I would like to offer my great thanks to the staff of Takeda Laboratory, Toda Laboratory, and RWDC Leading Program for their generous support. I would like to offer my special thanks to Dr. Kazuhiro Kobayashi, CEO at TARVO Co., Ltd., and Dr. Tomoki Hayashi, COO at Human Dataware Lab. Co., Ltd., for their technical advise and comments. I would also like to thank colleagues in Takeda Laboratory, and Toda Laboratory for their kind collaborations and open-minded discussions.

Finally, I would like to acknowledge my family and friends for their devoted support and encouragement.

Reference

[1] M. Cooke, G. J. Brown, M. Crawford, and P. Green, "Computational auditory scene analysis: Listening to several things at once," *Endeavour*, vol. 17, no. 4, pp. 186–190, 1993.

[2] M. Cooke and G. J. Brown, "Computational auditory scene analysis: Exploiting principles of perceived continuity," *Speech Communication*, vol. 13, no. 3-4, pp. 391–399, 1993.

[3] C. Jutten and J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal processing*, vol. 24, no. 1, pp. 1–10, 1991.

[4] F. Jelinek, *Statistical methods for speech recognition.* MIT press, 1997.

[5] H. Møller, "Fundamentals of binaural technology," *Applied acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis.* John Wiley & Sons, 2004, vol. 46.

[8] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation.* Springer, 2007, vol. 615.

[9] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ica to multivariate components," in *International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2006, pp. 165–172.

[10] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2006, pp. 601–608.

[11] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux, and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 245–253.

[12] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[13] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1622–1637, 2016.

[14] 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.

[15] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[16] C. M. Bishop, *Pattern recognition and machine learning.* springer, 2006.

[17] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

[18] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

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

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

[21] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* John Wiley & Sons, 2009.

[22] L. Li, H. Kameoka, T. Higuchi, and H. Saruwatari, "Semi-supervised joint enhancement of spectral and cepstral sequences of noisy speech," in *Interspeech*, 2016, pp. 3753–3757.

[23] P. C. Loizou, *Speech enhancement: theory and practice.* CRC press, 2013.

[24] H. Kameoka, “Multi-resolution signal decomposition with time-domain spectrogram factorization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2015, pp. 86–90.

[25] H. Kagami, H. Kameoka, and M. Yukawa, “A majorization-minimization algorithm with projected gradient updates for time-domain spectrogram factorization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2017, pp. 561–565.

[26] H. Kameoka, L. Li, S. Inoue, and S. Makino, “Semi-blind source separation with multichannel variational autoencoder,” *arXiv preprint arXiv:1808.00892*, 2018.

[27] ——, “Supervised determined source separation with multichannel variational autoencoder,” *Neural computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[28] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *International Conference on Learning Representations (ICLR)*, 2014.

[29] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, “Semi-supervised learning with deep generative models,” in *Advances in Neural Information Processing Systems*, 2014, pp. 3581–3589.

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

[31] B. Arons, “A review of the cocktail party effect,” *Journal of the American Voice I/O Society*, vol. 12, no. 7, pp. 35–50, 1992.

[32] A. R. Conway, N. Cowan, and M. F. Bunting, “The cocktail party phenomenon revisited: The importance of working memory capacity,” *Psychonomic bulletin & review*, vol. 8, no. 2, pp. 331–335, 2001.

[33] M. Kawamoto, K. Matsuoka, and N. Ohnishi, “A method of blind separation for convolved non-stationary signals,” *Neurocomputing*, vol. 22, no. 1-3, pp. 157–171, 1998.

[34] N. Murata and S. Ikeda, “An on-line algorithm for blind source separation on speech signals,” in *Proc. NOLTA98*, vol. 3, 1998, pp. 923–926.

[35] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[36] R. Scheibler and N. Ono, “Independent vector analysis with more microphones than sources,” *arXiv preprint arXiv:1905.07880*, 2019.

[37] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, no. 3, 2003, pp. 177–180.

[38] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in *International Conference on Independent Component Analysis and Signal Separation (ICA)*. Springer, 2004, pp. 494–499.

[39] M. N. Schmidt and M. Mørup, “Nonnegative matrix factor 2-d deconvolution for blind single channel source separation,” in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 700–707.

[40] S. A. Raczyński, N. Ono, and S. Sagayama, “Multipitch analysis with harmonic nonnegative matrix approximation,” in *8th International Society for Music Information Retrieval Conference (ISMIR).* Citeseer, 2007.

[41] R. Kompass, “A generalized divergence measure for nonnegative matrix factorization,” *Neural computation*, vol. 19, no. 3, pp. 780–791, 2007.

[42] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, “Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence,” in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP).* IEEE, 2010, pp. 283–288.

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

[44] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* The MIT Press, 1964.

[45] P. Smaragdis, B. Raj, and M. Shashanka, “Supervised and semi-supervised separation of sounds from single-channel mixtures,” in *International Conference on Independent Component Analysis and Signal Separation (ICA).* Springer, 2007, pp. 414–421.

[46] N. J. Bryan and G. J. Mysore, “Interactive refinement of supervised and semi-supervised sound source separation estimates,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2013, pp. 883–887.

[47] D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, “Music signal separation by supervised nonnegative matrix factorization with basis defor-

mation," in *IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2013, pp. 1–6.

[48] Aurora4-database. [Online]. Available: http://aurora.hsnr.de/aurora-4.html

[49] The 4th chime speech separation nad recognition challeng. [Online]. Available: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/

[50] The 5th chime speech separation nad recognition challeng. [Online]. Available: http://spandh.dcs.shef.ac.uk/chime_challenge/

[51] Corpus of spontaneous japanese. [Online]. Available: https://pj.ninjal.ac.jp/corpus_center/csj/en/

[52] Timit acoustic-phonetic continuous speech corpus. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1

[53] Csr-i (wsj). [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A

[54] Cambridge music technology. [Online]. Available: http://www.cambridge-mt.com/ms/mtk/

[55] Dsd100. [Online]. Available: https://sigsep.github.io/datasets/dsd100.html

[56] Musdb18. [Online]. Available: https://sigsep.github.io/datasets/musdb.html

[57] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3734–3738.

[58] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 2135–2139.

[59] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[60] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for multichannel audio source separation," in *European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1557–1561.

[61] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2018, pp. 696–700.

[62] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[63] L. Li, H. Kameoka, and S. Makino, "Fast mvae: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 546–550.

[64] S. Inoue, H. Kameoka, L. Li, S. Seki, and S. Makino, "Joint separation and dereverberation of reverberant mixtures with multichannel variational autoencoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 96–100.

[65] J. De Leeuw and W. J. Heiser, "Convergence of correction matrix algorithms for multidimensional scaling," *Geometric representations of relational data*, pp. 735–752, 1977.

[66] D. R. Hunter and K. Lange, “A tutorial on mm algorithms,” *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[67] C. Févotte and J.-F. Cardoso, “Maximum likelihood approach for blind audio source separation using time-frequency gaussian source models,” in *IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 78–81.

[68] E. Vincent, S. Arberet, and R. Gribonval, “Underdetermined instantaneous audio source separation via local gaussian modeling,” in *International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 775–782.

[69] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, “Probabilistic modeling paradigms for audio source separation,” in *Machine Audition: Principles, Algorithms and Systems.* IGI global, 2011, pp. 162–185.

[70] N. Q. Duong, E. Vincent, and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[71] N. Ito, S. Araki, and T. Nakatani, “Fastfca: Joint diagonalization based acceleration of audio source separation using a full-rank spatial covariance model,” in *European Signal Processing Conference (EUSIPCO).* IEEE, 2018, pp. 1667–1671.

[72] N. Ito and T. Nakatani, “Fastfca-as: Joint diagonalization based acceleration of full-rank spatial covariance analysis for separating any number of sources,”

in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC).* IEEE, 2018, pp. 151–155.

[73] ——, “Multiplicative updates and joint diagonalization based acceleration for under-determined bss using a full-rank spatial covariance model,” in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP).* IEEE, 2018, pp. 231–235.

[74] ——, “Fastmnmf: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 371–375.

[75] K. Yoshii, “Correlated tensor factorization for audio source separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 731–735.

[76] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning.* MIT press, 2016.

[77] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, “Improved mvdr beamforming using single-channel mask prediction networks,” in *Interspeech*, 2016, pp. 1981–1985.

[78] M. Togami, “Multi-channel itakura saito distance minimization with deep neural network,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 536–540.

[79] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2016, pp. 31–35.

[80] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Alternative objective functions for deep clustering,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 686–690.

[81] ——, “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[82] Z. Chen, Y. Luo, and N. Mesgarani, “Deep attractor network for single-microphone speaker separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[83] P. Seetharaman, G. Wichern, S. Venkataramani, and J. Le Roux, “Class-conditional embeddings for music source separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 301–305.

[84] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[85] D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” in *Advances in Neural Information Processing Systems*, 2018, pp. 10 215–10 224.

[86] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, “Statistical speech enhancement based on probabilistic integration of variational autoencoder

and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 716–720.

[87] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018, pp. 1–6.

[88] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2018, pp. 1233–1239.

[89] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 101–105.

[90] S. Leglaive, U. Şimşekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 541–545.

[91] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," *arXiv preprint arXiv:1910.10942*, 2019.

[92] 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.

[93] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *6th International Society for Music Information Retrieval Conference (ISMIR).* Citeseer, 2005, pp. 337–344.

[94] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent f0 estimation and source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2015, pp. 574–578.

[95] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on speech and audio processing*, vol. 7, no. 2, pp. 126–137, 1999.

[96] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

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

[98] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for time-frequency representations of audio signals," *Journal of signal processing systems*, vol. 65, no. 3, pp. 361–370, 2011.

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

[100] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “Atr japanese speech database as a tool of speech recognition and synthesis,” *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[101] M. D. Zeiler, “Adadelta: an adaptive learning rate method,” *arXiv preprint arXiv:1212.5701*, 2012.

[102] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, “Student’s t multichannel nonnegative matrix factorization for blind source separation,” in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[103] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, and N. Ono, “Independent low-rank matrix analysis based on complex student’s t-distribution for blind audio source separation,” in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.

[104] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation with independent low-rank matrix analysis,” in *Audio Source Separation.* Springer, 2018, pp. 125–155.

[105] H. Kameoka, H. Sawada, and T. Higuchi, “General formulation of multichannel extensions of nmf variants,” in *Audio Source Separation.* Springer, 2018, pp. 95–124.

[106] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, “The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods,” *arXiv preprint arXiv:1804.04262*, 2018.

[107] D. P. Kingma and J. L. Ba, “Adam: Amethod for stochastic optimization,” in *International Conference on Learning Representations*, 2015.

[108] Air. [Online]. Available: https://www.iks.rwth-aachen.de/fileadmin/user_upload/downloads/_forschung/tools-downloads/air_database_release_1_4.zip

[109] Ace. [Online]. Available: https://acecorpus.ee.ic.ac.uk/

[110] Reverb. [Online]. Available: https://catalog.ldc.upenn.edu/LDC95S24

[111] Rwcp. [Online]. Available: http://research.nii.ac.jp/src/en/RWCP-SSD.html

[112] But reverbdb. [Online]. Available: https://speech.fit.vutbr.cz/software/but-speech-fit-reverb-database

[113] https://pyroomacoustics.readthedocs.io/en/pypi-release/index.html. [Online]. Available: https://pyroomacoustics.readthedocs.io/en/pypi-release/index.html

[114] “Medleydb: A dataset of multitrack audio for music research.” [Online]. Available: https://medleydb.weebly.com

[115] Y. Bando, Y. Sasaki, and K. Yoshii, “Deep bayesian unsupervised source separation based on a complex gaussian mixture model,” in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP).* IEEE, 2019, pp. 1–6.

List of Publications

Journal Papers

1. Shogo Seki, Hirokazu Kameoka, Li Li, Tomoki Toda, and Kazuya Takeda, "Underdetermined Source Separation Based on Generalized Multichannel Variational Autoencoder," IEEE Access, Vol. 7, pp. 168104–168115. Nov. 2019.
2. Shogo Seki, Tomoki Toda, and Kazuya Takeda, "Stereophonic Music Separation Based on Non-negative Tensor Factorization with Cepstral Distance Regularization," IEICE Transactions on Fundamentals of Electronics, Vol. E101-A, No. 7, pp. 1057–1064, Jul. 2018.

International Conferences

1. Shogo Seki, Hirokazu Kameoka, Li Li, Tomoki Toda, and Kazuya Takeda, "Generalized Multichannel Variational Autoencoder for Underdetermined Source Separation," Proceedings of the 27th European Signal Processing Conference (EUSIPCO), pp. 1973–1977, Sep. 2019.
2. Shota Inoue, Hirokazu Kameoka, Li Li, Shogo Seki, and Shoji Makino, "Joint Separation and Dereverberation of Reverberant Mixtures with Multichannel Variational Autoencoder," Proceedings of IEEE International Conference on Acoustics,

Speech and Signal Processing (ICASSP), pp. 96–100, May 2019.

3. Moe Takada, Shogo Seki, and Tomoki Toda, “Self-produced Speech Enhancement and Suppression Method Using Air- and Body-conductive Microphones,” Proceedings of the Asia-Pacific Signal and Information Processing Association (APSIPA), pp. 1240–1245, Nov. 2018.
4. Shogo Seki, Hirokazu Kameoka, Tomoki Toda, and Kazuya Takeda, “Missing Component Restoration for Masked Speech Signals Based on Time-domain Spectrogram Factorization,” Proceedings of the 27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 6 pages, Sep. 2017.
5. Shogo Seki, Tomoki Toda, and Kazuya Takeda, “Stereophonic Music Separation Based on Non-negative Tensor Factorization with Cepstrum Regularization,” Proceedings of the 25th European Signal Processing Conference (EUSIPCO), pp. 1011–1015, Aug. 2017.
6. Shogo Seki, Tomoki Toda, and Kazuya, Takeda, “Stereo Channel Music Signal Separation Based on Non-negative Tensor Factorization with Cepstrum Regularization,” The Journal of the Acoustical Society of America, vol. 140, no. 4, pp. 2967, Nov. 2016.
7. Atsunori Ogawa, Shogo Seki, Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, and Kazuya Takeda, “Robust Example Search Using Bottleneck Features for Example-Based Speech Enhancement,” Proceedings of the 17th Annual Conference on International Speech Communication Association (Interspeech), Sep. 2016.

Domestic Conferences

1. 大竹 徹郎, 関 翔悟, 戸田 智基, “楽曲音源分離のための個別音源マスク推定ネットワークの統合法,” 日本音響学会秋季研究発表会, pp. 165–166, Sep. 2019.
2. 高田 萌絵, 関 翔悟, ルンバントビン パトリック, 戸田 智基, “空気／体内伝導音の対応関係を活用した自己発声音強調／抑圧法,” 日本音響学会秋季研究発表会, pp. 173–174, Sep. 2019.
3. 彦坂 秀, 小林 和弘, 林 知樹, 関 翔悟, 武田 一哉, 坂野 秀樹, 戸田 智基, “模擬難聴処理を活用した補聴器フィルタ設計,” 日本音響学会秋季研究発表会, pp. 567–568, Sep. 2019.
4. 関 翔悟, 亀岡 弘和, 李 莉, 戸田 智基, 武田 一哉, “多チャンネル変分自己符号化器を用いた劣決定音源分離,” 日本音響学会春季研究発表会, pp. 229–230, Mar. 2019.
5. 井上 翔太, 亀岡 弘和, 李 莉, 関 翔悟, 牧野 昭二, “多チャンネル変分自己符号化器を用いた音源分離と残響除去の統合的アプローチ,” 日本音響学会春季研究発表会, pp. 399–402, Mar. 2019.
6. 山田 智也, 関 翔悟, 小林 和弘, 戸田 智基, “楽曲中歌声加工における声質変換精度向上のための歌声・伴奏分離法,” 信号処理シンポジウム, pp. 258–263, Nov. 2018.
7. 高田 萌絵, 関 翔悟, 戸田 智基, “空気／体内伝導マイクロフォンを用いた雑音環境下における自己発声音強調／抑圧法,” 日本音響学会秋季研究発表会, pp. 173–174, Sep. 2018.
8. 関 翔悟, 林 知樹, 武田 一哉,戸田 智基, “WaveNetに基づく振幅スペクトログラムからの波形生成,” 日本音響学会秋季研究発表会, pp. 281–282, Sep. 2018.
9. Mohammad Eshghi, Shogo Seki,Kazuhiro Kobayashi,Tomoki Toda, “Electrolaryngeal Speech Enhancement by Using Attached Microphones onto Electrolarynx,” 日本音響学会秋季研究発表会, pp. 1023–1024, Sep. 2018.

10. 関 翔悟, 亀岡 弘和, 戸田 智基, 武田 一哉, “時間領域低ランクスペクトログラム近似法に基づくマスキング音声の欠損成分復元,” 日本音響学会春季研究発表会, pp. 421–422, Mar. 2017.
11. 関 翔悟, 大谷 健登, 戸田 智基, 武田 一哉, “ケプストラム正則化NTFによるステレオチャネル楽曲音源分離,” 日本音響学会秋季研究発表会, pp. 333–334, Sep. 2016.

Technical Reports

1. 彦坂 秀, 小林 和弘, 林 知樹, 関 翔悟, 武田 一哉, 坂野 秀樹, 戸田 智基, “模擬難聴処理を活用した音声波形加工に基づく明瞭度改善,” 電子情報通信学会技術研究報告, Vol. 119, No. 188, pp. 25–29, Aug. 2019.
2. 関 翔悟, 亀岡 弘和, 李 莉, 戸田 智基, 武田 一哉, “多チャンネル変分自己符号化器に基づく劣決定音源分離の評価,” 電子情報通信学会技術研究報告, Vol. 118, No. 497, pp. 323–328, Mar. 2019.
3. 高田 萌絵, 関 翔悟, 戸田 智基, “ウェアラブルな空気／体内伝導マイクロフォンを用いた自己発声音強調／抑圧法,” 電子情報通信学会技術研究報告, Vol. 118, No. 190, pp. 7–12, Aug. 2018.
4. 山田 智也, 関 翔悟, 小林 和弘, 戸田 智基, “統計的手法に基づく楽曲中の歌声加工のための歌声分離法の検討,” 電子情報通信学会技術研究報告, Vol. 117, No. 517, pp. 209–214, Mar. 2018.
5. 関 翔悟, 戸田 智基, 武田 一哉, “ケプストラム距離正則化を用いた半教師ありステレオチャネル楽曲音源分離,” 情報処理学会技術研究報告, Vol. 2017-MUS-115, No. 18, 6 pages, June 2017.
6. 山田 智也, 関 翔悟, 小林 和弘, 戸田 智基, “歌声分離ならびに統計的歌声声質変換に基づく楽曲中の歌声加工,” 情報処理学会技術研究報告, Vol. 2017-MUS-115, No. 30, 6 Pages, June 2017.

7. 関 翔悟，亀岡 弘和，戸田 智基，武田 一哉, “時間領域信号推定に基づく音声スペクトログラムの欠損成分復元,” 電子情報通信学会技術研究報告, Vol. 116, No. 475, pp. 19–24, Mar. 2017.

Awards

1. Best Student Paper Award Nominee, The 27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Sep. 2017.
2. The 20th Best Presentation Award, The Tokai Chapter of Acoustical Society of Japan, Mar. 2017.
3. The 14th Student Presentation Award, The Acoustical Society of Japan, Mar. 2017.