Chapter 8
Itakura-Saito Nonnegative Matrix Two-Dimensional Factorizations for Blind Single Channel Audio Separation

Bin Gao and Wai Lok Woo

Abstract A new blind single channel source separation method is presented. The proposed method does not require training knowledge and the separation system is based on nonuniform time-frequency (TF) analysis and feature extraction. Unlike conventional researches that concentrate on the use of spectrogram or its variants, we develop our separation algorithms using an alternative TF representation based on the gammatone filterbank. In particular, we show that the monaural mixed audio signal is considerably more separable in this nonuniform TF domain. We also provide the analysis of signal separability to verify this finding. In addition, we derive two new algorithms that extend the recently published Itakura-Saito nonnegative matrix factorization to the case of convolutive model for the nonstationary source signals. These formulations are based on the Quasi-EM framework and the Multiplicative Gradient Descent (MGD) rule, respectively. Experimental tests have been conducted which show that the proposed method is efficient in extracting the sources' spectral–temporal features that are characterized by large dynamic range of energy, and thus lead to significant improvement in source separation performance.

8.1 Introduction

The principal aim of blind source separation (BSS) is to extract the underlying source signals from only a set of observations. Due to the diverse promising and exciting applications, BSS has attracted a substantial amount of attention in both the academic field as well as the industry. During the last decade, tremendous developments have been achieved in the application of BSS, particularly in wireless

B. Gao · W. L. Woo (✉)
School of Electrical and Electronic Engineering, Newcastle University, England, UK
e-mail: lok.woo@ncl.ac.uk

B. Gao
e-mail: bin.gao@ncl.ac.uk

G. R. Naik and W. Wang (eds.), *Blind Source Separation*,
Signals and Communication Technology, DOI: 10.1007/978-3-642-55016-4_8,

communication, medical signal processing, geophysical exploration, and image enhancement/recognition. The so-called cocktail-party problem within the BSS context refers to the phenomenon of extracting original voice signals of the speakers from the signals recorded from several microphones. Similar examples in the field of radio communication involve the observations that correspond to the outputs of several antenna elements in response to several transmitters that represent the original signals. In the analysis of medical signals, electroencephalography (EEG), magnetoencephalography (MEG), and electrocardiogram (ECG) data represent the observations and BSS is used as a signal processing tool to assist noninvasive medical diagnosis. BSS has also been applied to the data analysis in other areas such as telecommunication, finance, and seismology. Further evidence of these applications can be found in [1–6]. A review of the current literature shows that there are three main classifications of BSS. These include linear and nonlinear, instantaneous and convolutive, overcomplete and underdetermined. In the first classification, linear algorithms dominate the BSS research field due to its simplicity in analysis and its explicit separability. Linear BSS assumes that the mixture is represented by a linear combination of sources. Extension of BSS for solving nonlinear mixtures has also been introduced. This model takes nonlinear distorted signals into consideration and offers a more accurate representation of a realistic environment. In the second classification, when the observed signals consist of combinations of multiple time-delayed versions of the original sources and/or mixed signals themselves, the system is referred as the convolutive mixture. Otherwise, the absence of time delays results in the instantaneous mixture of observed signals. Finally, when the number of observed signals exceeds the number of sources, this refers to the overcomplete BSS. Conversely, when the number of observed signals is less than the number of sources, this becomes the underdetermined BSS.

In general and for many practical applications, the challenging case for source separation is when only one monaural recording is available. This leads to the single channel blind source separation (SCBSS) where the problem can be stated as one observation mixed with several unknown sources. In this work, we consider the case of two sources, namely

$$y(t) = x_1(t) + x_2(t) \tag{8.1}$$

where $t = 1, 2, \ldots, T$ denotes time index and the goal is to estimate the two sources $x_1(t)$ and $x_2(t)$ given only the observation signal $y(t)$. Unlike conventional assumption used in BSS where the sources are assumed to be statistical independent which is rather too restrictive, in this chapter, the sources are characterized as nonstationary processes with time-varying spectra [7]. This assumption is practically justified since most signals encountered in applications are nonstationary with time-varying spectra. Examples include speech, audio, EEG, stock market index, and seismic trace.

Solutions to SCBSS using nonnegative matrix factorization (NMF) [8] have recently gained popularity. They exploit an appropriate time-frequency (TF) analysis on the mono input recordings, yielding a TF representation that can be decomposed as

$$|\mathbf{Y}|^{.2} \approx \mathbf{DH} \tag{8.2}$$

where $|\mathbf{Y}|^{.2} \in \Re_{+}^{F \times T_s}$ is the power TF representation of the mixture $y(t)$ which is factorized as the product of two nonnegative matrices , $\mathbf{D} \in \Re_{+}^{F \times I}$ and $\mathbf{H} \in \Re_{+}^{I \times T_s}$. The superscript '$\cdot$' represents element-wise operation. F and T_s represent the total frequency units and time slots in the TF domain, respectively. If I is chosen to be $I = T_s$, no benefit is achieved in terms of representation. Thus the idea is to determine $I < T_s$ so the matrix $\mathbf{D}$ can be compressed and reduced to its integral components so that it contains only a set of spectral basis vectors, and $\mathbf{H}$ is an encoding matrix that describes the amplitude of each basis vector at each time point. Because NMF gives a parts-based decomposition [8, 9], it has recently been proposed for separating drums from polyphonic music [10] and automatic transcription of polyphonic music [11]. Commonly used cost functions for NMF are the generalized Kullback-Leibler (KL) divergence and Least Square (LS) distance [8]. A sparseness constraint [12] can be added to these cost functions for optimizing $\mathbf{D}$ and $\mathbf{H}$. Other cost functions for audio spectrograms factorization have also been introduced such as that of [13] that assume multiplicative gamma-distributed noise in power spectrograms, while [14] attempts to incorporate phase into the factorization by using a probabilistic phase model. Notwithstanding the above, families of parameterized cost functions, such as the Beta divergence [15] and Csiszar's divergences [16], have also been presented for the source separation. However, they have some crucial limitations that explicitly use training knowledge of the sources [17]. As a consequence, these methods are only able to deal with a very specific set of signals and situations.

Model-based techniques have also been proposed for SCSS which usually require training a set of isolated recordings. The sources are trained by using a Hidden Markov model (HMM) based on Gaussian Mixture Model (GMM) and they are combined in a factorial HMM to separate the mixture [18]. Good separation requires detailed source models that might use thousands of full spectral states, e.g., in [19] HMMs with 8,000 states were required to accurately represent one person's speech for a source separation task. The large state space is required because it attempts to capture every possible instance of the signal. These model-based techniques, however, consume a long time not only in training the prior parameters but also presenting many difficult challenges during the inference stage.

From the above, it is clear that existing solutions to SCBSS are still practically limited and fall short of the success enjoyed in other areas of source separation. In this chapter, a novel separation system is proposed and the contributions are summarized as follows:

(i) A separability analysis in the TF domain for SCBSS and development a quantitative performance measure to evaluate the degree of "separateness" in the monaural mixed signal.
(ii) A separation framework based on the cochleagram. Unlike the spectrogram that deals only with uniform resolution, the gammatone filterbank produces nonuniform TF domain (termed as the cochleagram) whereby each TF unit has different resolution. We prove that the mixed signal is more separable in the cochleagram than the spectrogram and the log-frequency.

(iii) Development of two-dimensional NMF (NMF2D) signal model optimized under the Itakura-Saito (IS) divergence with Quasi-EM and MGD updates (IS-NMF2D). Two new algorithms have been developed to estimate the spectral and temporal features of the audio source model. The first algorithm is founded on the framework of Quasi-EM (Expectation-Maximization) while the second algorithm is based on the multiplicative gradient decent (MGD) update rule. Both algorithms have the unique property of scale-invariant whereby the lower energy components in the TF domain can be treated with equal importance as the higher energy components. This is to be contrasted with other methods based on LS distance [20] and KL divergence [21], which favor the high-energy components but neglect the low-energy components.

The chapter is organized as follows: Sect. 8.2 introduces the TF matrix representation using the gammatone filterbank. Section 8.3 delves into the separability analysis of the single-channel mixture in the nonuniform TF domain. In Sect. 8.4, the two new algorithms are derived and the proposed separation system is developed. Experimental results and a series of performance comparison with methods are presented in Sect. 8.5. Finally, Sect. 8.6 concludes the chapter.

8.2 Time-Frequency Representation

In the task of audio source separation, one critical decision is to choose a suitable TF domain to represent the time-varying contents of the signals. There are several types of TF representations and the most widely used ones are spectrogram [22] and log-frequency spectrogram (using constant-Q transform) [23]. This is documented over the last few years in the research of audio source separation [10–21]. In this work, however, we develop our separation algorithms using a TF representation based on the gammatone filterbank.

8.2.1 Gammatone Filterbank and Cochleagram

The Gammatone filterbank [24] is a cochlear filtering model which decomposes an input signal into the time-frequency domain using a set of gammatone filters. The specific steps of generate cochleagram are summarized as (Table 8.1).

In [25, 26], it was noted that some crucial differences exist in the TF representation of how sound is analyzed by the ear. In particular, the ear's frequency subbands get wider for higher frequencies, whereas the classical spectrogram as computed by the Short-Time Fourier Transform (STFT) has an equal-spaced bandwidth across all frequency channels. Since speech signals are characterized as highly nonstationary and nonperiodic whereas music changes continuously, therefore, application of the Fourier transform will produce errors when complicated transient phenomena such

Table 8.1 Cochleagram computation

1. Give impulse response of a gammatone filter:
   $g(f,t) = t^{h-1}e^{-2\pi vt}\cos(2\pi ft), \quad t \geq 0$ (8.3)
2. The filter output response $x(c,t)$ can be expressed as:
   $x(c,t) = \int_{-\infty}^{\infty} x(\tau) g_{f_c}(t-\tau)\, d\tau$ (8.4)
3. The output of each filter channel is divided into time frames with 50 % overlap between consecutive frames
4. The time-frequency spectra of all the filter outputs are then constructed to form the cochleagram

as the mixture of speech and music is contained in the analyzed signal. Unlike the spectrogram, the log-frequency spectrogram possesses nonuniform TF resolution. However, it does not exactly match the nonlinear resolution of the cochlear since their center frequencies are distributed logarithmically along the frequency axis and all filters have constant-Q factor [23]. On a separate hand, the gammatone filters used in the cochlear model (3) are approximately *logarithmically* spaced with constant-Q for frequencies from $f_s/10$ to $f_s/2$ (f_s denotes the sampling frequency), and approximately *linearly* spaced for frequencies below $f_s/10$. Hence, this characteristic results in selective *nonuniform* resolution in the TF representation of the analyzed audio signal. Figure 8.1 shows the frequency response of a general gammatone filterbank for $f_s = 16\,\text{kHz}$. It is seen that the higher frequencies correspond to the wider frequency subbands which resemble closely to the human perception of frequencies [27]. Therefore, the cochleagram is developed as an alternative TF analysis tool for source separation to overcome the limitations associated with the Fourier transform approach.

8.3 Single Channel Source Separability Analysis

For separation, one generates the TF mask corresponding to each source and applies the generated mask to the mixture to obtain the estimated source TF representation. In particular, when the sources do not overlap in the TF domain, an optimum mask $M_i^{\text{opt}}(f,t_s)$ exists which allows one to extract the ith original source from the mixture as

$$X_i(f,t_s) = M_i^{\text{opt}}(f,t_s)Y(f,t_s) \tag{8.5}$$

Given any TF mask $M_i(f,t_s)$ such that $0 \leq M_i(f,t_s) \leq 1$ for all (f,t_s), we define the separability for the target source $x_i(t)$ in the presence of the interfering sources $p_i(t) = \sum_{j=1, j\neq i}^{N} x_j(t)$ as

$$S_{M_i}^{Y\to X_i,P_i} = \frac{\|M_i(f,t_s)X_i(f,t_s)\|_F^2}{\|X_i(f,t_s)\|_F^2} - \frac{\|M_i(f,t_s)P_i(f,t_s)\|_F^2}{\|X_i(f,t_s)\|_F^2} \tag{8.6}$$

where $X_i(f, t_s)$ and $P_i(f, t_s)$ are the TF representations of $x_i(t)$ and $p_i(t)$, respectively. $\|\cdot\|_F$ is the Frobenius norm. We also define the separability of the mixture with respect to all the N sources as:

$$S_{M_1,\ldots,M_N}^{Y\to X_1,\ldots,X_N} = \frac{1}{N}\sum_{i=1}^{N} S_{M_i}^{Y\to X_i,P_i} \tag{8.7}$$

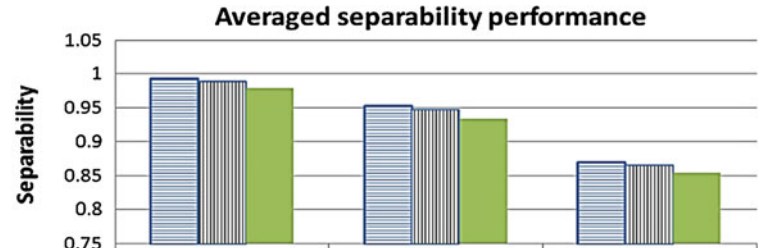
Equation (8.6) is equivalent to measuring the success of extracting the ith source $X_i(f, t_s)$ from the mixture $Y(f, t_s)$ given the TF mask $M_i(f, t_s)$. Similarly, (8.7) measures the success of extracting all the N sources simultaneously from the mixture. To further analyze the separability, we invoke the following: (i) Preserved signal ratio (PSR) that determines how well the mask preserves the source of interest and (ii) Signal-to-interference ratio (SIR) that indicates how well the mask suppresses the interfering sources:

$$\begin{aligned} PSR_{M_i}^{X_i} &= \frac{\|M_i(f,t_s)X_i(f,t_s)\|_F^2}{\|X_i(f,t_s)\|_F^2} \\ SIR_{M_i}^{X_i} &= \frac{\|M_i(f,t_s)X_i(f,t_s)\|_F^2}{\|M_i(f,t_s)P_i(f,t_s)\|_F^2} \end{aligned} \tag{8.8}$$

Using (8.8), it can be shown that (8.7) can be expressed as $S_{M_i}^{Y\to X_i,P_i} = PSR_{M_i}^{X_i} - PSR_{M_i}^{X_i}/SIR_{M_i}^{X_i}$. Analyzing the terms in (8.6), we have

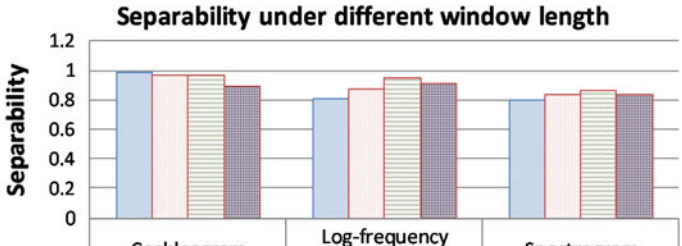
$$\begin{aligned} PSR_{M_i}^{X_i} &:= \begin{cases} 1, & \text{if } \operatorname{sup} pM_i^{\text{opt}} = \operatorname{sup} pM_i \\ <1, & \text{if } \operatorname{sup} pM_i^{\text{opt}} \subset \operatorname{sup} pM_i \end{cases} \\ SIR_{M_i}^{X_i} &:= \begin{cases} \infty, & \text{if } \operatorname{sup} p\,[M_iX_i] \cap \operatorname{sup} pP_i = \oslash \\ \text{finite}, & \text{if } \operatorname{sup} p\,[M_iX_i] \cap \operatorname{sup} pP_i \neq \oslash \end{cases} \end{aligned} \tag{8.9}$$

where 'supp' denotes the support. When $S_{M_i}^{Y\to X_i,P_i} = 1$ (i.e. $PSR_{M_i}^{X_i} = 1$ and $SIR_{M_i}^{X_i} = \infty$), this indicates that the mixture $y(t)$ is separable with respect to the i^{th} source $x_i(t)$. In other words, $X_i(f, t_s)$ does not overlap with $P_i(f, t_s)$ and the TF mask $M_i(f, t_s)$ has perfectly separated the i^{th} source $X_i(f, t_s)$ from the mixture $Y(f, t_s)$. This corresponds to $M_i(f, t_s) = M_i^{opt}(f, t_s)$ in (8.5). Hence, this is the maximum attainable $S_{M_i}^{Y\to X_i,P_i}$ value. For other cases of $PSR_{M_i}^{X_i}$ and $SIR_{M_i}^{X_i}$, we have $S_{M_i}^{Y\to X_i,P_i} < 1$. Using this concept, we can extend the analysis for the case of separating N sources. A mixture is fully separable to all the N sources if and only if $S_{M_1,\ldots,M_N}^{Y\to X_1,\ldots,X_N} = 1$ in (8.7). For the case $S_{M_1,\ldots,M_N}^{Y\to X_1,\ldots,X_N} < 1$, this implies that some of the sources overlap with each other in the TF domain and therefore, they cannot be fully separated. Thus, $S_{M_1,\ldots,M_N}^{Y\to X_1,\ldots,X_N}$ provides the quantitative performance measure for evaluating how separable is the mixture in the TF domain. In our comparison, the following TF representations are used to test the mixture's separability: spectrogram, log-frequency spectrogram, and cochleagram. In the log-frequency spectrogram, the frequency scale is set to logarithmic and grouped into 175 frequency bins in the



	Cochleagram	Log-frequency spectrogram	Spectrogram
M and M	0.993	0.953	0.869
M and S	0.989	0.947	0.865
S and S	0.979	0.934	0.854

Fig. 8.1 Averaged separability performance



	Cochleagram	Log-frequency spectrogram	Spectrogram
20ms	0.985	0.813	0.801
32ms	0.972	0.874	0.834
64ms	0.965	0.948	0.864
128ms	0.892	0.912	0.842

Fig. 8.2 Separability under different window length

range of 50–8 kHz with 24 bins per octave while the bandwidth follows the constant-Q rule [23]. To ensure fair comparison, we generate the ideal binary mask (IBM) [27] directly from the original sources. To reiterate our aim, the separability analysis is undertaken without recourse to any separation algorithms but utilizing only the energy of the sources to ascertain the degree of "separateness" of the mixture in different TF domains. These results have been tabulated in Fig. 8.1. The symbols 'M' and 'S' denotes music and speech, respectively.

In Fig. 8.1, three types of mixture have been used: (i) music mixed with music, (ii) speech mixed with music, and (iii) speech mixed with speech. The speech signals are selected from 10 male and 10 female speeches taken from TIMIT database and are normalized to unit energy. The 10 music sources are selected from the RWC database [28] and also normalized to unit energy. Two sources are randomly chosen from the databases and the mixed signal is generated by adding the sources. All mixed signals

are sampled at 16 kHz sampling rate. TF representation using different window length has also been investigated and the results are tabulated in Fig. 8.2.

Figure 8.2 shows the average separability results for all types of the mixture based on different window length. The bracketed number shows the number of data points corresponding to the particular window length. It is clear that, for both spectrogram and log-frequency spectrogram settings, the STFT with 1024-point window length is the best setting to analyze the separability performance. The results of PSR, SIR, and separability for each TF domain are obtained by averaging over 300 realizations. Following the listening performance test proposed in [29], we conclude that $S_{M_i}^{Y\to X_i,P_i} > 0.8$ leads to acceptable separation performance. Therefore, all TF representations satisfy this condition. While this is true, the spectrogram gives only a mediocre level of separability with averaged $S_{M_1,M_2}^{Y\to X_1,X_2} \approx 0.86$ while the log-frequency spectrogram shows a better result with $S_{M_1,M_2}^{Y\to X_1,X_2} \approx 0.94$. Nevertheless, the cochleagram yields the best separability with $S_{M_1,M_2}^{Y\to X_1,X_2} \approx 0.98$. Notwithstanding this, it is also seen that the average SIR of the cochleagram exhibits a much higher value than those of spectrogram and log-frequency spectrogram. This implies that the amount of interference between any two sources is lesser in the cochleagram.

8.4 The Proposed Method

In this section, two new algorithms are developed, namely the *Quasi-EM IS-NMF2D* and the *MGD IS-NMF2D*. The former algorithm optimizes the parameters of the signal model using the Expectation-Maximization approach, whereas the latter is directly based on the multiplicative gradient descent. To facilitate the derivation of these algorithms, we first consider the signal model in terms of the power TF representation

8.4.1 Signal Models

Since the sources have time-varying spectra, it is befitting to adopt a model whose power spectra can be described separately in terms of time and frequency. Although conventional NMF model can still be used, it will need a large number of spectral components and requires a clustering step to group and assign each spectral component to the appropriate source. As a result, the NMF model may not always yield the optimal results. An alternative model is to use the two-dimensional NMF model (NMF2D) [2, 3, 30, 31]. This model extends the basic NMF to be a two-dimensional convolution of $\mathbf{D}$ and $\mathbf{H}$ i.e. $|\mathbf{Y}|^{\cdot 2} \approx \sum_{\tau,\phi} \overset{\downarrow\phi}{\mathbf{D}}{}^{\tau} \overset{\to\tau}{\mathbf{H}}{}^{\phi}$ where the vertical arrow in $\overset{\downarrow\phi}{\mathbf{D}}{}^{\tau}$ denotes the downward shift that moves each

element in the matrix down by ϕ rows, and the horizontal arrow in $\overrightarrow{\mathbf{H}^{\phi}}^{\tau}$ denotes the right shift operator that moves each element in the matrix to the right by τ columns. In scalar representation, the (f, t_s)th element in $|\mathbf{Y}|^{.2}$ is given by $\left|\mathbf{Y}_{f,t_s}\right|^2 \approx \sum_{i=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\mathbf{D}_{f-\phi,i}^{\tau}\mathbf{H}_{i,t_s-\tau}^{\phi}$ where $\mathbf{D}_{f',i'}^{\tau'}$ is the $\left(f', \tau', i'\right)$th element of $\mathbf{D}$ and $\mathbf{H}_{i',t_s'}^{\phi'}$ is the $\left(i', \phi', t_s'\right)$th element of $\mathbf{H}$. In source separation, this model compactly represents the characteristics of the nonstationary sources by a time-frequency profile convolved in both time and frequency by a time-frequency weight matrix. $\mathbf{D}_i^{\tau}$ represents the spectral basis of ith source in the TF domain and $\mathbf{H}_i^{\phi}$ represents the corresponding temporal code for each spectral basis.

The TF representation of the mixture in (8.1) is given by $Y(f, t_s) = X_1(f, t_s) + X_2(f, t_s)$ where $Y(f, t_s)$, $X_1(f, t_s)$ and $X_2(f, t_s)$ denote the TF components that are obtained by applying the gammatone filterbank to the mixture. The time slots are given by $t_s = 1, 2, \ldots, T_s$ while frequencies by $f = 1, 2, \ldots, F$. Since each component is a function of t_s and f, we represent this as a $F \times T_s$ matrix $\mathbf{Y} = [Y(f, t_s)]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$ and $\mathbf{X}_i = [X_i(f, t_s)]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$. It is shown in Sect. 8.3 that the sources are almost perfectly separable in the cochleagram. This therefore enable us to express the power TF representation as $|\mathbf{Y}|^{.2} \approx \sum_{i=1}^{I}|\mathbf{X}_i|^{.2}$ which we will model as $\left|\mathbf{Y}_{f,t_s}\right|^2 \approx \sum_{i=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\mathbf{D}_{f-\phi,i}^{\tau}\mathbf{H}_{i,t_s-\tau}^{\phi}$. The source we seek to determine are $\left\{|X_i(f, t_s)|^{.2}\right\}_{i=1}^{I}$ and this will be obtained by using the matrix factorization as $\left|\tilde{X}_i(f, t_s)\right|^{.2} = \sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\mathbf{D}_{f-\phi,i}^{\tau}\mathbf{H}_{i,t_s-\tau}^{\phi}$. In the following, we propose two novel algorithms to estimate $\mathbf{D}_{f,i}^{\tau}$ and $\mathbf{H}_{i,t_s}^{\phi}$ from the mixture signal.

8.4.2 Algorithm 1: Quasi-EM Formulation of IS-NMF2D (Quasi-EM IS-NMF2D)

We consider the following generative model defined as:

$$\mathbf{y}_{t_s} = \sum_{k=1}^{K}\mathbf{c}_{k,t_s}, \forall t_s = 1, \ldots, T_s \mathbf{c}_{k,t_s} = \left[c_{k,1,t_s}, \ldots, c_{F,1,t_s}\right]^{\mathbf{T}}$$

$$c_{k,f,t_s} \sim N_c\left(0\,,\ \sum_{\tau,\phi}\mathbf{H}_{k,t_s-\tau}^{\phi}\mathbf{D}_{f-\phi,k}^{\tau}\right) \tag{8.10}$$

where $\mathbf{y}_{t_s} \in C^{F\times 1}, \mathbf{c}_{k,t_s} \in C^{F\times 1}$ and $N_c\,(u,\,\Sigma)$ denotes the proper complex Gaussian distribution and the components $\mathbf{c}_{1,t_s}, \ldots, \mathbf{c}_{K,t_s}$ are both mutually and individually independent. The Expectation-Maximization (EM) framework is developed for the ML estimation of $\theta = \left\{\mathbf{D}^{\tau}, \mathbf{H}^{\phi}\right\}$. Due to the additive structure of the generative

model (8.10), the parameters describing each component $\mathbf{C}_k = \left[\mathbf{c}_{k,1,} \ldots, \mathbf{c}_{k,T_s}\right]$ can be updated separately. We now consider a partition of the parameter space $\theta = \bigcup_{k=1}^{K} \theta_k$ as $\theta_k = \left\{\mathbf{D}_k^{\tau}, \mathbf{H}_k^{\phi}\right\}$ where $\mathbf{D}_k^{\tau}$ is the kth column of $\mathbf{D}^{\tau}$ and $\mathbf{H}_k^{\phi}$ is the k^{th} row of $\mathbf{H}^{\phi}$. The EM algorithm works by formulating the conditional expectation of the negative log likelihood of $\mathbf{C}_k$ as

$$Q_k^{ML}\left(\theta_k|\theta'\right) = -\int_{\mathbf{C}_k} p\left(\mathbf{C}_k|\mathbf{Y}, \theta'\right) \log p(\mathbf{C}_k|\theta_k)\, d\mathbf{C}_k \tag{8.11}$$

where θ' always contains the most recent parameter values of $\left\{\mathbf{D}^{\tau}, \mathbf{H}^{\phi}\right\}$.

8.4.2.1 Expressions of the E- and M-step

One iteration of the EM algorithm includes computing the E-step and maximizing the M-step $Q_k^{ML}\left(\theta_k|\theta'\right)$ for $k = 1, \ldots, K$. The minus hidden-data log likelihood is defined as

$$\begin{aligned} -\log p\left(\mathbf{C}_k|\theta_k\right) &= -\sum_{t_s=1}^{T_s}\sum_{f=1}^{F} \log N_c\left(c_{k,f,t_s} \,\middle|\, 0,\ \sum_{\tau,\phi} \mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}\right) \\ &\doteq \sum_{t_s=1}^{T_s}\sum_{f=1}^{F} \log\left(\sum_{\tau,\phi} \mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}\right) + \frac{\left|c_{k,f,t_s}\right|^2}{\sum\limits_{\tau,\phi} \mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}} \end{aligned} \tag{8.12}$$

where '$\doteq$' in the second line denotes equality up to constant terms. Then, by virtue of (10), the hidden-data posterior also has a Gaussian form as $p\left(\mathbf{C}_k|\mathbf{Y}, \theta\right) = \prod\limits_{t_s=1}^{T_s}\prod\limits_{f=1}^{F} N_c\left(c_{k,f,t_s} \left| u_{k,f,t_s}^{post}\lambda_{k,f,t_s}^{post}\right.\right)$ where u_{k,f,t_s}^{post} and λ_{k,f,t_s}^{post} are the posterior mean and variance of c_{k,f,t_s} given as:

$$u_{k,f,t_s}^{post} = \frac{\sum\limits_{\tau,\phi} \mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}}{\sum\limits_{\tau,\phi,l} \mathbf{D}_{f-\phi,l}^{\tau}\mathbf{H}_{l,t_s-\tau}^{\phi}}\mathbf{Y}_{f,t_s}$$

$$\lambda_{k,f,t_s}^{post} = \frac{\sum\limits_{\tau,\phi} \mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}}{\sum\limits_{\tau,\phi,l} \mathbf{D}_{f-\phi,l}^{\tau}\mathbf{H}_{l,t_s-\tau}^{\phi}} \sum_{\tau,\phi,l\neq k} \mathbf{D}_{f-\phi,l}^{\tau}\mathbf{H}_{l,t_s-\tau}^{\phi} \tag{8.13}$$

Thus, the E-step merely includes computing the posterior power $\mathbf{V}_k$ of component $\mathbf{C}_k$, defined as $[\mathbf{V}_k]_{f,t_s} = v_{k,f,t_s} = \left|u_{k,f,t_s}^{post}\right|^2 + \lambda_{k,f,t_s}^{post}$. The M-step can be treated as one-component NMF problem:

$$\begin{aligned} Q_k^{ML}\left(\theta_k|\theta'\right) &\doteq \sum_{t_s=1}^{T_s}\sum_{f=1}^{F}\log\left(\sum_{\tau,\phi}\mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}\right) + \frac{\left|u_{k,f,t_s}^{post'}\right|^2 + \lambda_{k,f,t_s}^{post'}}{\sum_{\tau,\phi}\mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}} \\ &\doteq \sum_{t_s=1}^{T_s}\sum_{f=1}^{F} d_{IS}\left(|u_{k,f,t_s}^{post'}|^2 + \lambda_{k,f,t_s}^{post'} \;\middle|\; \sum_{\tau,\phi}\mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}\right) \end{aligned} \tag{8.14}$$

where $d_{IS}(\cdot|\cdot)$ is the IS divergence [32] and is formally defined as $d_{IS}(a|b) = (a/b) - \log(a/b) - 1$. The IS divergence has the property of scale invariant, i.e., $d_{IS}(\kappa\, a|\kappa\, b) = d_{IS}(a|b)$ for any κ. This implies that any low energy components (a, b)will bear the same relative importance as the high energy ones $(\kappa\, a, \kappa\, b)$. This is particularly important in situations where$|\mathbf{Y}|^{\cdot 2}$ is characterized by a large dynamic range such as the audio short-term spectra.

8.4.2.2 Estimation of the Spectral Basis and Temporal Code Using Quasi-EM Method

The spectral basis and temporal code can be obtained from (8.14). The derivative of a given element of $g_{k,f,t_s} = \sum_{\tau,\phi}\mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}$ with respect to $\mathbf{D}_{f,k}^{\tau}$ and $\mathbf{H}_{k,t_s}^{\phi}$ is given by:

$$\begin{aligned} \frac{\partial g_{k,f,t_s}}{\partial \mathbf{D}_{f',k'}^{\tau'}} &= \frac{\partial \sum_{\tau,\phi}\mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}}{\partial \mathbf{D}_{f',k'}^{\tau'}} = \mathbf{H}_{k',t_s-\tau'}^{f-f'} \\ \frac{\partial g_{k,f,t_s}}{\partial \mathbf{H}_{k',t_s'}^{\phi'}} &= \frac{\partial \sum_{\tau,\phi}\mathbf{D}_{f-\phi,k}^{\tau}\mathbf{H}_{k,t_s-\tau}^{\phi}}{\partial \mathbf{H}_{k',t_s'}^{\phi'}} = \mathbf{D}_{f-\phi',k'}^{t_s-t_s'} \end{aligned} \tag{8.15}$$

The derivatives of (8.14) corresponding to $\mathbf{D}_{f,k}^{\tau}$ and $\mathbf{H}_{k,t_s}^{\phi}$ is then obtained as

$$\begin{aligned} \frac{\partial Q_k^{ML}(\theta_k|\theta')}{\partial \mathbf{D}_{f',k'}^{\tau'}} &= \frac{\partial}{\partial \mathbf{D}_{f',k'}^{\tau'}}\sum_{f,t_s}\log\left(g_{k,f,t_s}\right) + \frac{v'_{k,f,t_s}}{g_{k,f,t_s}} \\ &= \sum_{\phi,t_s}\left(\frac{g_{k,f'+\phi,t_s} - v'_{k,f'+\phi,t_s}}{g_{k,f'+\phi,t_s}^2}\right)\mathbf{H}_{k',t_s-\tau'}^{\phi} \\ \frac{\partial Q_k^{ML}(\theta_k|\theta')}{\partial \mathbf{H}_{k',t_s'}^{\phi'}} &= \frac{\partial}{\partial \mathbf{H}_{k',t_s'}^{\phi'}}\sum_{f,t_s}\log\left(g_{k,f,t_s}\right) + \frac{v'_{k,f,t_s}}{g_{k,f,t_s}} \\ &= \sum_{\tau,f}\left(\frac{g_{k,f,t_s'+\tau} - v'_{k,f,t_s'+\tau}}{g_{k,f,t_s'+\tau}^2}\right)\mathbf{D}_{f-\phi',k'}^{\tau} \end{aligned} \tag{8.16}$$

Unlike the conventional EM algorithm, it is not possible to directly set $\partial Q_k^{ML}\left(\theta_k|\theta'\right)/\mathbf{D}_{f',k'}^{\tau'} = 0$ and $\partial Q_k^{ML}\left(\theta_k|\theta'\right)/\mathbf{H}_{k',t_s'}^{\phi'} = 0$ because of the nonlinear coupling between and via v'_{k,f,t_s}. Thus, closed-form expressions for estimating $\mathbf{D}_{f,k}^{\tau}$ and $\mathbf{H}_{k,t_s}^{\phi}$ cannot be accomplished. To overcome this problem, we use the following update rules and unify it as part of the M-step:

$$\theta_k \leftarrow \theta_k \cdot \left(\frac{\left[\nabla Q_k^{ML}\left(\theta_k|\theta'\right)\right]_-}{\left[\nabla Q_k^{ML}\left(\theta_k|\theta'\right)\right]_+}\right) \tag{8.17}$$

where $\nabla Q_k^{ML}\left(\theta_k|\theta'\right) = \left[\nabla Q_k^{ML}\left(\theta_k|\theta'\right)\right]_+ - \left[\nabla Q_k^{ML}\left(\theta_k|\theta'\right)\right]_-$. For each $\mathbf{D}_k^{\tau}$ and $\mathbf{H}_k^{\phi}$ variables, we have:

$$\begin{aligned}\left[\nabla Q_k^{ML}\left(\theta_k|\theta'\right)\right]_-^{\mathbf{D}} &= \sum_{\phi,t_s}\left(g_{k,f'+\phi,t_s}\right)^{-2} v'_{k,f'+\phi,t_s}\mathbf{H}_{k',t_s-\tau'}^{\phi} \\ \left[\nabla Q_k^{ML}\left(\theta_k|\theta'\right)\right]_+^{\mathbf{D}} &= \sum_{\phi,t_s}\left(g_{k,f'+\phi,t_s}\right)^{-1}\mathbf{H}_{k',t_s-\tau'}^{\phi}\end{aligned} \tag{8.18}$$

and

$$\begin{aligned}\left[\nabla Q_k^{ML}\left(\theta_k|\theta'\right)\right]_-^{\mathbf{H}} &= \sum_{\tau,f}\mathbf{D}_{f-\phi',k'}^{\tau}\left(g_{k,f,t_s'+\tau}\right)^{-2} v'_{k,f,t_s'+\tau} \\ \left[\nabla Q_k^{ML}\left(\theta_k|\theta'\right)\right]_+^{\mathbf{H}} &= \sum_{\tau,f}\mathbf{D}_{f-\phi',k'}^{\tau}\left(g_{k,f,t_s'+\tau}\right)^{-1}\end{aligned} \tag{8.19}$$

Inserting (8.18) and (8.19) into (8.17) leads to

$$\mathbf{D}_{f',k'}^{\tau'} \leftarrow \mathbf{D}_{f',k'}^{\tau'}\frac{\sum_{\phi,t_s}\left(g_{k,f'+\phi,t_s}\right)^{-2} v'_{k,f'+\phi,t_s}\mathbf{H}_{k',t_s-\tau'}^{\phi}}{\sum_{\phi,t_s}\left(g_{k,f'+\phi,t_s}\right)^{-1}\mathbf{H}_{k',t_s-\tau'}^{\phi}} \tag{8.20}$$

Similarly, the update rules in $\mathbf{H}_{k',t_s'}^{\phi'}$ writes

$$\mathbf{H}_{k',t_s'}^{\phi'} \leftarrow \mathbf{H}_{k',t_s'}^{\phi'}\frac{\sum_{\tau,f}\mathbf{D}_{f-\phi',k'}^{\tau}\left(g_{k,f,t_s'+\tau}\right)^{-2} v'_{k,f,t_s'+\tau}}{\sum_{\tau,f}\mathbf{D}_{f-\phi',k'}^{\tau}\left(g_{k,f,t_s'+\tau}\right)^{-1}} \tag{8.21}$$

It can be verified that the above update rules have an advantage of ensuring the nonnegativity constraints of $\mathbf{D}_{f,k}^{\tau}$ and $\mathbf{H}_{k,t_s}^{\phi}$ are always maintained during every iteration.

8.4.3 Algorithm 2: Multiplicative Gradient Descent Formulation of IS-NMF2D (MGD IS-NMF2D)

We consider the following generative model defined as:

$$\left|\mathbf{Y}_{f,t_s}\right|^2 = \left(\sum_{i=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\mathbf{D}^{\tau}_{f-\phi,i}\mathbf{H}^{\phi}_{i,t_s-\tau}\right)\bullet\mathbf{E}_{f,t_s} \tag{8.22}$$

where $\mathbf{E}_{f,t_s}$ is a scalar of multiplicative independent and identically distributed (i.i.d.) Gamma noise with unit mean, i.e., $p(\mathbf{E}_{f,t_s}) = \xi(\mathbf{E}_{f,t_s}|\alpha,\beta)$ where $\xi(\mathbf{E}_{f,t_s}|\alpha,\beta)$ denotes the Gamma probability density function (pdf) defined as: $\xi(\mathbf{E}_{f,t_s}|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\left(\mathbf{E}_{f,t_s}\right)^{\alpha-1}\exp\left(-\beta\mathbf{E}_{f,t_s}\right),\ \mathbf{E}_{f,t_s}\geq 0$. Next, we define $\mathbf{D} = \left[\mathbf{D}^1\mathbf{D}^2\cdots\mathbf{D}^{\tau_{\max}}\right]$ and $\mathbf{H} = \left[\mathbf{H}^1\mathbf{H}^2\cdots\mathbf{H}^{\phi_{\max}}\right]$. Under the independent and identically distributed (i.i.d.) noise assumption, the term $-\log p(\mathbf{Y}\,|\mathbf{D},\mathbf{H})$ becomes

$$-\log p\left(\mathbf{Y}\,|\mathbf{D},\mathbf{H}\right) = \frac{-\sum_{t_s=1}^{Ts}\sum_{f=1}^{F}\log\xi\left(\frac{|\mathbf{Y}|^{\cdot 2}_{f,t_s}}{\sum_{i=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\mathbf{D}^{\tau}_{f-\phi,i}\mathbf{H}^{\phi}_{i,t_s-\tau}}\,\middle|\,\alpha,\beta\right)}{\sum_{i=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\mathbf{D}^{\tau}_{f-\phi,i}\mathbf{H}^{\phi}_{i,t_s-\tau}}$$

$$\doteq d_{IS}\left(|\mathbf{Y}|^{\cdot 2}_{f,t_s}\,\middle|\,\sum_{i=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\mathbf{D}^{\tau}_{f-\phi,i}\mathbf{H}^{\phi}_{i,t_s-\tau}\right) \tag{8.23}$$

where $\doteq$ in the second line denotes equality up to constant terms. Thus, the cost function is $C^{NMF2D}_{IS} = -\log p\left(\mathbf{Y}\,|\mathbf{D},\mathbf{H}\right)$. The derivatives of (23) corresponding to $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$are given by

$$\begin{aligned}\frac{\partial C^{NMF2D}_{IS}}{\partial\mathbf{D}^{\tau'}_{f',i'}} &= \frac{\partial}{\partial\mathbf{D}^{\tau'}_{f',i'}}\sum_{f,t_s}\left(\frac{|\mathbf{Y}|^2_{f,t_s}}{\mathbf{Z}_{f,t_s}} - \log\frac{|\mathbf{Y}|^2_{f,t_s}}{\mathbf{Z}_{f,t_s}} - 1\right) \\ &= -\sum_{\phi,t_s}\left(\left(\mathbf{Z}_{f'+\phi,t_s}\right)^{-2}\left(|\mathbf{Y}|^2_{f'+\phi,t_s} - \mathbf{Z}_{f'+\phi,t_s}\right)\right)\mathbf{H}^{\phi}_{i',t_s-\tau'}\end{aligned} \tag{8.24}$$

$$\begin{aligned}\frac{\partial C^{NMF2D}_{IS}}{\partial\mathbf{H}^{\phi'}_{i',t'_s}} &= \sum_{f,t_s}\mathbf{D}^{t_s-t'_s}_{f-\phi',i'}\left(\left(\mathbf{Z}_{f,t_s}\right)^{-2}\left(\mathbf{Z}_{f,t_s} - |\mathbf{Y}|^2_{f,t_s}\right)\right) \\ &= -\sum_{\tau,f}\mathbf{D}^{\tau}_{f-\phi',i'}\left(\left(\mathbf{Z}_{f,t'_s+\tau}\right)^{-2}\left(|\mathbf{Y}|^2_{f,t'_s+\tau} - \mathbf{Z}_{f,t'_s+\tau}\right)\right)\end{aligned} \tag{8.25}$$

where $\mathbf{Z} = \sum_{\tau}\sum_{\phi} \overset{\downarrow\phi}{\mathbf{D}^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}^{\phi}}$. The standard gradient decent approach gives

$$\mathbf{D}^{\tau'}_{f',i'} \leftarrow \mathbf{D}^{\tau'}_{f',i'} - \eta_D \frac{\partial Cost^{NMF2D}_{IS}}{\partial \mathbf{D}^{\tau'}_{f',i'}} \quad \text{and} \quad \mathbf{H}^{\phi'}_{i',t'_s} \leftarrow \mathbf{H}^{\phi'}_{i',t'_s} - \eta_H \frac{\partial Cost^{NMF2D}_{IS}}{\partial \mathbf{H}^{\phi'}_{i',t'_s}} \tag{8.26}$$

where η_D and η_H are positive learning rates and can be obtained as

$$\eta_D = \frac{\mathbf{D}^{\tau'}_{f',i'}}{\sum\limits_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s}\right)^{-1} \mathbf{H}^{\phi}_{i',t_s-\tau'}} \quad \text{and} \quad \eta_H = \frac{\mathbf{H}^{\phi'}_{i',t'_s}}{\sum\limits_{\tau,f} \mathbf{D}^{\tau}_{f-\phi',i'} \left(\mathbf{Z}_{f,t'_s+\tau}\right)^{-1}} \tag{8.27}$$

Inserting (8.27) into (8.26) gives the multiplicative gradient decent rules

$$\mathbf{D}^{\tau'}_{f',i'} \leftarrow \mathbf{D}^{\tau'}_{f',i'} \frac{\sum\limits_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s}\right)^{-2} |\mathbf{Y}|^2_{f'+\phi,t_s} \mathbf{H}^{\phi}_{i',t_s-\tau'}}{\sum\limits_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s}\right)^{-1} \mathbf{H}^{\phi}_{i',t_s-\tau'}} \tag{8.28}$$

and

$$\mathbf{H}^{\phi'}_{i',t'_s} \leftarrow \mathbf{H}^{\phi'}_{i',t'_s} \frac{\sum\limits_{\phi,t_s} \left(\mathbf{Z}_{f,t'_s+\tau}\right)^{-2} |\mathbf{Y}|^2_{f,t'_s+\tau} \mathbf{D}^{\tau}_{f-\phi',i'}}{\sum\limits_{\tau,f} \mathbf{D}^{\tau}_{f-\phi',i'} \left(\mathbf{Z}_{f,t'_s+\tau}\right)^{-1}} \tag{8.29}$$

The key difference between both algorithms is that the Quasi-EM IS-NMF2D algorithm prevents zeros in the factors, i.e., $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ cannot take entries equal to zero. On the contrary, this is not a feature shared by the MGD IS-NMF2D algorithm since zero coefficients are invariant under MGD updates. If the MGD IS-NMF2D algorithm attains a fixed point solution with zero entries, then it cannot be determined since the limit point is a stationary point [33]. Consequently, the resulting factorizations rendered by these algorithms are not equivalent. For this reason, the Quasi-EM IS-NMF2D algorithm can be considered more reliable for updating $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$. We have summarized both proposed algorithms in Table 8.2. Details of the source separation performance between these algorithms will be shown in Sect. 8.5 where $\psi = 10^{-6}$ is the threshold for ascertaining the convergence.

8.4.4 Estimation of Sources

The two matrices that we seek to separate from $\left|\mathbf{Y}_{f,t_s}\right|^2$ are $\left|\tilde{X}_1(f,t_s)\right|^{.2}$ and $\left|\tilde{X}_2(f,t_s)\right|^{.2}$. These matrices are estimated as $\left|\tilde{X}_1(f,t_s)\right|^{.2} = \sum\limits_{\tau=0}^{\tau_{\max}} \sum\limits_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{f-\phi,1} \mathbf{H}^{\phi}_{1,t_s-\tau}$

and $\left|\tilde{X}_2(f,t_s)\right|^{.2} = \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{f-\phi,2} \mathbf{H}^{\phi}_{2,t_s-\tau}$ [29] which are then used to generate the binary mask as $\mathbf{mask}_i(f,t_s) = 1$ if $\left|\tilde{X}_i(f,t_s)\right|^{.2} > \left|\tilde{X}_j(f,t_s)\right|^{.2}$ and zero otherwise. Finally, the estimated time-domain sources are obtained as $\tilde{x}_i =$ Resynthesize $(\mathbf{mask}_i \cdot \mathbf{Y})$ for $i = 1, 2$ where $\tilde{x}_i = [\tilde{x}_i(1), \ldots, \tilde{x}_i(T)]^{\mathbf{T}}$ denotes the i^{th} estimated source. The time-domain estimated sources are resynthesized using the approach in [22] by weighting the mixture cochleagram by the mask and correcting phase shifts introduced during the gammatone filtering.

8.5 Experimental Results and Analysis

The proposed separation system is tested on recorded audio signals. All recordings and processing are conducted using a PC with Intel Core 2 CPU 6600 @ 2.4 GHz and 2 GB RAM. For mixture generation, three types of mixtures are used, i.e., mixture of music and speech, mixture of different kinds of music, and mixture of different kinds of speech. The speech sources (male and female) are selected from the TIMIT speech database while the music sources (jazz and piano) from the RWC database [28]. All mixtures are sampled at 16 kHz sampling rate. In all cases, the sources are mixed with equal average power over the duration of the signals. As for our proposed algorithms, the convolutive components are selected as follows:

(i) For jazz and speech mixture, $\tau = \{0, \ldots, 4\}$ and $\phi = \{0, \ldots, 4\}$.
(ii) For jazz and piano mixture, $\tau = \{0, \ldots, 6\}$ and $\phi = \{0, \ldots, 9\}$.
(iii) For piano and speech mixture, $\tau = \{0, \ldots, 6\}$ and $\phi = \{0, \ldots, 9\}$.
(iv) For speech and speech mixture, $\tau = \{0, 1\}$ and $\phi = \{0, 1, 2\}$.

These parameters are selected after conducting Monte Carlo tests over 100 realizations of audio mixture. We have evaluated our separation performance in terms of the Signal-to-Distortion ratio (SDR) that unifies the Signal-to-Interference ratio (SIR) and Signal-to-Artifacts ratio (SAR). MATLAB routines for computing these criteria are obtained from the SiSEC'08 webpage [34].

8.5.1 Separation Performance Under Different TF Representations

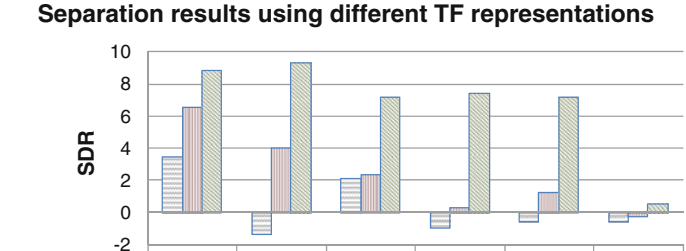
In Sect. 8.2, the separability analysis was undertaken by using the IBM to determine the "separateness" of the mixture without recourse to the separation algorithms. In this section, the impact of separation algorithm is analyzed. Instead of using the IBM, the Quasi-EM IS-NMF2D algorithm is now used to estimate the mask according to Sect. 8.4. In this situation, we are investigating the performance of mixture separation (rather than mixture separability). Speech signals and music are used to generate the

Table 8.2 Pseudo codes for Quasi-EM IS-NMF2D and IS-NMF2D (MGD) algorithms

Quasi-EM IS-NMF2D algorithm	MGD IS-NMF2D algorithm
Input: $\|\mathbf{Y}\|^{\cdot 2}$, random nonnegative matrix $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$, ϕ, τ Output: $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$	Input: $\|\mathbf{Y}\|^{\cdot 2}$, random nonnegative matrix $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$, ϕ, τ Output: $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$
Procedure: Compute initialize cost value $Cost(1)$ using (8.12)	Procedure: Compute initialize cost value $Cost(1)$ using (8.23)
for n=1: max number of iterations	for n=1: max number of iterations
for k=1:K	Compute $\mathbf{Z} = \sum_{\tau}\sum_{\phi} \mathbf{D}^{\tau}_{f-\phi}\mathbf{H}^{\phi}_{t_s-\tau}$
(E-step): Compute $v_{k,f,t_s} = \left\|u^{post}_{k,f,t_s}\right\|^2 + \lambda^{post}_{k,f,t_s}$ using (8.13)	• Update $\mathbf{D}^{\tau'}_{f',i'}$ using (8.28) for all τ, ϕ
(M-step): Iterate convergence is achieved	Normalize $\mathbf{D}^{\tau'}_{f',i'}$
• Update $\mathbf{D}^{\tau'}_{f',k'}$ using (8.20) for all τ, ϕ	Compute $\mathbf{Z} = \sum_{\tau}\sum_{\phi} \mathbf{D}^{\tau}_{f-\phi}\mathbf{H}^{\phi}_{t_s-\tau}$
Normalize $\mathbf{D}^{\tau'}_{f',k'}$	• Update $\mathbf{H}^{\phi'}_{i',t'_s}$ using (8.29) for all τ, ϕ
• Update $\mathbf{H}^{\phi'}_{k',t'_s}$ using (8.21) for all τ, ϕ	Normalize $\mathbf{H}^{\phi'}_{i',t'_s}$
Normalize $\mathbf{H}^{\phi'}_{k',t'_s}$	Compute cost value using (8.23)
end	end
end	Stopping criterion: $\frac{Cost(n-1)-Cost(n)}{Cost(n)} < \psi$
Stopping criterion: $\frac{Cost(n-1)-Cost(n)}{Cost(n)} < \psi$	

monoaural mixture recording. The separation performance is evaluated by using three types of TF representation: (i) spectrogram (STFT with 1024-point Hamming windowed FFT and 50 % overlap), (ii) log-frequency spectrogram (as described in Sect. 8.3 with 1024-point Hamming windowed FFT), and (iii) cochleagram based on Gammatone filterbank of 128 channels, filter order of 4 (i.e., $h = 4$ in (4)), and each filter output is divided into 20 ms time frames with 50 % overlap. To validate the parameters setting of cochleagram (e.g. h and v), we have constructed an experiment based on three speech sources and tested the result by fixing the parameter h in (3) to unity. The experiment is then repeated by progressively increasing h from 2 to 10. Over this range, the optimal separability is obtained when $h = 4$. The parameter v determines the rate of decay of the impulse response of the gammatone filters. In most audio processing tasks, it is set to $v(f) = 1.019ERB(f)$ where $ERB(f) = 24.7 + 0.108f$ is the equivalent rectangular bandwidth of the filter with the center frequency f. A range of values for v has been tested, i.e., $v(f) = (1.019 + c)ERB(f)$ where c ranges from -0.5 to 0.5 with increment of 0.1. The obtained result indicates that the optimal separability is obtained by setting $c = 0$. As c moves away from 0, the separability result progressively deteriorates. This confirms the validity of setting $v(f) = 1.019ERB(f)$ for the cochleagram.

where 'J', 'M', 'F', 'P', 'S' denote jazz, male speech, female speech, piano, and speech, respectively.



	J and M	J and F	P and M	P and F	J and P	S and S
spectrogram	3.47	-1.41	2.1	-1.01	-0.59	-0.6
log-frequency spectrogram	6.54	3.97	2.31	0.27	1.21	-0.3
cochleagram	8.87	9.34	7.16	7.44	7.21	0.5

Fig. 8.3 Separation results using different TF representations

Figure 8.3 shows the comparison of our proposed algorithm based on the spectrogram, log-frequency spectrogram, and cochleagram under various audio mixtures. The separation results for all mixture types based on the spectrogram gives an average SDR of 0.51 dB while the log-frequency spectrogram an average SDR of 2.8 dB. However, a significantly higher performance is attained by the cochleagram with an average SDR of 8 dB. This leads to a substantial improvement gain of 7.5 dB and 5.2 dB, respectively. The major reason for the large discrepancy is due to the mixing ambiguity between $|\mathbf{X}_1|^{\cdot 2}$ and $|\mathbf{X}_2|^{\cdot 2}$. The larger the mixing ambiguity between $|\mathbf{X}_1|^{\cdot 2}$ and $|\mathbf{X}_2|^{\cdot 2}$, the more TF units will be ambiguous which subsequently decreases the probability of correct assignment of each unit to the sources and eventually results in poorer separation performance. To validate this, Fig. 8.4 shows the spectrogram of the original sources, the mixed signal, and the estimated sources using the proposed Quasi-EM IS-NMF2D algorithm. Both figures indicate that the STFT lacks provision for further low-level information of a TF unit and therefore, the resulting spectrogram fails to infer the dominating source. This leads to high degree of ambiguity in TF domain and causes lack of uniqueness in extracting the spectral-temporal features of the sources

Similar to the above, Fig. 8.5 shows the separation results based on the log-frequency spectrogram. Compared with spectrogram, the separation performance is better since log-frequency spectrogram has the propensity of nonuniform time frequency resolution. However, the transform operation used by the log-frequency spectrogram is still based on the Fourier Transform which may not be an optimal option. On the other hand, the results of separation in the cochleagram have led to significant SDR improvement. The cochleagram enables the mixed signal to be more separable and thus reduces the mixing ambiguity between $|\mathbf{X}_1|^{\cdot 2}$ and $|\mathbf{X}_2|^{\cdot 2}$.

This explains the average performance of separating mixture jazz music and female utterance is the highest among all the mixtures because both sources have very distinguishable TF patterns in the cochleagram. This is evident in Fig. 8.6, which shows the separation results in the cochleagram. The plot clearly shows that the

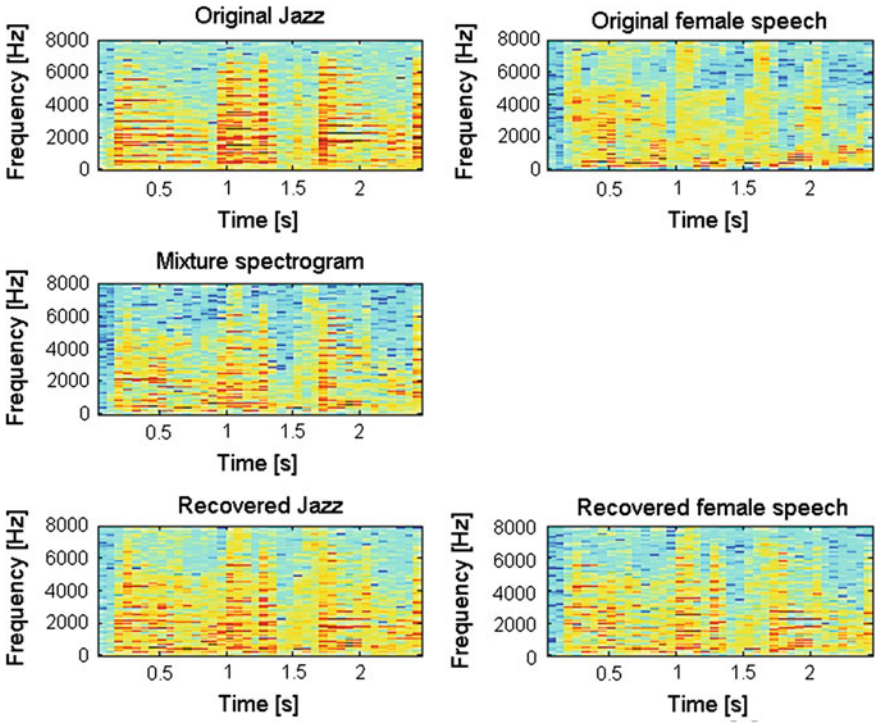


Fig. 8.4 Separation results in spectrogram

spectral energy of the two audio sources has been clustered at different frequencies in the cochleagram due to their different fundamental frequencies. These prominent features have been separated using our proposed Quasi-EM IS-NMF2D algorithm.

The performance of source separation also depends on how accurate the spectral bases are estimated. Given the different types of TF representation, a question arises as to which set of estimated spectral bases have yielded better approximation to the respective original sources' spectral bases. Figure 8.7 shows the results of the original and the estimated spectral basis $\mathbf{D}_i^{\tau}$ for the above mixture when the factorization is performed in the cochleagram. In Fig. 8.7, panels (a and b) refer to the original spectral bases of the jazz music and female utterance, respectively. Panels (c and d) refer to the estimated spectral bases. In comparison, we have also included similar factorization results of the same mixture in the spectrogram and log-frequency spectrogram. These are shown in Figs. 8.8 and 8.9, respectively. In sharp contrast with Fig. 8.7, it is noted that the estimated spectral bases in Figs. 8.8 and 8.9 are quite dissimilar to the original spectral bases. Thus, the construction of the separating mask will inevitably introduce errors in assigning the TF units to the respective sources. Therefore, the recovered sources are very coarse with very low values of SDR in Fig. 8.3.

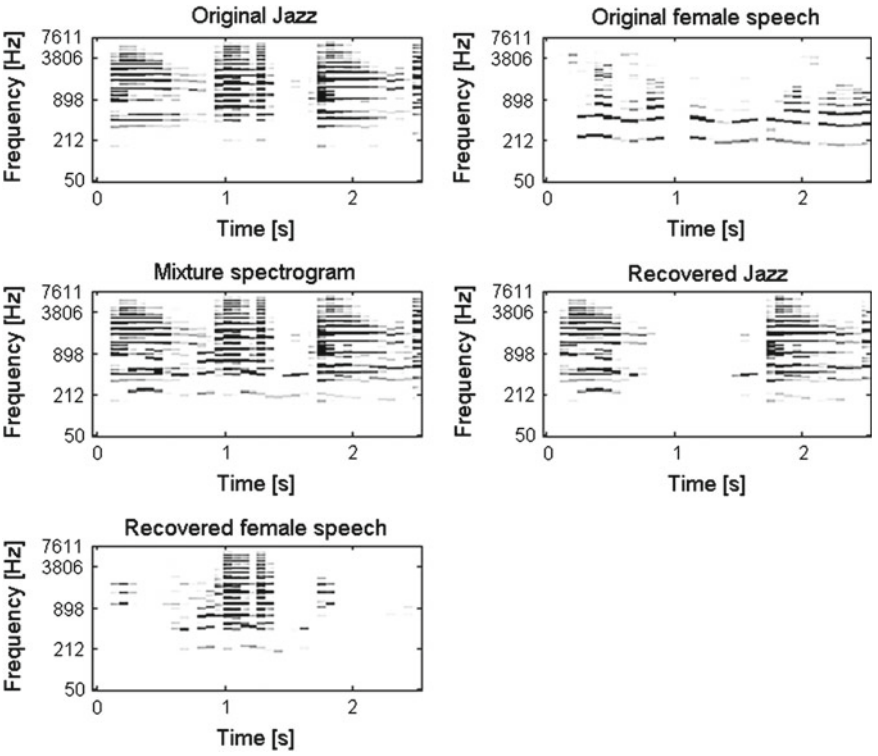


Fig. 8.5 Separation results in log-frequency spectrogram

8.5.2 Comparison Between Different Cost Functions

In the following, experiments are conducted to evaluate the efficiency of the proposed algorithm under different cost functions. Here, we consider the Least Square (LS) distance and Kullback-Leibler (KL) divergence. Figure 8.10 shows the separation results in the cochleagram based on LS, KL, and IS cost functions. In Fig. 8.10, it is noted that Quasi-EM IS-NMF2D algorithm outperforms those of LS distance and KL divergence with an average SDR of 3.1 and 1.8 dB, respectively. This is evidenced by the fact that the IS divergence holds a desirable property of scale invariant so that low energy components can be precisely estimated and they bear the same relative importance as the high energy ones. On the contrary, factorizations obtained with LS distance and KL divergence tend to favor the high energy components at the expense of disregarding the low energy ones. In the cochleagram, the dynamic range of the mixture signal can be considerably large such that the dominating signal at a particular TF unit can manifest either as low or high energy components. In addition, these components tend to exist as clusters. As such, when either LS distance- or KL divergence is used, clusters with low energy tend to be ignored in favor of the

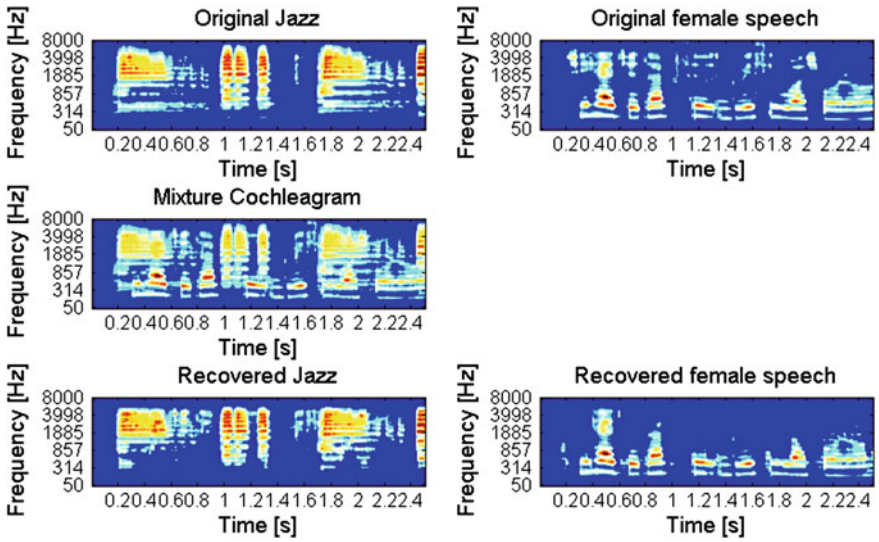


Fig. 8.6 Separation results in cochleagram

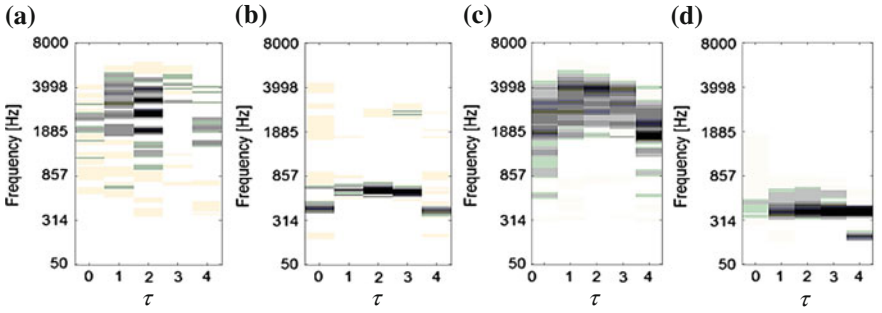


Fig. 8.7 **a–b** Original spectral bases of jazz music and female utterance in the cochleagram. **c–d** The corresponding estimated spectral bases

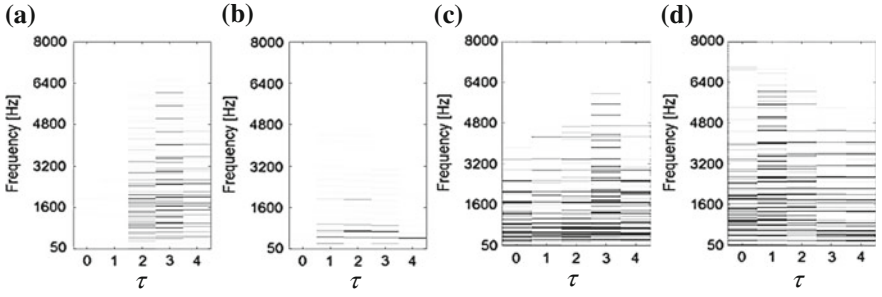


Fig. 8.8 **a–b** Original spectral bases of jazz music and female utterance in the spectrogram. **c–d** The corresponding estimated spectral bases

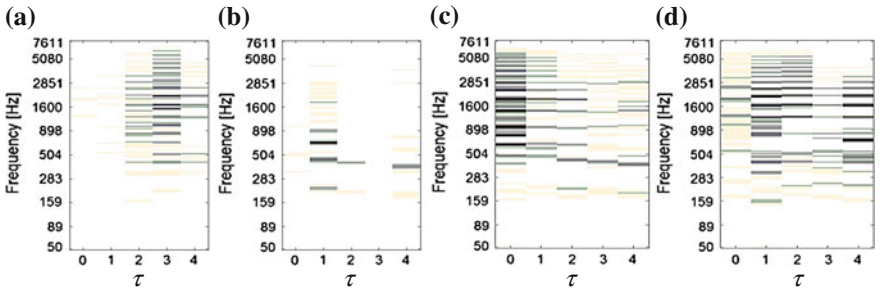
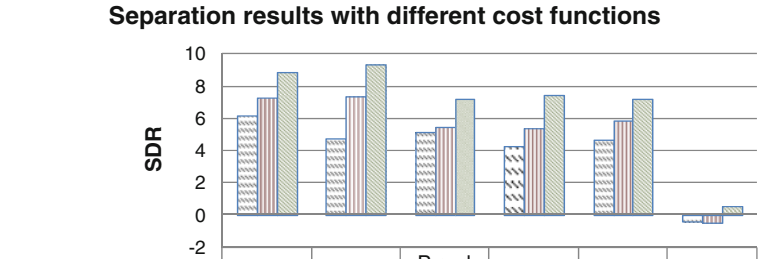


Fig. 8.9 **a**–**b** Original spectral bases of jazz music and female utterance in the log-frequency spectrogram. **c**–**d** The corresponding estimated spectral bases



	J and M	J and F	P and M	P and F	J and P	S and S
LS-NMF2D	6.15	4.69	5.11	4.21	4.61	-0.4
KL-NMF2D	7.24	7.35	5.42	5.38	5.86	-0.5
Quasi-EM IS-NMF2D	8.87	9.34	7.16	7.44	7.21	0.5

Fig. 8.10 Separation results with different cost functions

high energy ones. This leads to mixing ambiguities especially for low energy ones in which case when they are subsumed together leads to significant lost of spectral–temporal information of the sources. Figure 8.11 shows how different cost functions have impacted the separation performance. It is clearly seen that the LS-NMF2D algorithm fails to determine the correct TF components of each source. Panels (a and b) show a considerable level of mixing ambiguities (red box marked area) that have not been accurately resolved by the LS-NMF2D algorithm. The KL-NMF2D exhibits better performance but ignores some low energy TF components in the red box marked area of (c). On the other hand, the proposed algorithm has successfully extracted the low energy components for both female speech and jazz music with high accuracy.

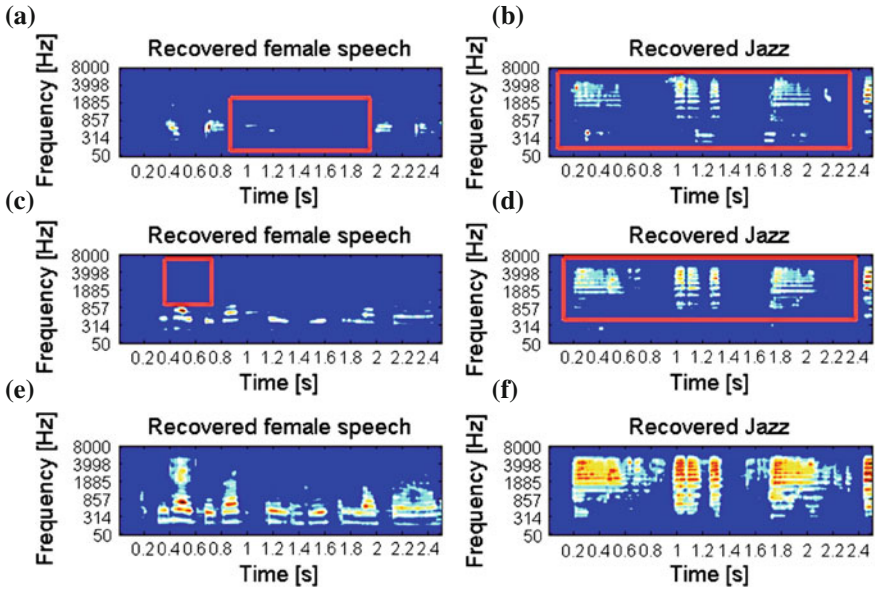


Fig. 8.11 Separation results: **a**–**b**, **c**–**d** and **e**–**f** denote the recovered female speech and jazz music in the cochleagram by using the algorithms with different cost function

8.5.3 Comparing with Different SCBSS Methods

We have made comparison with the recently published EMD SCBSS [35], which first decomposes the given signal into spectrally independent modes using EMD algorithm, and then, ICA is applied to extract statistically independent sources. All the above single channel BSS methods will be tested across all types of mixture and compared in terms of SDR. Table 8.3 summarizes the comparison results. In comparison, the Quasi-EM IS-NMF2D with cochleagram leads to the best separation performance for all types of the mixture. The EMD SCBSS also performs with relative acceptable results compared with Quasi-EM IS-NMF2D. However, it is interesting to point out that the advantage of using Quasi-EM IS-NMF2D with cochleagram is that this method is less complex than the EMD SCBSS and simultaneously it retains a higher level of the separation performance.

8.5.4 Separating More than Two Sources

The proposed method can be extended to the case when $i > 2$ sources. If more than two sources are mixed in a single channel, this requires specifying the number of sources to be separated. Since the method is blind, the separability of the complex

Table 8.3 Separation results using different SCBSS methods

Mixtures	Method	SDR
Jazz and male	EMD SCBSS	6.3
	Quasi-EM IS-NMF2D	8.8
Jazz and female	EMD SCBSS	5.2
	Quasi-EM IS-NMF2D	9.3
Piano and male	EMD SCBSS	5.2
	Quasi-EM IS-NMF2D	7.1
Piano and female	EMD SCBSS	6.6
	Quasi-EM IS-NMF2D	7.4
Jazz and piano	EMD SCBSS	6.6
	Quasi-EM IS-NMF2D	8.5
Speech and speech	EMD SCBSS	0.4
	Quasi-EM IS-NMF2D	0.5

mixture depends highly on how accurate the spectral bases $\mathbf{D}_i^{\tau}$ can be estimated from the TF mixture. Consequently, a set of distinguishable spectral basis of each source for a generic case is a necessary condition to achieve good separation performance. Thus, we adopt three different types of sources, e.g., jazz, piano, and trumpet to generate a complex mixture. The convolutive components in the proposed algorithm are selected as $\tau = \{0, \ldots, 3\}$ and $\phi = \{0, \ldots, 31\}$. Table 8.4 shows the overall separation results. It is seen that mixtures generated by all music sources have been recovered quite successfully. Figure 8.12 shows an example of separating the mixture of Jazz, piano, and trumpet music. It can be seen that three music sources are almost completely separated by using the proposed method. In addition, it is noted that the separation performance has deteriorated when the number of sources increases from two. Increased number of sources will mean that there exists more interference in separating every target source and hence results in higher probability of incurring an error. Comparing the results in the table, mixtures containing speech somehow results in slightly poorer performance than mixtures of music sources only. One reason is the seemingly more overlaps in the TF domain between the speech and music sources. It is observed from Fig. 8.6 that music pitches tend to jump discretely while speech pitches do not. Consequently, this leads to less efficiency in the estimation of the spectral basis from the mixture signal. In addition, we have tested the performance of the proposed method on recordings mixed with $i > 3$ sources. We have found that the proposed method works well for mixtures of music sources that are characterized with distinguishable spectral basis. However, the performance shows degradation when separating mixture contains speech sources.

8.5.5 Separating Real Music Recording

In the final experiment, the proposed method is tested on professionally produced music recordings of the well-known song namely "You raise me up" by Kenny G. The

Table 8.4 Separation results of three sources

Mixtures: $y = x_1 + x_2 + x_3$			SDR of $\hat{x}_1$	SDR of $\hat{x}_2$	SDR of $\hat{x}_3$
x_1	x_2	x_3			
Jazz	Piano	Trumpet	6.51	5.61	5.65
Male	Jazz	Piano	5.23	5.73	4.13
Male	Jazz	Trumpet	5.18	5.65	5.21
Male	Piano	Trumpet	5.20	4.09	4.53
Female	Jazz	Piano	5.36	5.47	4.24
Female	Jazz	Trumpet	5.02	5.51	5.10
Female	Piano	Trumpet	5.02	4.32	4.28
Male	Female	Male	−0.8	1.3	−1.6

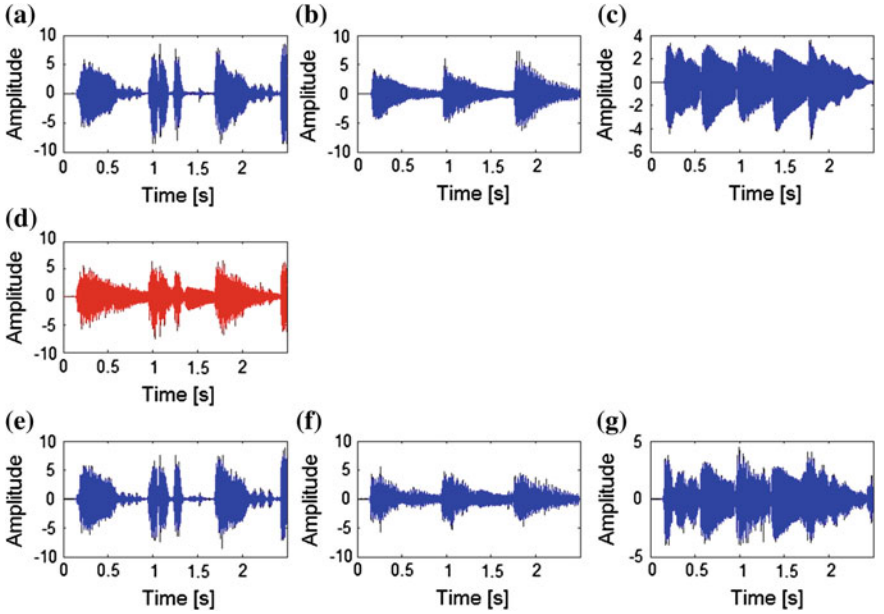


Fig. 8.12 Decomposition results. **a**–**c** denote the original Jazz, piano, and trumpet music, **d** is the mixture and **e**–**g** denote the recovered sources using the proposed method

music consists of two excerpts of length approximately 23 s on mono channel and resampled to 16 kHz. The song is an instrumental music consisting of saxophone and piano sound. The factors of τ and ϕ shifts are set to have $\tau_{\max} = 8$ and $\phi_{\max} = 32$. Since the original source spatial images are not available for this experiment, the separation performance is assessed perceptually and informally by analyzing the log-frequency spectrogram of the estimated source images and listening to the separated sound. This task was a tough task since the instruments play many different notes in the recording. Figure 8.13 shows the separation results of the saxophone and piano

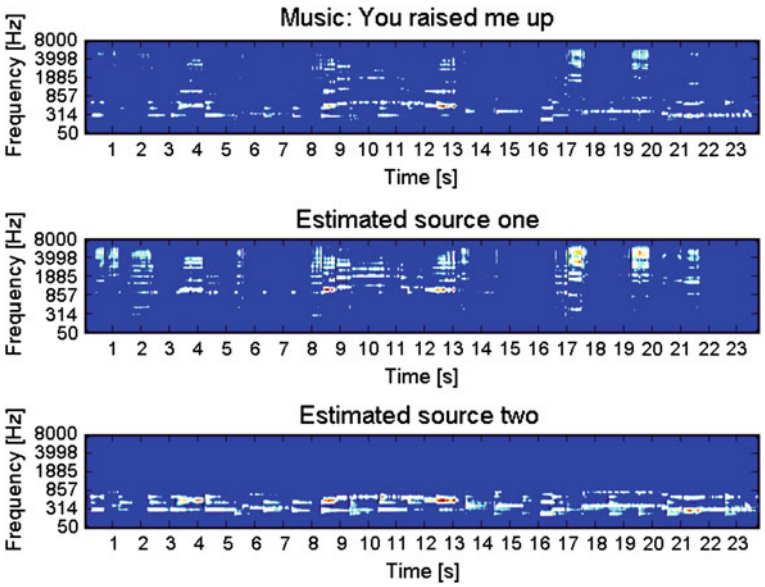


Fig. 8.13 Separation result for song "You raised me up" by Kenny G. *Top* Recorded music. *Middle* Separated saxophone sound. *Bottom* Separated piano sound

sound. The high pitch of continuous saxophone sound is shown in the middle panel of Fig. 8.13 while the notes of the piano are evidently present in Fig. 8.11 bottom panel. Overall, our proposed method successfully separated the professionally produced music recordings and gives a perceptually pleasant listening experience.

8.6 Conclusion

In this chapter, a novel method to solve the single channel audio source separation is proposed. In addition, two algorithms for nonnegative matrix two-dimensional factorization optimized using the Itakura-Saito divergence are presented: Quasi-EM IS-NMF2D and MGD IS-NMF2D. Coupled with the theoretical support of signal separability in the TF domain, the separation system using the gammatone filterbank with these algorithms have shown to yield considerable success. The proposed method enjoys at least three significant advantages: First, it avoids strong constraints of separating sources without training knowledge. Second, the cochleagram rendered by the gammatone filterbank has nonuniform TF resolution which enables the mixed signal to be more separable and thus improves the efficiency of source separation. Finally, the method holds a desirable property of scale invariant which enables low energy components in the cochleagram to bear the same relative importance as the high energy ones. The proposed cochleagram-based IS-NMF2D method in partic-

ular using the Quasi-EM algorithm has yielded significant improvements in source separation compared with other nonnegative matrix factorizations.

References

1. Lee, T.W.: Blind source separation of nonlinear mixing models. Neural Netw. **7**, 121–131 (199**7**)
2. Gao, B., Woo, W.L., Dlay, S.S.: Unsupervised single channel separation of non-stationary signals using gammatone filterbank and Itakura-Saito nonnegative matrix two-dimensional factorizations. IEEE Trans. Circuits Syst. I **60**(3), 662–675 (2013)
3. Gao, B., Woo, W.L., Dlay, S.S.: Variational regularized two-dimensional nonnegative matrix factorization. IEEE Trans. Neural Netw. Learn. Syst. **23**(5), 703–716 (2012)
4. Hyvarinen, A., Karhunen, J., Oja, E.: Independent component analysis and blind source separation, pp. 20–60. Wiley, New York (2001)
5. Cichocki, A., Amari, S.I.: Adaptive Blind Signal and Image Processing—Learning Algorithms and Applications. Wiley (2003)
6. Hyvarinen, A.: Survey on independent component analysis. Neural Comput. Surv. **1**, 94–128 (1999)
7. Taleb, A., Jutten, C.: Source separation in post-nonlinear mixtures. IEEE Trans. Sign. Process. **47**(10), 2807–2820 (1999)
8. Lee, D., Seung, H.: Learning the parts of objects by nonnegative matrix factorisation. Nature **401**(6755), 788–791 (1999)
9. Xie, S., Yang, Z.Y., Fu, Y.L.: Nonnegative matrix factorization applied to nonlinear speech and image Cryptosystems. IEEE Trans. on Circuits Syst. I **55**, 2356–2367 (2008)
10. Helén, M., Virtanen, T.: Separation of drums from polyphonic music using nonnegative matrix factorization and support vector machine. In: Proceedings of 13th European Signal Processing. Turkey (2005)
11. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 177–180. (2003)
12. Rickard, S., Cichocki, A.: When is non-negative matrix decomposition unique? In: 42nd Annual Conference on Information Sciences and Systems (CISS), pp. 1091–1092. (2008)
13. Abdallah, S.A., Plumbley, M.D.: Polyphonic transcription by non-negative sparse coding of power spectra. In: Proceedings of 5th International Confernence on Music Information Retrieval (ISMIR '04), pp. 318–325. Spain (2004)
14. Parry, R.M., Essa, I.: Incorporating phase information for source separation via spectrogram factorization. In: Proceedings of Conference on Acoustics, Speech and Signal Processing (ICASSP'07), pp. 661–664. Hawaii (2007)
15. Kompass, R.: A generalized divergence measure for nonnegative matrix factorization. Neural Comput. **19**(3), 780–791 (2007)
16. Cichocki, A., Zdunek, R., Amari, S.-I.: Csisz'ar's divergences for non-negative matrix factorization: family of new algorithms. In: Proceedings of 6th International Conference on Independent Component Analysis and Signal Separation (ICA '06), pp. 32–39. Charleston (2006)
17. Virtanen, T.: Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. **15**(3), 1066–1074 (2007)
18. Radfa, M.H., Dansereau, R.M.: Single-channel speech separation using soft mask filtering. IEEE Trans. Audio Speech Lang. Process. **15**(6) (2007)
19. Roweis, S.: One microphone source separation. In: Proceedings of Neural Information Processing, pp. 793–799 (2000)

20. Morup, M., Schmidt, M.N.: Sparse Non-negative Matrix Factor 2-D Deconvolution. Technical Report, Denmark (2006)
21. Schmidt, M.N., Morup, M.: Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In: Proceedings 6th Internatinal Conference on Independent Component Analysis and Signal Separation (ICA '06), pp. 700–707. Charleston (2006)
22. Gröchenig, K.: Foundations of Time-Frequency Analysis. Birkhauser, Boston (2001)
23. Brown, Judith C.: Calculation of a constant Q spectral transform. J. Acoust. Soc. Am. **89**(1), 425–434 (1991)
24. Hu, G., Wang, D.L.: Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Networks **15**(5), 1135–1150 (2004)
25. Roads, C., et al.: The computer music tutorial. The MIT Press, Cambridge (1996)
26. Schulz, S., Herfet, t.: Binaural source separation in non-ideal reverberant environments. In: Proceedings of 10th International Conference on Digital Audio Effects (DAFx-07), pp. 10–15. Bordeaux, France (2007)
27. Wang, D.L.: On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P. (ed.) Speech Separation by Humans and Machines, pp. 181–197. Norwell, Kluwer (2005)
28. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC music database: music genre database and musical instrument sound database. In: Proceedings of International Symposium on Music Information Retrieval (ISMIR), pp. 229–230. Baltimore, Maryland (2003)
29. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Sign. Process. **52**(7), 1830–1847 (2004)
30. Gao, B., Woo, W.L., Dlay, S.S.: Single channel source separation using EMD-subband variable regularized sparse features. IEEE Trans. Audio Speech Lang. Process. **19**, 961–976 (2011)
31. Gao, B., Woo, W.L., Dlay, S.S.: Adaptive sparsity non-negative matrix factorization for single channel source separation. IEEE J. Sel. Top. Sign. Process, **5**, 1932–4553 (2011)
32. Itakura, F., Saito, S.: Analysis synthesis telephony based on the maximum likelihood method. In: Proceedings of 6th International Congress on Acoustics, pp. C-17–C-20. Tokyo, Aug 1968
33. Fevotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
34. Signal Separation Evaluation Campaign (SiSEC 2008) (2008) Available http://sisec.wiki.irisa.fr