Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation

Bin Gao
School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, People's Republic of China

W. L. Woo[a] and L. C. Khor
School of Electrical and Electronic Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom

(Received 6 November 2012; revised 15 January 2014; accepted 17 January 2014)

An unsupervised single channel audio separation method from pattern recognition viewpoint is presented. The proposed method does not require training knowledge and the separation system is based on non-uniform time-frequency (TF) analysis and feature extraction. Unlike conventional research that concentrates on the use of spectrogram or its variants, the proposed separation algorithm uses an alternative TF representation based on the gammatone filterbank. In particular, the monaural mixed audio signal is shown to be considerably more separable in this non-uniform TF domain. The analysis of signal separability to verify this finding is provided. In addition, a variational Bayesian approach is derived to learn the sparsity parameters for optimizing the matrix factorization. Experimental tests have been conducted, which show that the extraction of the spectral dictionary and temporal codes is more efficient using sparsity learning and subsequently leads to better separation performance.
© *2014 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4864294]

PACS number(s): 43.60.Ac, 43.60.Hj, 43.60.Pt, 43.60.Uv [SAF] Pages: 1171–1185

I. INTRODUCTION

Although blind source separation (BSS) has gained immense popularity since the last decades, the case of underdetermined blind source separation (UBSS) is still less addressed than the over-determined case. Methods for solving the UBSS problem[1–6] without prior knowledge on the source distribution have been proposed by exploiting the sparseness of the non-stationary sources (e.g., audio sources). For audio sources, this requires transforming the signals to the time-frequency domain and using a computational model to replicate the auditory patterns. This BSS method is considered as time-frequency (TF) pattern separation. The aim is to extract meaningful pattern about the audio contents and provide a better description of the audio signals. This is important, for example, in automatic music transcription where the musicians often spend large amounts of time listening to a song to learn the part of specific instrument by ear. If the instrument of interest can be extracted, it simplifies the task of the musician. Therefore, it is highly desirable to design a machine learning algorithm that is able to offer the separated sound of each instrument.

In general and for many practical applications, one of the most challenging UBSS problems for audio signal separation is when only a monaural recording is available. This leads to the single channel audio separation (SCAS) problem. The problem can be stated as one observation mixed with several unknown audio signals, namely,

$$y(t) = \sum_i x_i(t), \tag{1}$$

where $t = 1, 2, \ldots, T$ denotes time index and the goal is to estimate the ith audio patterns $x_i(t)$ given only the observation signal $y(t)$. Unlike random signals, audio signals tend to have patterns that are characteristically unique to the type of audio or instruments. We solve the SCAS problem from the viewpoint of unsupervised pattern recognition. In particular, Eq. (1) can be converted into the single channel audio pattern separation (SCAPS) problem by transforming the mixture into the time-frequency domain. Conventional methods used in SCAS assume that the audio signals are statistically independent of each other. This condition is rather too restrictive, and in this paper, we treat the audio signals as characterized by non-stationary power spectra processes[7] which can be suitably modeled as patterns in some two-dimensional spaces (as described in Sec. III).

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

$$\mathbf{Y} \approx \mathbf{DH}, \tag{2}$$

where $\mathbf{Y} \in \Re_+^{K\times L}$ is the power time-frequency (TF) representation of the mixture $y(t)$ which is factorized as the product of two nonnegative matrices, $\mathbf{D} \in \Re_+^{K\times I}$ and $\mathbf{H} \in \Re_+^{I\times L}$. K and L represent the total frequency units and time slots in the TF domain, respectively. The idea is to determine $I < L$ so the matrix $\mathbf{D}$ can be compressed and reduced to its integral components. The matrix $\mathbf{D}$ will therefore contain only a

[a] Author to whom correspondence should be addressed. Electronic mail: w.l.woo@ncl.ac.uk

set of spectral basis vectors, and **H** is an encoding matrix which describes the amplitude of each basis vector at each time point. As NMF gives a parts-based decomposition,[8,9] it has recently been applied for separating drums from polyphonic music[10] and automatic transcription of polyphonic music. Commonly used cost functions for NMF are the generalized Kullback−Leibler (KL) divergence and least square (LS) distance.[8] A sparsity constraint[11,12] can be added to these cost functions for optimizing **D** and **H**. Other cost functions for audio spectrograms factorization have also been introduced that assume multiplicative gamma-distributed noise in power spectrograms[13] while others attempt to incorporate phase into the factorization by using a probabilistic phase model.[14] Notwithstanding the above, families of parameterized cost functions such as Beta divergence[15] and Csiszar's divergences[16] have also been presented for audio pattern separation. However, they have some crucial limitations that explicitly require training knowledge of the audio patterns.[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 SCAPS which usually require training a set of isolated recordings. The audio patterns are used for training a hidden Markov model (HMM) based on Gaussian mixture models (GMM), and they are then combined in a factorial HMM to separate the mixture.[18] Good separation requires detailed audio pattern models that might use thousands of full spectral states, e.g., 8000-state HMMs were required to accurately represent one person's speech for the audio separation task.[19] 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 present many difficult challenges during the inference stage.

It is clear from the above that existing solutions to SCAPS are still practically limited and fall short of the success enjoyed in other areas of audio pattern separation. In this paper, a novel separation system is proposed and the contributions are summarized as follows: First, we derive a separability analysis in the TF domain for SCAPS and develop a quantitative performance index to evaluate the degree of "separateness" in the monaural mixed signal. In particular, we have identified the ideal condition when the audio patterns are perfectly separable. Second, a novel development of separation framework based on the gammatone filterbank is proposed. Unlike the spectrogram which deals only with uniform resolution, the gammatone filterbank produces non-uniform TF domain (termed the cochleagram) whereby each TF unit has different resolutions. We prove that the mixed signal is significantly more separable in the cochleagram than in the spectrogram. Finally, a novel sparsity learning two-dimensional non-negative matrix factorization is proposed. Our proposed model allows: (i) overcomplete representation by allowing many spectral and temporal shifts which are not inherent in the NMF and SNMF models. Thus, imposing sparsity is necessary to give unique and realistic representations of the non-stationary audio signals. Unlike the SNMF2D, our model imposes sparsity on **H** element-wise so that *each individual code* has its own distribution. Therefore, the sparsity parameter can be individually optimized for each code. This overcomes the problem of under- and over-sparse factorization. (ii) Each sparsity parameter in our model is learned and adapted as part of the matrix factorization. This bypasses the need of manual selection as in the case of two-dimensional sparse NMF deconvolution (SNMF2D).[20]

The paper is organized as follows: Sec. II introduces the TF matrix representation using the cochleagram and the separability analysis of the single-channel mixture in the non-uniform TF domain. In Sec. III, the new algorithms are derived and the proposed separation system is developed. Experimental results and a series of performance comparison with methods are presented in Sec. IV. Finally, Sec. V concludes the paper.

II. PROPOSED FRAMEWORK

In the task of audio pattern 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 one is the spectrogram.[22] This is documented over the last few years in the research of audio separation.[10–23] In this work, however, we develop our separation algorithm using a non-uniform TF representation based on the cochleagram.

A. Non-uniform time-frequency domain

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. It was noted that some crucial differences exist in the TF representation of how sound is analyzed by the ear.[25,26] In particular, the ear's frequency subbands get wider for higher frequencies, whereas the classical spectrogram computed by the short-time Fourier transform (STFT) has an equal-spaced bandwidth across all frequency channels. Since speech signals are characterized as highly non-stationary and non-periodic whereas music changes continuously; therefore, application of the Fourier transform will produce errors when complicated transient phenomena such as the mixture of speech and music is contained in the analyzed signal. Unlike the spectrogram, the gammatone filters used in the cochlear model are approximately *logarithmically* spaced with constant-Q (Ref. 23) 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 has resulted in selective *non-uniform* resolution in the TF representation of the analyzed audio signal and the higher frequencies correspond to the wider frequency subbands which closely resemble the human perception of frequencies.[27] Therefore, the cochleagram is developed as an alternative TF analysis tool for audio pattern separation to overcome the limitations associated with the Fourier transform approach. The specific steps to generate the cochleagram have been summarized in Table I.

From Fig. 1, it can be seen that the classic spectrogram based on the STFT yields a time-frequency representation with only uniform time and frequency resolution. On the

TABLE I. Cochleagram computation.

1. Generate the impulse response of a gammatone filter:
$g(f,t) = t^{h-1}e^{-2\pi vt}\cos(2\pi ft)$, $t \geq 0$,
where $h = 4$ denotes the order of filter, $v(f) = 1.019\mathrm{ERB}(f)$ represents the rectangular bandwidth and $\mathrm{ERB}(f) = 24.7 + 0.108f$.
2. Compute the filter output response $x(c,t)$ as filtering:
$x(c,t) = \int_{-\infty}^{\infty} x(\tau) g_{f_c}(t-\tau)d\tau$.
3. Divide the output of each filter channel 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.

other hand, the cochleagram based on gammatone filter bank has non-uniform time-frequency resolutions between the high and low frequency regions. A visual example of cochleagram for audio signals can be seen in Fig. 14. It shows the cochleagram of the jazz music, female speech and its mixture. Two noticeable features are apparent from the figure. First, the overlapping region of the mixture's cochleagram is considerably less. Second, the pitches assemble in more distinguishable regions corresponding to different types of patterns. These advantages are not available to other forms of TF transform such as spectrogram or its variants.

B. Time-frequency separability analysis

For separation, one generates the TF mask corresponding to each audio pattern and applies the generated mask to the mixture to obtain the estimated audio pattern TF representation. In particular, when the patterns do not overlap in the TF domain, an optimum mask $M_i^{\mathrm{opt}}(f,t_s)$ exists which allows one to extract the ith original audio pattern from the mixture as $X_i(f,t_s) = M_i^{\mathrm{opt}}(f,t_s)Y(f,t_s)$ with $t_s = 1,2,\ldots,T_s$ representing the time slots and $f = 1,2,\ldots,F$ representing the frequencies. 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 (i) preserved signal ratio (PSR) which determines how well the mask preserves the source of interest, and (ii) signal-to-interference ratio (SIR) which indicates how well the mask suppresses the interfering sources and the separability for the target audio pattern $x_i(t)$ in the presence of the interfering audio patterns $p_i(t) = \sum_{j=1,j\neq i}^{N} x_j(t)$ as

$$\mathrm{Sep}_i = \mathrm{PSR}_i - \frac{\mathrm{PSR}_i}{\mathrm{SIR}_i}, \qquad (3)$$

where

$$\mathrm{PSR}_i = \frac{\|M_i(f,t_s)X_i(f,t_s)\|_F^2}{\|X_i(f,t_s)\|_F^2}$$

and

$$\mathrm{SIR}_i = \frac{\|M_i(f,t_s)X_i(f,t_s)\|_F^2}{\|M_i(f,t_s)P_i(f,t_s)\|_F^2} \quad X_i(f,t_s)$$

and $P_i(f,t_s)$ are the TF representations of $x_i(t)$ and $p_i(t)$, respectively. $\|\,.\,\|_F$ is the Frobenius norm which is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_f^F \sum_{t_s}^{T_s} a_{f,t_s}}$, where $\mathbf{A}$ is an $F \times T_s$ matrix. We also define the separability of the mixture with respect to all the N audio patterns as

$$\mathrm{Sep} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Sep}_i. \qquad (4)$$

Equation (3) is equivalent to measuring the success of extracting the ith audio pattern $X_i(f,t_s)$ from the mixture $Y(f,t_s)$ given the TF mask $M_i(f,t_s)$. Similarly, Eq. (4) measures the success of extracting all the N patterns simultaneously from the mixture.

We can analyze the time-frequency separability of any mixture as follows:

(1) When $\mathrm{Sep}_i = 1$, this indicates that the mixture $Y(f,t_s)$ is separable with respect to the ith pattern $X_i(f,t_s)$. 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 ith pattern $X_i(f,t_s)$ from the mixture $Y(f,t_s)$. This corresponds to $M_i(f,t_s) = M_i^{\mathrm{opt}}(f,t_s)$. Hence, this is the maximum attainable Sep_i value.

(2) For the case $\mathrm{Sep} < 1$, this implies that at least one $\mathrm{Sep}_i < 1$. In this situation, the ith audio pattern overlaps

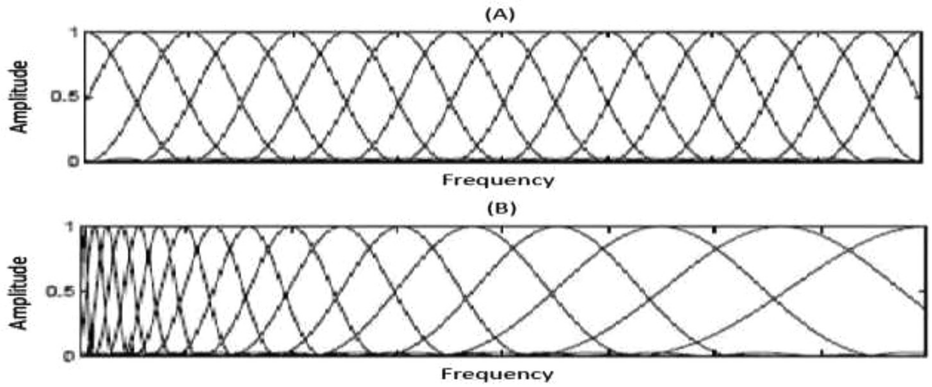


FIG. 1. (a) Normalized frequency responses of 17-channel STFT filter bank. (b) Normalized frequency responses of 17-channel gammatone filter bank.

with other sources in the TF domain and therefore, they cannot be fully separated.

Using the above concept, we can extend the analysis for the case of separating N patterns. A mixture $y(t)$ is fully separable to all the N patterns if and only if $\text{Sep} = 1$ in (4). Thus, Sep provides the quantitative performance measure for evaluating how separable the mixture is in the TF domain. In our comparison, both the spectrogram and cochleagram TF representations are used to test the mixture's separability. All comparison details will be discussed in Sec. IV C.

III. AUDIO PATTERN SEPARATION

In this paper, we derive the *sparsity-learning* two-dimensional non-negative matrix factorization as a pattern separation method. To facilitate the derivation of the algorithm, we consider the signal model in terms of the power TF representation.

A. Audio pattern model

Since the audio signals have time-varying spectra, it is fitting 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 pattern (i.e., original signal). 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).[20,21] This model extends the basic NMF to a two-dimensional convolution of $\mathbf{D}$ and $\mathbf{H}$, i.e., $\mathbf{Y} \approx \sum_{\tau,\phi} \overset{\downarrow\phi}{\mathbf{D}^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}^{\phi}}$, where the vertical arrow in $\overset{\downarrow\phi}{\mathbf{D}^{\tau}}$ denotes the downward shift which moves each element in the matrix down by ϕ rows, and the horizontal arrow in $\overset{\rightarrow\tau}{\mathbf{H}^{\phi}}$ denotes the right shift operator which moves each element in the matrix to the right by τ columns. In scalar representation, the $(k,l)^{\text{th}}$ element in $\mathbf{Y}$ is given by $\mathbf{Y}_{k,l} \approx \sum_{d=1}^{d_{\max}} \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{k-\phi,d} \mathbf{H}^{\phi}_{d,l-\tau}$ where $\mathbf{D}^{\tau}_{k',d'}$ is the (k',τ,d')th element of $\mathbf{D}$ and $\mathbf{H}^{\phi'}_{d',l'}$ is the (d',ϕ',l')th element of $\mathbf{H}$. In pattern separation, this model compactly represents the characteristics of the non-stationary patterns by a time-frequency profile convolved in both time and frequency by a time-frequency weight matrix. $\mathbf{D}^{\tau}_{i}$ represents the spectral basis of the ith pattern in the TF domain and $\mathbf{H}^{\phi}_{i}$ represents the corresponding temporal code for each spectral basis.

The TF representation of the mixture in (1) is given by $Y(k,l) = \sum_i X_i(k,l)$, where $Y(k,l)$, $X_i(k,l)$ denote the TF components which are obtained by applying the gammatone filterbank to the mixture and pattern. Since each component is a function of l and k, we represent this as a $K \times L$ matrix $\mathbf{Y} = [|Y(k,l)|^2]_{l=1,2,\ldots,l_{\max}}^{k=1,2,\ldots,K}$ and $\mathbf{X}_i = [|X_i(k,l)|^2]_{t_s=1,2,\ldots,l_{\max}}^{k=1,2,\ldots,K}$. This therefore enables us to express the power TF representation as $\mathbf{Y} \approx \sum_{i=1}^{I} \mathbf{X}_i$ which we will model as $\mathbf{Y}_{k,l} \approx \sum_{i=1}^{I} \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{k-\phi,i} \mathbf{H}^{\phi}_{i,l-\tau}$. The patterns we seek to determine are $\{|X_i(k,l)|^{.2}\}_{i=1}^{I}$ and this will be obtained by using the matrix factorization as $|\tilde{X}_i(k,l)|^{.2} = \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{k-\phi,i} \mathbf{H}^{\phi}_{i,l-\tau}$. In the following, we propose a novel algorithm to estimate $\mathbf{D}^{\tau}_{k,i}$ and $\mathbf{H}^{\phi}_{i,l}$ from the mixture signal.

B. Algorithm: Sparsity-learning two-dimensional non-negative matrix factorization

To derive the sparsity learning algorithm, we consider the following generative model defined as

$$\mathbf{Y} = \sum_{d=1}^{d_{\max}} \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \overset{\downarrow\phi}{\mathbf{D}^{\tau}_{d}} \overset{\rightarrow\tau}{\mathbf{H}^{\phi}_{d}} + \mathbf{V}. \tag{5}$$

In Eq. (5), it is worth pointing out that *each individual element* in $\mathbf{H}^{\phi}$ is constrained to an exponential distribution $f(\mathbf{H}^{\phi}_{d,l};\lambda^{\phi}_{d,l}) = \lambda^{\phi}_{d,l} e^{-\lambda^{\phi}_{d,l}\mathbf{H}^{\phi}_{d,l}}$ with independent decay parameter $\lambda^{\phi}_{d,l}$. Here, $\mathbf{D}^{\tau}_{d}$ is the dth column of $\mathbf{D}^{\tau}$, $\mathbf{H}^{\phi}_{d}$ is the dth row of $\mathbf{H}^{\phi}$ and $\mathbf{V}$ is assumed to be independently and identically distributed (i.i.d.) as the Gaussian distribution with noise having variance σ^2. The terms $d_{\max}$, $\tau_{\max}$, $\phi_{\max}$, and $l_{\max}$ are the maximum number of columns in $\mathbf{D}^{\tau}$, τ shifts, ϕ shifts, and column length in $\mathbf{Y}$, respectively. This is in contrast with the conventional SNMF2D, where $\lambda^{\phi}_{d,l}$ is simply set to a fixed constant, i.e., $\lambda^{\phi}_{d,l} = \lambda$ for all d, l, ϕ. Such setting imposes uniform constant sparsity on all temporal codes $\mathbf{H}^{\phi}$ which enforces each temporal code to be identical to a fixed distribution according to the selected constant sparsity parameter. The consequence of this uniform constant sparsity has already been discussed in Sec. I. In Sec. IV, we will present the details of the sparsity analysis for pattern separation and evaluate its performance against other existing methods.

C. Formulation of the sparsity learning NMF2D

To facilitate such spectral dictionaries with sparse learning coding, we define $\mathbf{D} = \{\mathbf{D}^{\tau}\}_{\tau=0}^{\tau_{\max}}$, $\mathbf{H} = \{\mathbf{H}^{\phi}\}_{\phi=0}^{\phi_{\max}}$ and $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{\phi}\}_{\phi=0}^{\phi_{\max}}$, and then choose a prior distribution $p(\mathbf{D},\mathbf{H})$ over the factors $\{\mathbf{D},\mathbf{H}\}$ in the analysis equation. According to Bayes' theorem, the log-posterior can be expressed as

$$\log p(\mathbf{D},\mathbf{H}|\mathbf{Y},\sigma^2,\boldsymbol{\lambda}) = \log p(\mathbf{Y}|\mathbf{D},\mathbf{H},\sigma^2) + \log p(\mathbf{D},\mathbf{H}|\boldsymbol{\lambda}) + \text{const.} \tag{6}$$

Thus, the likelihood of the factors $\mathbf{D}$ and $\mathbf{H}$ can be written as

$$p(\mathbf{Y}|\mathbf{D},\mathbf{H},\sigma^2) = N\left(\mathbf{Y}_{k,l}; \sum_{d}\sum_{\tau}\sum_{\phi} \overset{\downarrow\phi}{\mathbf{D}^{\tau}_{d}} \overset{\rightarrow\tau}{\mathbf{H}^{\phi}_{d}}, \sigma^2\right), \tag{7}$$

where $N(\bullet)$ denotes the normal distribution and the second term in Eq. (6) consists of the prior distribution of $\mathbf{D}$ and $\mathbf{H}$ where they are jointly independent. Each element of $\mathbf{H}$ is constrained to be exponentially distributed with independent decay parameters, namely,

$$p(\mathbf{H}|\boldsymbol{\lambda}) = \prod_{\phi}\prod_{d}\prod_{l} \lambda_{d,l}^{\phi} e^{-\lambda_{d,l}^{\phi}\mathbf{H}_{d,l}^{\phi}}. \tag{8}$$

Hence, the negative log likelihood serves as the cost function defined as

$$\begin{aligned} L &\propto \frac{1}{2\sigma^2}\left\| \mathbf{Y} - \sum_{d}\sum_{\tau}\sum_{\phi} \overset{\downarrow\phi}{\mathbf{D}}{}_{d}^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}{}_{d}^{\phi} \right\|_{F}^{2} + f(\mathbf{H}) \\ &= \frac{1}{2\sigma^2}\left\| \mathbf{Y} - \sum_{d}\sum_{\tau}\sum_{\phi} \overset{\downarrow\phi}{\mathbf{D}}{}_{d}^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}{}_{d}^{\phi} \right\|_{F}^{2} \\ &\quad + \sum_{\phi,d,l} \lambda_{d,l}^{\phi}\mathbf{H}_{d,l}^{\phi}\ . \end{aligned} \tag{9}$$

The sparsity term $f(\mathbf{H}) = \sum_{\phi,d,l} \lambda_{d,l}^{\phi}\mathbf{H}_{d,l}^{\phi}$ forms the L_1-norm regularization which is used to resolve the ambiguity by forcing all structure in $\mathbf{H}$ onto $\mathbf{D}$. Therefore, the sparsity of the solution in Eq. (9) is highly dependent on the regularization parameter $\lambda_{d,l}^{\phi}$.

D. Estimation of the dictionary and temporal code

In Eq. (9), we constrain each spectral dictionary to unit length. This can be easily satisfied by normalizing each spectral dictionary according to $\tilde{\mathbf{D}}_{k,d}^{\tau} = \mathbf{D}_{k,d}^{\tau}/\sqrt{\sum_{\tau,k}(\mathbf{D}_{k,d}^{\tau})^2}$ for all $d \in [1\,,\ \ldots, d_{\max}]$. With this normalization, the two-dimensional convolution of the spectral dictionary and temporal codes is now represented as $\tilde{\mathbf{Z}} = \sum_{d}\sum_{\tau}\sum_{\phi} \overset{\downarrow\phi}{\tilde{\mathbf{D}}}{}_{d}^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}{}_{d}^{\phi}$. The derivatives of Eq. (9) corresponding to $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ of the sparsity learning factorization model are given by

$$\frac{\partial L}{\partial \mathbf{D}_{k',d'}^{\tau'}} = -\frac{1}{\sigma^2}\sum_{\phi,l}\left(\mathbf{Y}_{k'+\phi,l} - \tilde{\mathbf{Z}}_{k'+\phi,l}\right)\frac{\partial \tilde{\mathbf{Z}}_{k',l}}{\partial \mathbf{D}_{k',d'}^{\tau'}}, \tag{10}$$

$$\frac{\partial L}{\partial \mathbf{H}_{d',l'}^{\phi'}} = -\frac{1}{\sigma^2}\sum_{\tau,k}\tilde{\mathbf{D}}_{k-\phi',d'}^{\tau}\left(\mathbf{Y}_{k,l'+\tau} - \tilde{\mathbf{Z}}_{k,l'+\tau}\right) + \frac{\partial f(\mathbf{H})}{\partial \mathbf{H}_{d',l'}^{\phi'}}. \tag{11}$$

Thus, by following the approach of Lee and Seung,[8] in matrix notation, the multiplicative learning rules become

$$\mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi}\bullet\frac{\sum_{\tau} \overset{\downarrow\phi^{\mathrm{T}}}{\tilde{\mathbf{D}}}{}^{\tau} \overset{\leftarrow\tau}{\mathbf{Y}}}{\sum_{\tau} \overset{\downarrow\phi^{\mathrm{T}}}{\tilde{\boldsymbol{D}}}{}^{\tau} \overset{\leftarrow\tau}{\tilde{\mathbf{Z}}} + \lambda^{\phi}}, \tag{12}$$

$$\mathbf{D}^{\tau} \leftarrow \tilde{\mathbf{D}}^{\tau}\bullet\frac{\sum_{\phi} \overset{\uparrow\phi}{\mathbf{Y}} \overset{\rightarrow\tau^{\mathrm{T}}}{\mathbf{H}^{\phi}} + \tilde{\mathbf{D}}^{\tau}\mathrm{diag}\left\{\sum_{\tau}\mathbf{1}\left[\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \overset{\rightarrow\tau^{\mathrm{T}}}{\mathbf{H}^{\phi}}\right)\bullet\tilde{\mathbf{D}}^{\tau}\right]\right\}}{\sum_{\phi} \overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \overset{\rightarrow\tau^{\mathrm{T}}}{\mathbf{H}^{\phi}} + \tilde{\mathbf{D}}^{\tau}\ \mathrm{diag}\left\{\sum_{\tau}\mathbf{1}\left[\left(\overset{\uparrow\phi}{\mathbf{Y}} \overset{\rightarrow\tau^{\mathrm{T}}}{\mathbf{H}^{\phi}}\right)\bullet\tilde{\mathbf{D}}^{\tau}\right]\right\}}. \tag{13}$$

In Eq. (12), superscript "T" denotes matrix transpose, "•" is the element wise product and diag(·) denotes a matrix with the argument on the diagonal. The column vectors of $\mathbf{D}^{\tau}$ will be factor-wise normalized to unit length.

E. Estimation of the sparsity learning parameter

Since $\overset{\rightarrow\tau}{\mathbf{H}^{\phi}}$ is obtained directly from the original sparse code matrix $\overset{\rightarrow 0}{\mathbf{H}^{\phi}}$, it suffices to compute only the regularization parameters associated with $\overset{\rightarrow 0}{\mathbf{H}^{\phi}}$. Therefore, we can set the cost function in (9) with $\tau_{\max} = 0$ as

$$\begin{aligned} F(\mathbf{H}) &= \frac{1}{2\sigma^2}\left\| \mathrm{vec}(\mathbf{Y}) - \sum_{\phi=0}^{\phi_{\max}}\left(\mathbf{I}\otimes\overset{\downarrow\phi}{\mathbf{D}}\right)\mathrm{vec}(\mathbf{H}^{\phi}) \right\|_{F}^{2} \\ &\quad + \sum_{\phi=0}^{\phi_{\max}}(\underline{\boldsymbol{\lambda}}^{\phi})^{\mathbf{T}}\mathrm{vec}(\mathbf{H}^{\phi}), \end{aligned} \tag{14}$$

where $\|\ \cdot\ \|_F$ denotes the Frobenius norm, with Vec(·) representing the column vectorization, "⊗" is the Kronecker product, and $\mathbf{I}$ is the identity matrix. Defining the following terms: $\underline{\mathbf{y}} = \mathrm{vec}(\mathbf{Y})$; $\bar{\mathbf{D}} = [\mathbf{I}\otimes\overset{\downarrow 0}{\mathbf{D}}\,\vdots\,\mathbf{I}\otimes\overset{\downarrow 1}{\mathbf{D}}\,\vdots\cdots\ \vdots\,\mathbf{I}\otimes\overset{\downarrow\phi_{\max}}{\mathbf{D}}\,\vdots]$; $\underline{\mathbf{h}} = [\mathrm{vec}(\mathbf{H}^{0})^{\mathbf{T}}\vdots,\ldots,\vdots\,\mathrm{vec}(\mathbf{H}^{\phi_{\max}})^{\mathbf{T}}]^{\mathbf{T}}$ $\underline{\boldsymbol{\lambda}} = [\underline{\boldsymbol{\lambda}}^{0\mathbf{T}}\vdots,\ldots,\vdots\,\underline{\boldsymbol{\lambda}}^{\phi_{\max}\mathbf{T}}]^{\mathbf{T}}$; $\underline{\boldsymbol{\lambda}}^{\phi} = [\lambda_{1,1}^{\phi}, \lambda_{2,1}^{\phi}, \ldots, \lambda_{d_{\max},l_{\max}}^{\phi}]^{\mathbf{T}}$. Thus, Eq. (14) can be rewritten in terms of $\underline{\mathbf{h}}$ as

$$F(\underline{\mathbf{h}}) = \frac{1}{2\sigma^2}\|\underline{\mathbf{y}} - \bar{\mathbf{D}}\underline{\mathbf{h}}\|_{F}^{2} + \underline{\boldsymbol{\lambda}}^{\mathbf{T}}\underline{\mathbf{h}}. \tag{15}$$

Note that $\underline{\mathbf{h}}$ and $\underline{\boldsymbol{\lambda}}$ are vectors of dimension $R\times 1$, where $R = d_{\max}\times l_{\max}\times(\phi_{\max}+1)$. To determine $\underline{\boldsymbol{\lambda}}$, we use the expectation-maximization (EM) algorithm and treat $\underline{\mathbf{h}}$ as the hidden variable where the log-likelihood function can be optimized with respect to $\underline{\boldsymbol{\lambda}}$. Using Jensen's inequality, it can be shown that for any distribution $Q(\underline{\mathbf{h}})$, the log-likelihood function satisfies the following:

$$\ln p\left(\underline{\mathbf{y}}|\underline{\boldsymbol{\lambda}},\bar{\mathbf{D}},\sigma^2\right) \geq \int Q(\underline{\mathbf{h}})\ln\left(\frac{p\left(\underline{\mathbf{y}},\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}},\bar{\mathbf{D}},\sigma^2\right)}{Q(\underline{\mathbf{h}})}\right)d\underline{\mathbf{h}}. \tag{16}$$

One can easily check that the distribution that maximizes the right hand side of Eq. (16) is given by $Q(\underline{\mathbf{h}}) = p(\underline{\mathbf{h}}|\underline{\mathbf{y}},\underline{\boldsymbol{\lambda}},\bar{\mathbf{D}},\sigma^2)$ which is the posterior distribution of $\underline{\mathbf{h}}$. In this paper, we represent the posterior distribution in the form of Gibbs distribution:

$$Q(\underline{\mathbf{h}}) = \frac{1}{Z_h}\exp[-F(\underline{\mathbf{h}})] \quad \text{where} \quad Z_h = \int\exp[-F(\underline{\mathbf{h}})]d\underline{\mathbf{h}}. \tag{17}$$

The functional form of the Gibbs distribution in Eq. (17) is expressed in terms of $F(\underline{\mathbf{h}})$. This is crucial for the purpose of simplifying the variational optimization of $\underline{\boldsymbol{\lambda}}$. The maximum likelihood estimation of $\underline{\boldsymbol{\lambda}}$ can be expressed by

$$\underline{\boldsymbol{\lambda}}^{\mathrm{ML}} = \arg\max_{\underline{\boldsymbol{\lambda}}}\int Q(\underline{\mathbf{h}})\ln p(\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}})d\underline{\mathbf{h}}. \tag{18}$$

Similarly,

$$\sigma^2_{\mathrm{ML}} = \arg\max_{\sigma^2} \int Q(\underline{\mathbf{h}}) \ln p\left(\underline{\mathbf{y}} | \underline{\mathbf{h}}, \sigma^2, \bar{\mathbf{D}}\right) d\underline{\mathbf{h}}. \tag{19}$$

Since each element of $\mathbf{H}$ is constrained to be exponentially distributed with independent decay parameters, this gives $p(\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}}) = \prod_p \lambda_p \exp(-\lambda_p h_p)$. The Gibbs distribution $Q(\underline{\mathbf{h}})$ treats $\underline{\mathbf{h}}$ as the dependent variable while assuming all other parameters to be constant. As such, the functional optimization of $\underline{\boldsymbol{\lambda}}$ in Eq. (18) is obtained by differentiating the terms within the integral with respect to λ_p and the end result is given by

$$\lambda_p = \frac{1}{\int h_p Q(\underline{\mathbf{h}}) d\underline{\mathbf{h}}} \quad \text{for} \quad p = 1, 2, \ldots, R, \tag{20}$$

where λ_p is the pth element of $\underline{\boldsymbol{\lambda}}$. Since

$$p\left(\underline{\mathbf{y}} | \underline{\mathbf{h}}, \bar{\mathbf{D}}, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{N_0/2}} \exp\left(-\frac{1}{2\sigma^2} \|\underline{\mathbf{y}} - \bar{\mathbf{D}}\underline{\mathbf{h}}\|^2\right),$$

where $N_o = K \times L$, the iterative update rule for σ^2_{ML} is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N_0} \int Q(\underline{\mathbf{h}})(\|\underline{\mathbf{y}} - \bar{\mathbf{D}}\underline{\mathbf{h}}\|^2) d\underline{\mathbf{h}}. \tag{21}$$

Despite the simple form of Eqs. (20) and (21), the integral is difficult to compute analytically and therefore, we seek an approximation to $Q(\underline{\mathbf{h}})$. We note that the solution $\underline{\mathbf{h}}$ naturally partitions its elements into distinct subsets $\underline{\mathbf{h}}_P$ and $\underline{\mathbf{h}}_M$ consisting of components $\forall p \in P$ such that $h_p = 0$, and components $\forall m \in M$ such that $h_m > 0$. Thus, $F(\underline{\mathbf{h}})$ can be expressed as follows:

$$F(\underline{\mathbf{h}}) = F(\underline{\mathbf{h}}_M) + F(\underline{\mathbf{h}}_P) + G, \tag{22}$$

where $F(\underline{\mathbf{h}}_M) = 1/2\sigma^2 \|\underline{\mathbf{y}} - \bar{\mathbf{D}}_M \underline{\mathbf{h}}_M\|_F^2 + \underline{\boldsymbol{\lambda}}_M^{\mathbf{T}} \underline{\mathbf{h}}_M$, $F(\underline{\mathbf{h}}_P) = 1/2\sigma^2 \|\underline{\mathbf{y}} - \bar{\mathbf{D}}_P \underline{\mathbf{h}}_P\|_F^2 + \underline{\boldsymbol{\lambda}}_P^{\mathbf{T}} \underline{\mathbf{h}}_P$ and $G = 1/2\sigma^2 \left[2\left(\bar{\mathbf{D}}_M \underline{\mathbf{h}}_M\right)^{\mathbf{T}} \left(\bar{\mathbf{D}}_P \underline{\mathbf{h}}_P\right) - \|\underline{\mathbf{y}}\|^2\right]$.

The term $\|\underline{\mathbf{y}}\|^2$ in G is a constant and the cross-term $(\bar{\mathbf{D}}_M \underline{\mathbf{h}}_M)^{\mathbf{T}} (\bar{\mathbf{D}}_P \underline{\mathbf{h}}_P)$ measures the orthogonality between $\bar{\mathbf{D}}_M \underline{\mathbf{h}}_M$ and $\bar{\mathbf{D}}_P \underline{\mathbf{h}}_P$, where $\bar{\mathbf{D}}_P$ is the sub-matrix of $\bar{\mathbf{D}}$ that corresponds to $\underline{\mathbf{h}}_P$, $\bar{\mathbf{D}}_M$ is the sub-matrix of $\bar{\mathbf{D}}$ that corresponds to $\underline{\mathbf{h}}_M$. In this paper, we intend to simply the expression in Eq. (22) by discounting the contribution from these terms and let $F(\underline{\mathbf{h}})$ be approximated as $F(\underline{\mathbf{h}}) \approx F(\underline{\mathbf{h}}_M) + F(\underline{\mathbf{h}}_P)$. Given this approximation, $Q(\underline{\mathbf{h}})$ can be decomposed as

$$\begin{aligned} Q(\underline{\mathbf{h}}) &\approx \frac{1}{Z_P} \exp[-F(\underline{\mathbf{h}}_P)] \frac{1}{Z_M} \exp[-F(\underline{\mathbf{h}}_M)] \\ &= Q_P(\underline{\mathbf{h}}_P) Q_M(\underline{\mathbf{h}}_M) \end{aligned} \tag{23}$$

with $Z_P = \int \exp[-F(\underline{\mathbf{h}}_P)] d\underline{\mathbf{h}}_P$ and $Z_M = \int \exp[-F(\underline{\mathbf{h}}_M)] d\underline{\mathbf{h}}_M$. Since $\underline{\mathbf{h}}_P = \underline{\mathbf{0}}$ is on the boundary of the distribution, this distribution is represented by using the Taylor expansion about the MAP estimate, $\underline{\mathbf{h}}^{\mathrm{MAP}}$:

$$\begin{aligned} Q_P(\underline{\mathbf{h}}_P \geq 0) &\propto \exp\left\{-\left[\left(\frac{\partial F}{\partial \underline{\mathbf{h}}}\right)\bigg|_{\underline{\mathbf{h}}^{\mathrm{MAP}}}\right]_P^{\mathbf{T}} \underline{\mathbf{h}}_P - \frac{1}{2} \underline{\mathbf{h}}_P^{\mathbf{T}} \bar{\boldsymbol{\Lambda}}_P \underline{\mathbf{h}}_P\right\} \\ &= \exp\left[-\left(\bar{\boldsymbol{\Lambda}} \underline{\mathbf{h}}^{\mathrm{MAP}} - \frac{1}{\sigma^2} \bar{\mathbf{D}}^{\mathbf{T}} \underline{\mathbf{y}} + \underline{\boldsymbol{\lambda}}\right)_P^{\mathbf{T}} \underline{\mathbf{h}}_P \right. \\ &\quad \left. - \frac{1}{2} \underline{\mathbf{h}}_P^{\mathbf{T}} \bar{\boldsymbol{\Lambda}}_P \underline{\mathbf{h}}_P\right], \end{aligned} \tag{24}$$

where $\bar{\boldsymbol{\Lambda}}_P = \left(1/\sigma^2\right) \bar{\mathbf{D}}_P^{\mathbf{T}} \bar{\mathbf{D}}_P$, $\bar{\boldsymbol{\Lambda}} = \left(1/\sigma^2\right) \bar{\mathbf{D}}^{\mathbf{T}} \bar{\mathbf{D}}$. We perform variational approximation to $Q_P(\underline{\mathbf{h}}_P)$ by using the exponential distribution:

$$\hat{Q}_P(\underline{\mathbf{h}}_P \geq 0) = \prod_{p \in P} \frac{1}{u_p} \exp\left(-h_p/u_p\right). \tag{25}$$

The variational parameters $\underline{\mathbf{u}} = \{u_p\}$ for $\forall p \in P$ are obtained by minimizing the Kullback–Leibler divergence between Q_P and $\hat{Q}_P$:

$$\underline{\mathbf{u}} = \arg\min_{\underline{\mathbf{u}}} \int \hat{Q}_P(\underline{\mathbf{h}}_P) \ln \frac{\hat{Q}_P(\underline{\mathbf{h}}_P)}{Q_P(\underline{\mathbf{h}}_P)} d\underline{\mathbf{h}}_P, \tag{26}$$

which leads to

$$\min_{u_p} \hat{\underline{\mathbf{b}}}_P^{\mathbf{T}} \underline{\mathbf{u}} + \frac{1}{2} \underline{\mathbf{u}}^{\mathbf{T}} \hat{\boldsymbol{\Lambda}} \underline{\mathbf{u}} - \sum_{p \in P} \ln u_p, \tag{27}$$

where $\hat{\underline{\mathbf{b}}}_P = \left[\bar{\boldsymbol{\Lambda}} \underline{\mathbf{h}}^{\mathrm{MAP}} - \left(1/\sigma^2\right) \bar{\mathbf{D}}^{\mathbf{T}} \underline{\mathbf{y}} + \underline{\boldsymbol{\lambda}}\right]_P$ and $\hat{\boldsymbol{\Lambda}} = \bar{\boldsymbol{\Lambda}}_P + \mathrm{diag}(\bar{\boldsymbol{\Lambda}}_P)$. The optimization of Eq. (27) can be accomplished by using the non-negative quadratic programming method[28] or Gauss–Newton multiplicative updates.[29] As for components $\underline{\mathbf{h}}_M$, since none of the non-negative constraints are active, we approximate $Q_M(\underline{\mathbf{h}}_M)$ as unconstrained Gaussian with mean $\underline{\mathbf{h}}_M^{\mathrm{MAP}}$. Thus using the factorized approximation $Q(\underline{\mathbf{h}}) = \hat{Q}_P(\underline{\mathbf{h}}_P) Q_M(\underline{\mathbf{h}}_M)$ in Eq. (29), we obtain, $\lambda_p = 1/h_p^{\mathrm{MAP}}$ $\forall p \in M$ and $\lambda_p = 1/u_p \forall p \in P$. for $p = 1, 2, \ldots, R$ and h_p^{MAP} is the pth element of sparse code $\underline{\mathbf{h}}_P$ computed from Eq. (12) and its covariance $\mathbf{C}$ is given by $C_{pm} = \left(\bar{\boldsymbol{\Lambda}}_P^{-1}\right)_{pm}$ $\forall p, m \in M$ and $C_{pm} = u_p^2 \delta_{pm}$ $\forall p, m \in P$. Thus, the update rule for σ^2 can be obtained as

$$\sigma^2 = \frac{1}{N_0} \left[(\underline{\mathbf{y}} - \bar{\mathbf{D}} \widehat{\underline{\mathbf{h}}})^{\mathbf{T}} (\underline{\mathbf{y}} - \bar{\mathbf{D}} \widehat{\underline{\mathbf{h}}}) + \mathrm{Tr}\left(\bar{\mathbf{D}}^{\mathbf{T}} \bar{\mathbf{D}} \mathbf{C}\right)\right], \tag{28}$$

where $\widehat{h}_p = h_p^{\mathrm{MAP}} \forall p \in M$ and $\widehat{h}_p = u_p \forall p \in P$. In order to test the efficacy of our proposed method, we evaluate and compare the proposed method with other existing sparse NMF methods in the application of single channel audio pattern separation in the following section. The specific steps of the proposed method can be summarized in Table II. In Table II, $\psi = 10^{-6}$ is the threshold for ascertaining the convergence.

F. Estimation of audio patterns

The matrices which we seek to separate from $|\mathbf{Y}_{f,t_s}|^{2}$ using our proposed matrix factorization are $\left\{|\mathbf{X}_d|^{.2}\right\}_{d=1}^{d_{\max}}$

TABLE II. Pseudo codes for the proposed algorithm.

```
Input: $|\mathbf{Y}|^{\cdot 2}$, random nonnegative matrix $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$, $\phi$, $\tau$ Output: $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$
Procedure:
Compute initialize cost value $Cost(1)$ using Eq. (9)
for $n = 1$: number of iterations
    Compute $\tilde{\mathbf{Z}} = \sum_d \sum_\tau \sum_\phi \overset{\downarrow\phi}{\tilde{\mathbf{D}}_d^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}_d^{\phi}}$
    Minimize Eq. (27) with respect to $u_p$
    Calculate $\lambda_p$ and $\sigma^2$
    Update $\mathbf{H}_{d',t'_s}^{\phi'}$ using Eq. (12) for all$\tau$, $\phi$
    Compute $\tilde{\mathbf{Z}} = \sum_d \sum_\tau \sum_\phi \overset{\downarrow\phi}{\tilde{\mathbf{D}}_d^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}_d^{\phi}}$
    Update $\mathbf{D}_{f',d'}^{\tau'}$ using Eq. (13) for all $\tau$, $\phi$
    Normalize $\mathbf{D}_{f',i'}^{\tau'}$
    Compute cost value using Eq. (9)
end
stopping criterion: $\frac{\mathrm{Cost}(n-1)-\mathrm{Cost}(n)}{\mathrm{Cost}(n)} < \psi$
```

which is defined as $|\tilde{\mathbf{X}}_d|^{\cdot 2} = \sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{D}_d^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}_d^{\phi}}$ with $\mathbf{D}_d^{\tau}$ and $\mathbf{H}_d^{\phi}$ estimated using Eqs. (12) and (13). Once these matrices are estimated, we form the dth binary mask according to $\mathbf{mask}_d(f, t_s) = 1$ if $|\tilde{X}_d(f, t_s)|^{\cdot 2} > |\tilde{X}_j(f, t_s)|^{\cdot 2}$ $d \neq j$ and zero otherwise. Finally, the estimated time-domain patterns are obtained as $\tilde{\mathbf{x}}_d = \mathrm{Resynthesize}\,(\mathbf{mask}_d \bullet \mathbf{Y})$ where $\tilde{\mathbf{x}}_d = [\tilde{x}_d(1), \ldots, \tilde{x}_d(T)]^{\mathbf{T}}$ denotes the dth estimated pattern. The time-domain estimated patterns are re-synthesized using the approach in Ref. 22 by weighting the mixture cochleagram by the mask and correcting phase shifts introduced during the gammatone filtering.

IV. RESULTS AND ANALYSIS

The proposed separation system has been tested on recorded audio signals. All recordings and processing were conducted using a PC of Intel(R) Core(TM) i5-3317U CPU 1.70 GHz and 4GB RAM. For mixture generation, three types of mixtures are used, i.e., mixture of music and speech and mixture of different kinds of music. The speech signals (male and female) are selected from the TIMIT speech database, while the music signals (jazz and piano) are selected from the RWC database.[30,31] All mixtures are sampled at 16 kHz sampling rate. In all cases, the audio patterns 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 $\phi = \{0, \ldots, 9\}$.

These parameters are selected after conducting Monte-Carlo tests over 20 realizations of audio mixture. We have evaluated our separation performance in terms of the signal-to-distortion ratio (SDR) which 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.[32]

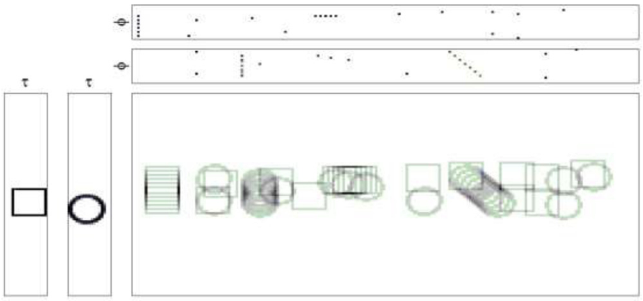


FIG. 2. (Color online) Real basis and code of the simulated mixed data.

A. Impact of automatic sparsity learning and fixed sparsity

In this implementation, several experiments have been conducted to compare the performance of the proposed method with SNMF2D under different sparsity regularizations. To investigate the impact of sparsity regularization on pattern separation performance, the following two cases are analyzed:

Case (i): Uniform constant sparsity, $\lambda_{d,l}^{\phi} = \lambda$ for all d, l, ϕ.

Case (ii): Proposed sparsity, $\lambda_p = 1/h_p^{\mathrm{MAP}}\,\forall p \in M$ and $\lambda_p = 1/u_p\,\forall p \in P$ according to Sec. III E.

B. Computer generated dataset

Figure 2 shows the real basis (i.e., vertical panels) and code (i.e., horizontal panels) of the simulated mixed pattern. The basis $\mathbf{D}$ consists of one circle and one triangle features. These features are convolved with the code $\mathbf{H}$ given at the top panels to yield the data matrix $\mathbf{Y}$ which is a mixture of both patterns.

Figures 3–5 show the matrix factorization results corresponding to each of the above experiments. It is seen that the uniform sparsity factorization with either too small or too large weight of sparsity has failed to identify the correct basis and code. The major reason stems from the high degree of pattern overlap between the circle and square features in the mixed dataset. Since the sparsity is uncontrolled, the larger parts of the overlap will cause increased errors in estimating the basis while the code tends to be more ambiguous.

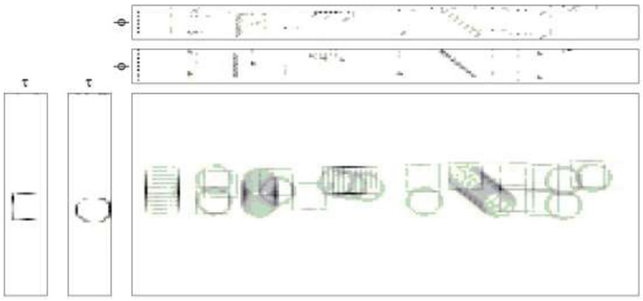


FIG. 3. (Color online) The estimated results based on uniform and constant sparsity factorization with small weight, i.e., $\lambda_{d,l}^{\phi} = 0.01$.

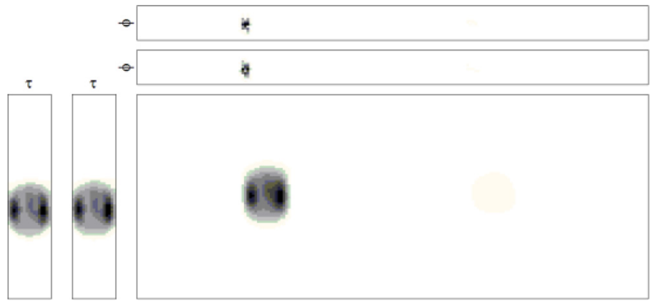


FIG. 4. (Color online) The estimated results based on uniform and constant sparsity factorization with large weight, i.e., $\lambda_{d,l}^{\phi} = 100$.

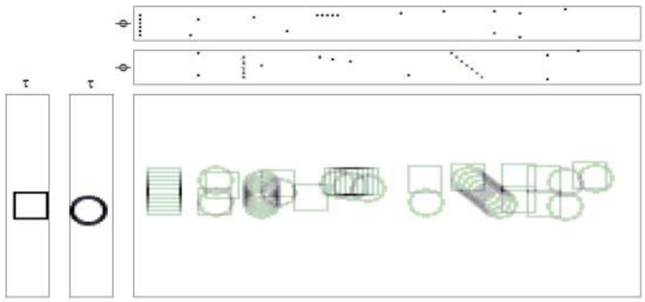


FIG. 5. (Color online) The estimated results based on sparsity learning factorization.

This decreases the possibility of correct assignment of the basis to each feature, and subsequently results in poorer extraction and reconstruction performance. This is shown in Figs. 3 and 4. For example, one could see the extracted codes (i.e., upper panels of Fig. 3) are almost identical and therefore parts of the square and circle features are missing from the figure. In Fig. 4, the codes are too sparse and thereby force the algorithm to extract similar features from the mixed patterns. On the other hand, Fig. 5 shows a better extraction result by using only the sparsity learning parameter. From the above, the analysis results have unanimously indicated the importance of selecting the correct sparsity $\lambda_{d,l}^{\phi}$ for each element code. In the next section, the proposed method will be further tested on the real application of single channel pattern separation. A series of performance comparisons with other matrix factorization methods will also be presented.

C. Single channel audio pattern separation

Figure 6 shows an example of separating a mixture of jazz and male speech by using the conventional SNMF2D

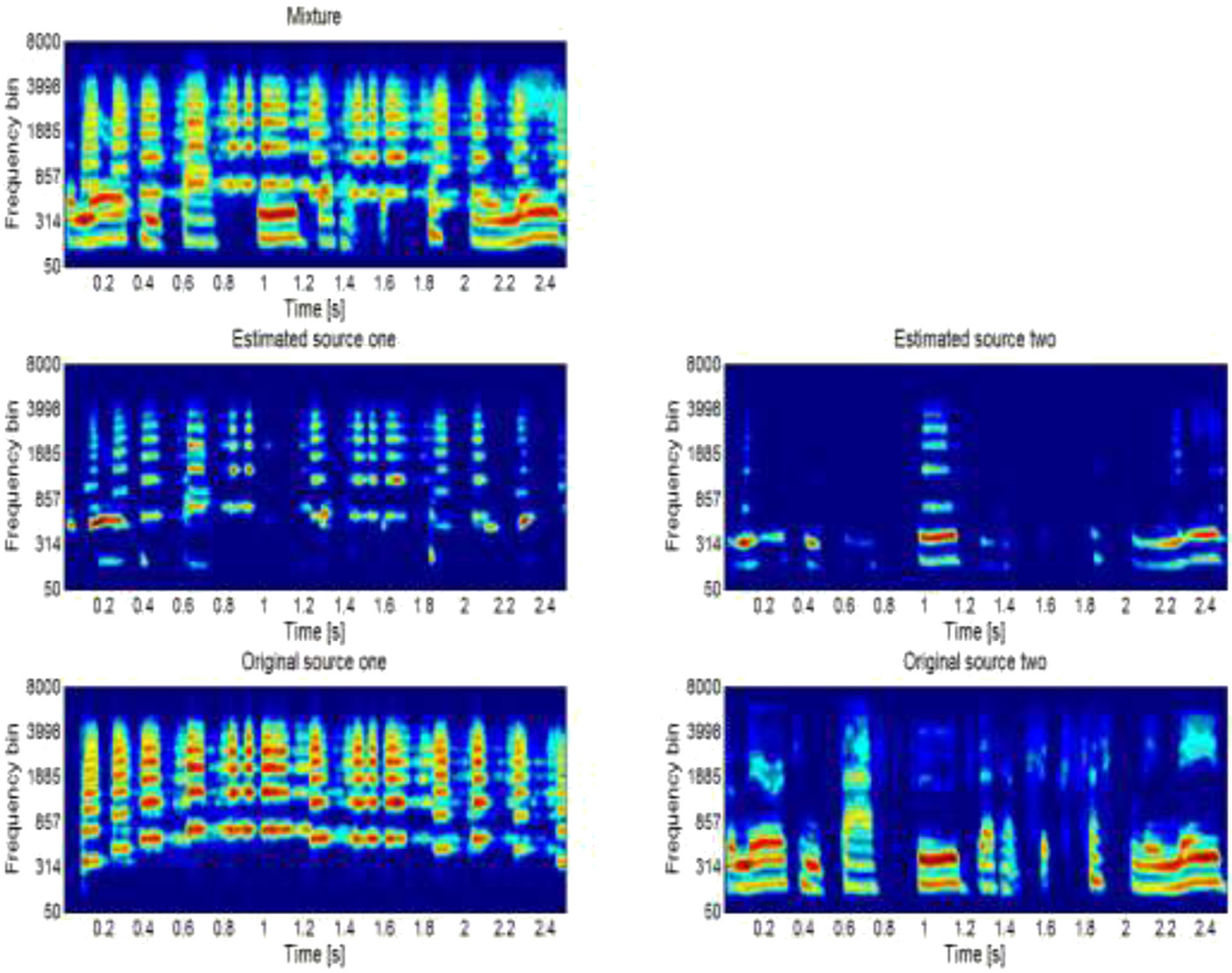


FIG. 6. (Color online) Separation results of case (i): top panel, middle panels, and bottom panels denote the mixture, estimated jazz and male speech, and original jazz and male speech, respectively.

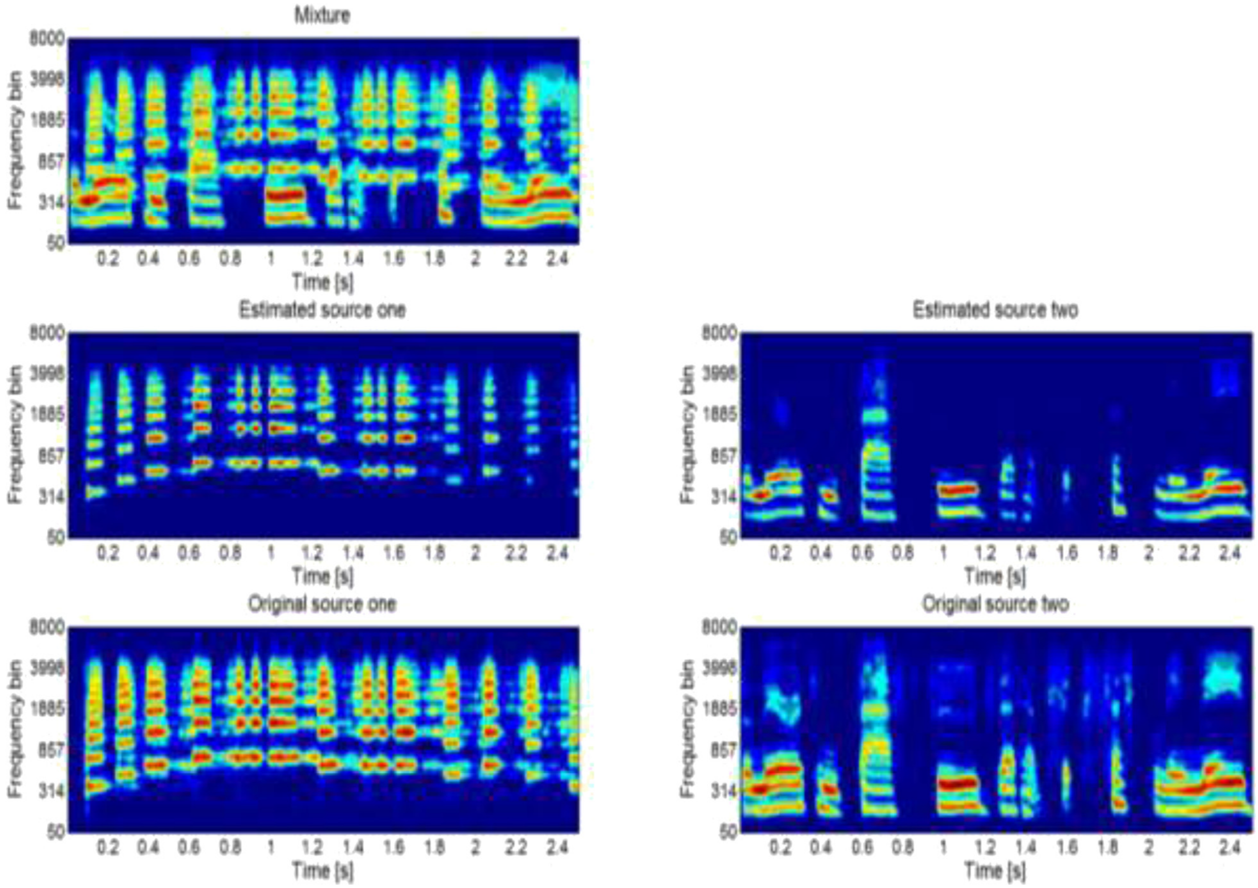
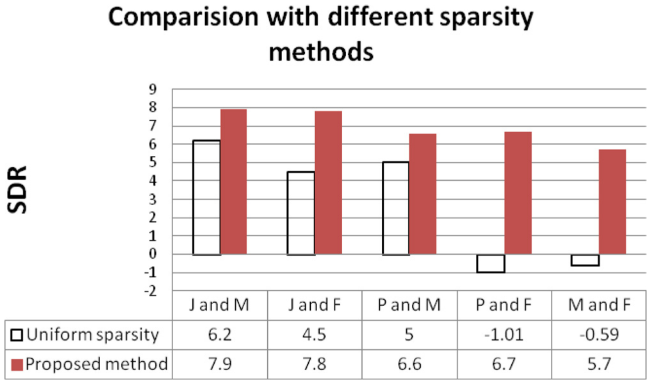


FIG. 7. (Color online) Separation results of case (ii): top panel, middle panels, and bottom panels denote the mixture, estimated jazz, and male speech and original jazz and male speech, respectively.

method. In this method, the sparsity is not fully controlled but is imposed uniformly on all the codes. The ensuing consequence is that the temporal codes are no longer optimal and this leads to "under-sparse" or "over-sparse" factorization which eventually results in inferior separation performance.

In the case of the proposed method as shown in Fig. 7, it assigns a regularization parameter to each temporal code which is individually and adaptively tuned to yield the optimal number of times the spectral dictionary of audio pattern recurs in the spectrogram. The sparsity on $\mathbf{H}_d^{\phi}$ is imposed *element-wise* in the proposed model so that each individual code in $\mathbf{H}_d^{\phi}$ is optimally sparse in the L_1-norm.

A study of the impact of sparsity regularization on the separation results in terms of the SDR under different uniform regularization has been undertaken and the results are summarized in Fig. 8. In this figure, "J" and "P" denotes jazz and piano music, respectively; "M" and "F" denotes male and female speech, respectively. In this implementation, the uniform regularization is chosen as $c = 0, 0.5, \ldots, 10$ for all sparsity parameters, i.e., $\lambda_{d,l}^{\phi} = \lambda = c$. The best result is retained and tabulated in Fig. 8. In our experiments, in the case of jazz and speech mixtures, the best performance is obtained when λ ranges from 1 to 3. As for jazz and piano mixtures, the best performance is obtained when λ ranges from 1.0 to 2.5 and for piano and speech mixtures, the best performance is obtained when λ ranges from 0.5 to 1. From above, it is evident that the uniform sparsity scheme gives varying performance depending on the value of λ which in turn depends on the type of mixture. Hence, this poses a practical difficulty in selecting the appropriate level of sparsity necessary to resolve the ambiguity between the patterns in the TF domain. According to the figure, SNMF2D with sparsity learning tends to yield better results than the uniform sparsity-based methods. We may summarize the average performance improvement of our method against the uniform constant sparsity method as follows: (i) For the jazz and speech, the improvement per source in terms of the SDR is 2.4 dB. (ii) For the piano and jazz music, the improvement per source in terms of SDR is 2.2 dB. (iii) For



	J and M	J and F	P and M	P and F	M and F
□ Uniform sparsity	6.2	4.5	5	-1.01	-0.59
🟥 Proposed method	7.9	7.8	6.6	6.7	5.7

FIG. 8. (Color online) Separation results with different sparsity methods.

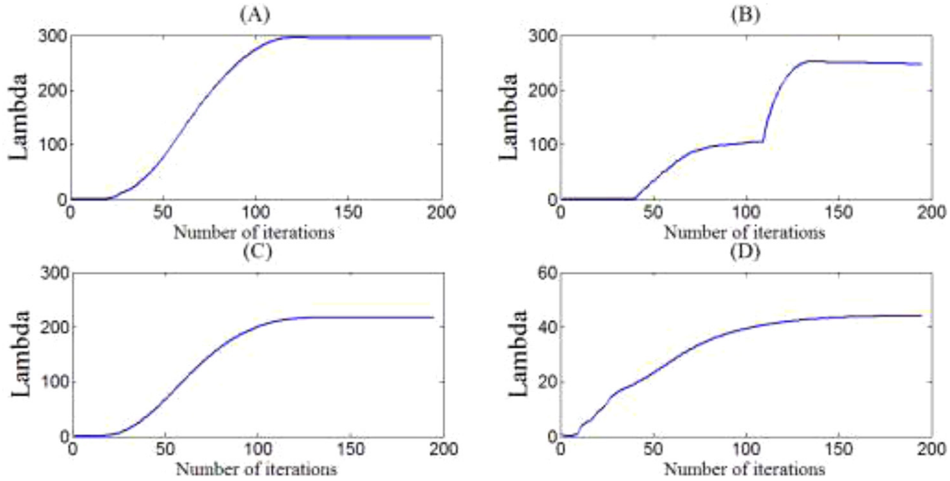


FIG. 9. (Color online) Convergence trajectory of the sparsity: (a) $\lambda_{1,1}^{\phi=0}$, (b) $\lambda_{1,5}^{\phi=0}$, (c) $\lambda_{1,10}^{\phi=0}$, (d) $\lambda_{1,15}^{\phi=0}$.

the piano and speech, the improvement per source in terms of SDR is 1.5 dB.

D. Adaptive behavior of sparsity parameter

In this sub-section, the adaptive behavior of the sparsity parameters by using the proposed method will be demonstrated. Several sparsity parameters have been selected to illustrate its adaptive behavior. Figure 9 shows the convergence trajectory of four sparsity learning parameters $\lambda_{1,1}^{\phi=0}$, $\lambda_{1,5}^{\phi=0}$, $\lambda_{1,10}^{\phi=0}$, and $\lambda_{1,15}^{\phi=0}$ corresponding to their respective element codes. All sparsity parameters are initialized as $\lambda_{d,l}^{\phi}=1$ for all d,l,ϕ and are subsequently adapted according to the proposed method. After 200 iterations, the above sparsity parameters converge to their steady-states. By examining Fig. 8, it is noted that the converged steady-state values are significantly different for each sparsity parameter, e.g., $\lambda_{1,1}^{\phi=0}=294$, $\lambda_{1,5}^{\phi=0}=247$, $\lambda_{1,10}^{\phi=0}=217$, and $\lambda_{1,15}^{\phi=0}=44$ even though they started at the same initial condition. This shows that each element code has its own sparsity. In addition, it is worth pointing out that in the case of jazz and female speech mixture the SDR result rises to 7.8 dB when $\lambda_{d,l}^{\phi}$ is learning. This represents a 3.3 dB per source improvement over the case of uniform constant sparsity (which is only 4.5 dB in Fig. 8). From above, the results suggest that the pattern separation performance has been undermined when the uniform constant sparsity scheme is used. On the contrary, improved performances can be obtained by allowing the sparsity parameters to be individually adapted for each element code. This is evident based on audio separation performance as indicated in Fig. 8.

In addition to the convergence trajectory, we have plotted the histogram of the converged sparsity learning parameters in Fig. 10. The figure suggests that the histogram can be represented very closely as a bimodal distribution. In addition, the figure clearly shows that the codes are optimally sparse as they are predominantly zero or large values.

E. Performance analysis under different TF domains

To ensure a fair comparison, we generate the ideal binary mask (IBM)[27] directly from the original audio signal. To reiterate our aim, the separability analysis is undertaken without recourse to any separation algorithms but utilizing only the pattern of the audio signal to ascertain the degree of "separateness" of the mixture in different TF domains. These results have been tabulated in Fig. 11. The symbols "M" and "S" denotes music and speech, respectively.

In Fig. 11, three types of mixture have been used: (i) music mixed with music, (ii) speech mixed with music, and

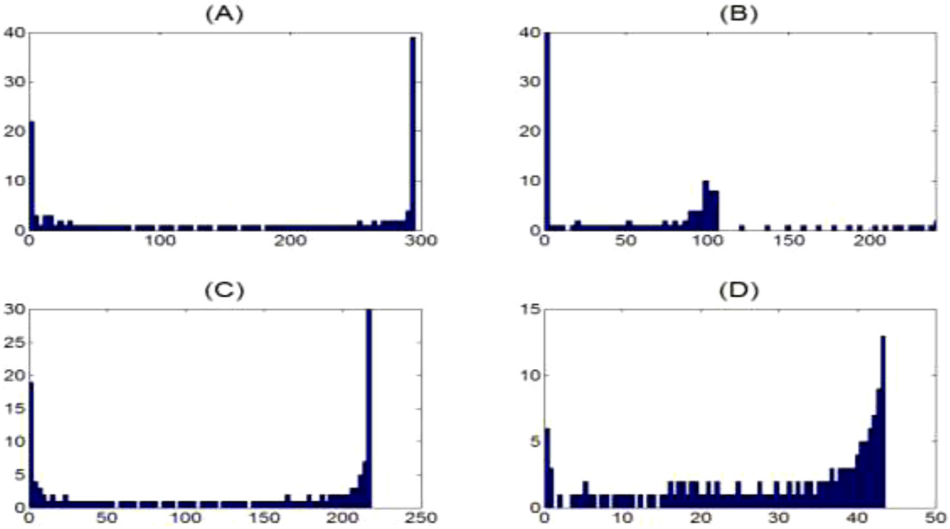
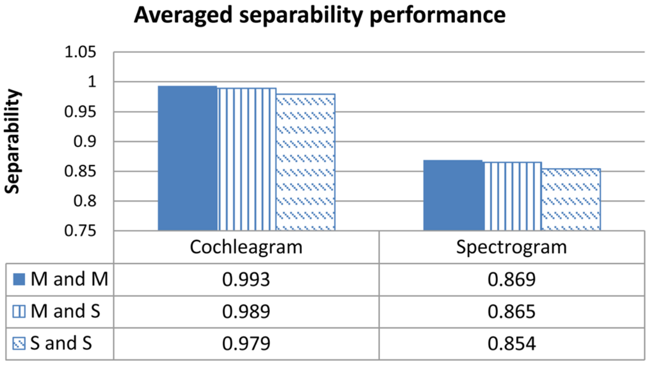


FIG. 10. (Color online) The histogram of the converged sparsity learning parameter.



	Cochleagram	Spectrogram
■ M and M	0.993	0.869
▥ M and S	0.989	0.865
▨ S and S	0.979	0.854

FIG. 11. (Color online) Averaged separability performance.

(iii) speech mixed with speech. The speech signals are selected from 10 male and 10 female speech taken from the TIMIT database and are normalized to unit energy. The 10 music audio patterns are selected from the RWC database[30] and are also normalized to unit energy. Two audio patterns are randomly chosen from the databases and the mixed signal is generated by adding the audio patterns. All mixed signals are sampled at 16 kHz sampling rate.

Following the listening performance test proposed by Yilmaz and Rickard,[32] we conclude that Sep > 0.8 leads to acceptable separation performance. Therefore, all TF representations in Fig. 11 satisfy this condition. While this is true, the spectrogram gives only a mediocre level of separability with averaged Sep ≈ 0.86 while the cochleagram yields the best separability with Sep ≈ 0.98. The analysis conducted above is based on a mixture of two audio patterns. In the following, we extend the separability analysis by increasing the number of audio patterns from two to eight. For mixture of music and speech audio patterns, the number of music audio patterns is selected equal to the number of speech audio patterns (e.g., for mixture of eight audio patterns, four are drawn from music and another four from speech; for mixture of seven patterns, either three (or four) are drawn from music and the remaining four (or three) from speech). The result is shown in Fig. 12. Similar to the above, the separability performance for each TF representation is obtained by averaging over 300 realizations. It is observed that for any number of audio patterns, the cochleagram has shown the best separability performance across all different types of mixture. It is worth pointing out that the cochleagram always retains a high level of separability even when the number of audio patterns increases. Also, the separability decreases steadily as the number of audio patterns increases. On the contrary, the spectrogram fails to separate the mixture when a large number of audio patterns is present, e.g., for mixture of music and speech (eight patterns mixed), Sep ≈ 0.65 for spectrogram. They are below the acceptable level of separability. On the other hand, the cochleagram maintains its performance at a good level with Sep ≈ 0.9 which is well above the rest. Finally, of all the mixture types only the cochleagram preserves the separability larger than 0.8 over the range of eight audio patterns. Therefore, based on this study, it can be concluded that the cochleagram is more suitable to audio TF transform than the spectrogram.

From the above analysis, 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 the separation algorithm is analyzed. Instead of using the IBM, the proposed algorithm is now used to estimate the mask according to Sec. III D. In this

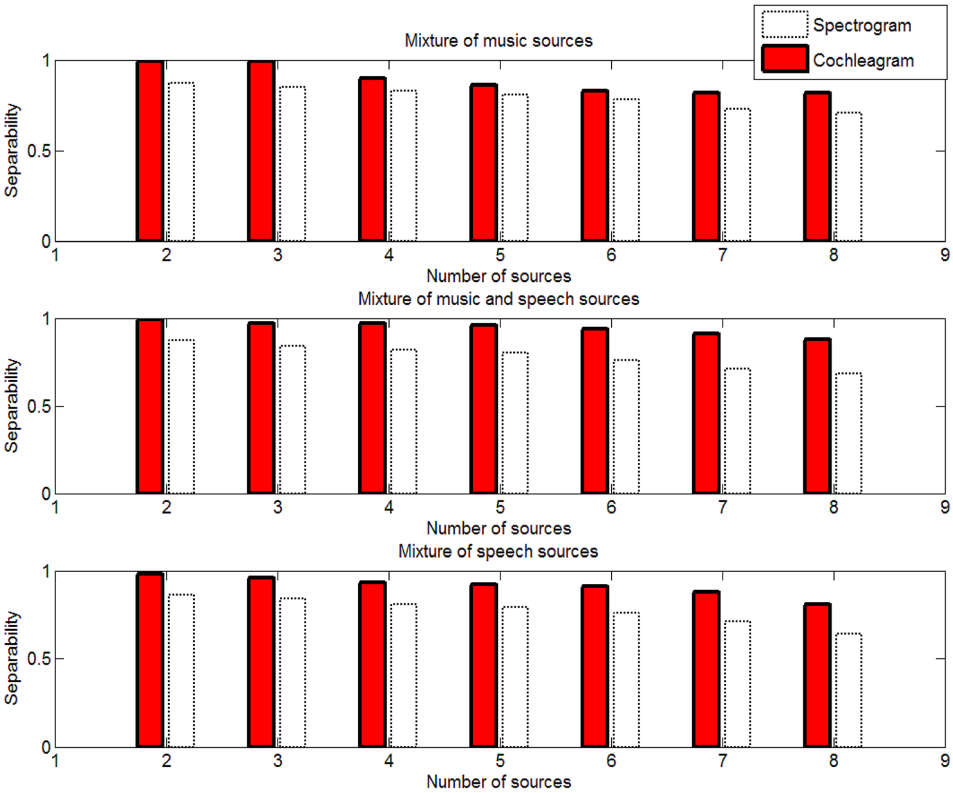


FIG. 12. (Color online) Overall separability for each mixture type.

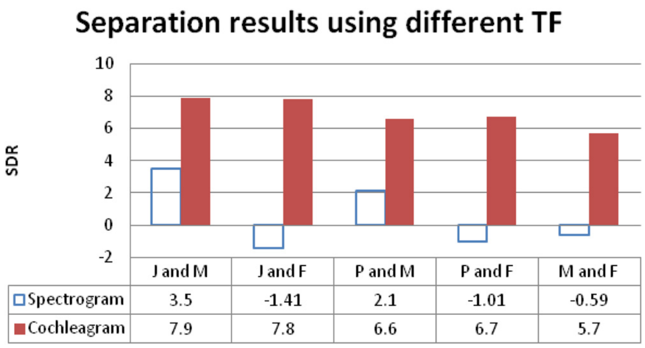


FIG. 13. (Color online) Separation results using different TF representation.

situation, we are investigating the performance of mixture separation (rather than mixture separability). Speech signals and music are used to generate the monoaural mixture recording. The separation performance is evaluated by using two types of TF representation: (i) spectrogram (STFT with 1024-point Hamming windowed FFT and 50% overlap) and (ii) cochleagram based on Gammatone filterbank of 128 channels, filter order of 4 (i.e., $h = 4$, and each filter output is divided into 20 ms time frame with 50% overlap. To validate the parameters setting of the cochleagram (e.g., h and v), we have constructed an experiment based on different audio patterns and tested the result by fixing the parameter h to unity. The experiment is then repeated by progressively increasing h from two 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 $v(f) = 1.019\mathrm{ERB}(f)$, where $\mathrm{ERB}(f) = 24.7 + 0.108f$ is the equivalent rectangular bandwidth of the filter with the center frequency f. A range of values for v has been tested, i.e., $v(f) = (1.019 + c)\mathrm{ERB}(f)$, 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 $v(f) = 1.019\mathrm{ERB}(f)$ for the cochleagram.

Figure 13 shows the comparison of our proposed algorithm based on the 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 a significantly higher performance is attained by the cochleagram with an average SDR of 6.8 dB. This leads to a substantial improvement gain of 6.7 dB. The major reason for the large discrepancy is due to the mixing ambiguity between $|\mathbf{X}_i|^{\cdot 2}$ and $|\mathbf{X}_j|^{\cdot 2}$. The larger the mixing ambiguity between $|\mathbf{X}_i|^{\cdot 2}$ and $|\mathbf{X}_j|^{\cdot 2}$, the more the TF units are ambiguous, which subsequently decreases the probability of correct assignment of each unit to the patterns and eventually results in poorer separation performance. The STFT-based spectrogram lacks any provision for further low-level information of a TF unit and therefore, the resulting spectrogram fails to infer the dominating audio pattern. This leads to a high degree of ambiguity in the TF domain and causes lack of uniqueness in extracting the spectral-temporal features of the patterns. 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 $|\mathbf{X}_i|^{\cdot 2}$ and $|\mathbf{X}_j|^{\cdot 2}$. This explains how the average performance of separating a mixture of jazz music and male utterance is highest among all the mixtures since both audio signals have very distinguishable TF patterns in the cochleagram. This is evident in Fig. 7 which shows the separation results in the cochleagram. The plot clearly shows that the spectral energy of the two audio signals has been clustered at different frequencies in the cochleagram due to their different fundamental frequencies.[33]

F. Pattern tracking analysis

In Sec. IV C, we show that the mixed signal is more separable in the cochleagram than in the spectrogram. In this section, we will analyze the performance of pattern tracking, i.e., identifying the TF patterns that belong to a particular original audio pattern signal, and investigate its performance in both cochleagram and spectrogram. To conduct this analysis, we first split the mixed signal into small segments where the audio mixture is divided into blocks of length 0.65 s with approximately 10^4 samples and then use these segments to illustrate the TF mixing ambiguity in the cochleagram and the spectrogram. Figure 14 shows the TF representations of the audio patterns and the mixture in cochleagram and spectrogram, respectively.

From Fig. 14, we observe two main advantages of using the cochleagram for pattern tracking: First, the mixing ambiguity region is less than those in the spectrograms, as highlighted with a black box in panel (C). Second, the pitches assemble in distinguishable regions corresponding to different types of patterns. However, this is not the case for the spectrogram as the pitches of each pattern (i.e., audio pattern signal) are diversely distributed in that domain. In the following, we investigate the performance of pattern tracking to identify the TF patterns of the audio patterns. Here, the ideal binary mask is used as a reference for comparison. The ideal binary mask for an audio pattern is found for each TF unit by comparing the energy of the target pattern to the energy of all the interfering patterns. Hence, the ideal binary mask produces the optimal SDR gain of all binary masks. We have simulated our proposed algorithm to estimate the optimal mask for the pattern in both cochleagram and spectrogram. Figure 15 shows the comparison of the obtained results.

In Fig. 15, the ideal binary mask results in separation performance with SDR 11.3 dB for spectrogram and SDR 12.8 dB for cochleagram, respectively. However, there is a large difference in the estimated results between the above TF representations (SDR 3.4 dB for spectrogram and SDR 8.8 dB for cochleagram) using the proposed algorithm. It is observed that the pattern tracking in the cochleagram has led to substantially higher accuracy with approximately 5 dB improvement over the spectrogram. In addition, the SDR difference between the ideal and estimated binary mask in the cochleagram is only 3.94 dB whereas the difference is 7.88 dB in the spectrogram which is almost six times more error than the cochleagram. Based on the above results, it is

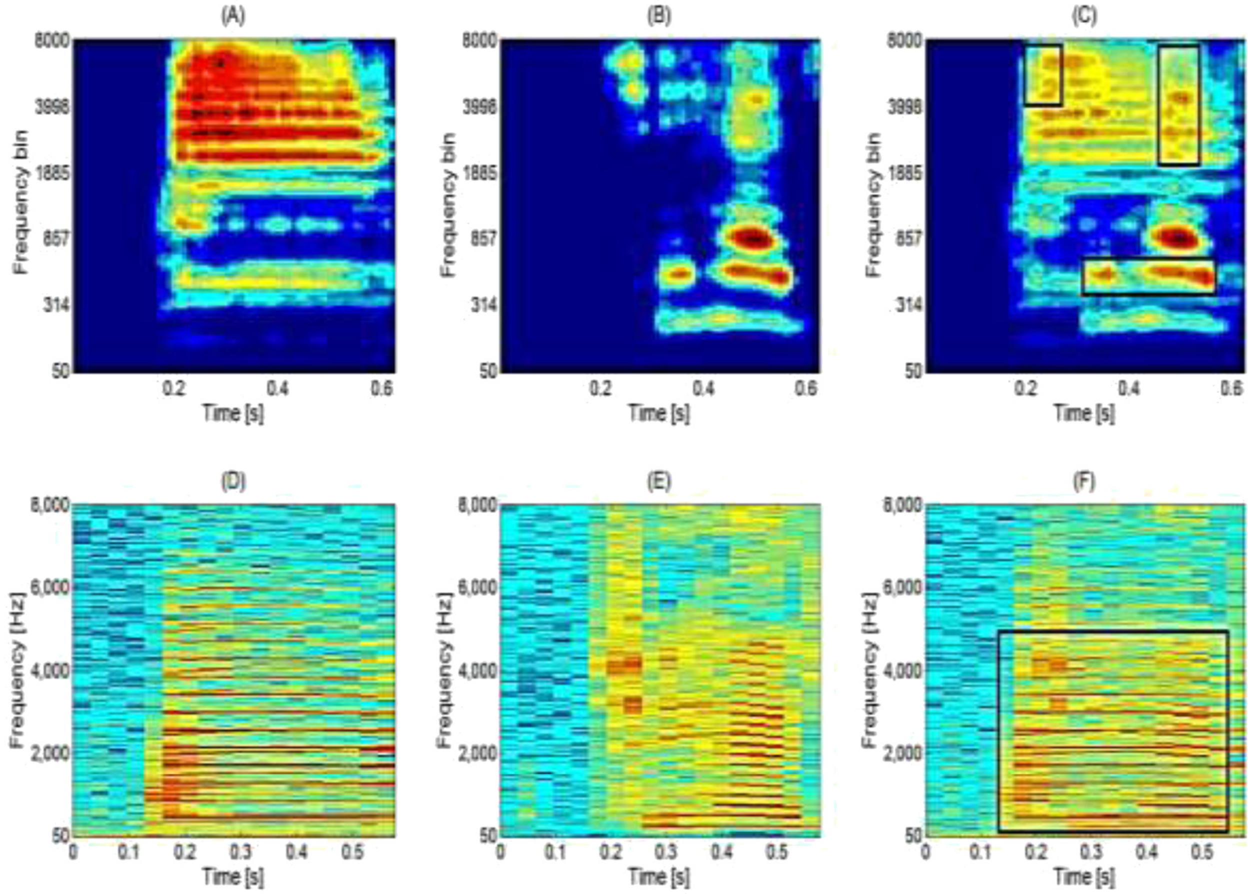


FIG. 14. (Color online) Panels (a)–(c) denote the cochleagram of the jazz music, female speech, and their mixture. Panels (d)–(f) denotes the spectrogram of the same jazz music, female speech, and their mixture.

evident that pattern tracking in the cochleagram is more effective than in the spectrogram. In addition, due to the non-uniform TF resolution, the contents in the cochleagram tend to exist in clusters which may manifest at low or/and high energy levels. This characteristic, however, is unlikely to occur in the spectrogram as the STFT scattered the contents throughout that TF domain. Thus, the use of the cochleagram for SCAPS offers a highly desirable advantage since the separation algorithm can concentrate on those clusters and extract the corresponding spectral-temporal features of each pattern, and finally re-synthesize them to form the estimated signals in the time domain.

G. Comparison with EMD based SCASS approach

We have included the comparison results between our previous paper[34] and the proposed method. The SDR results are summarized as Fig. 16. In the comparison, the proposed

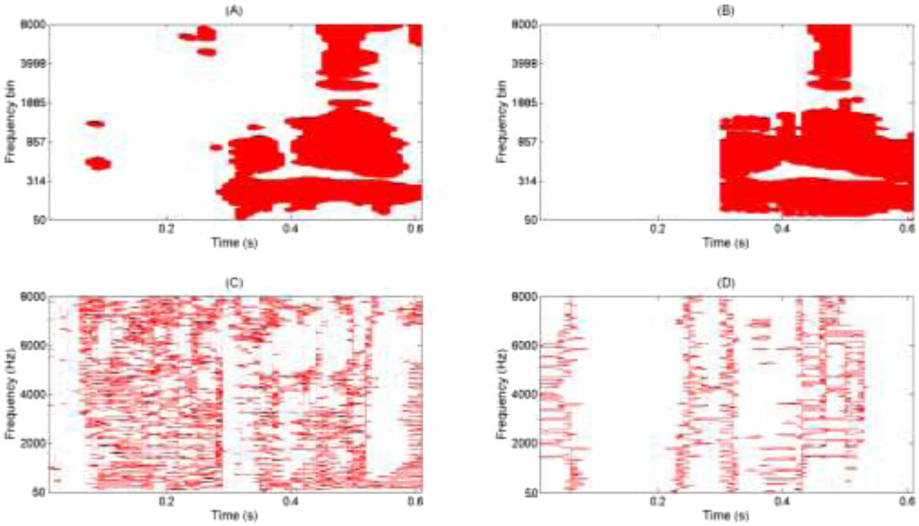
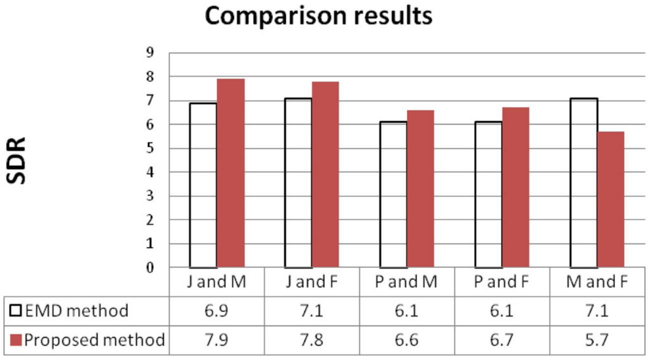


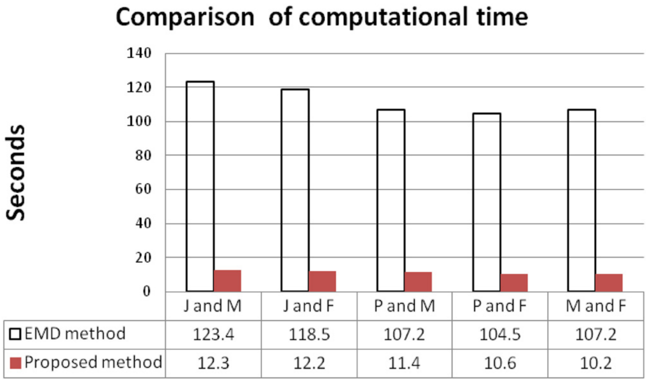
FIG. 15. (Color online) Panels (a)–(b) denote the ideal and estimated binary masks using the proposed algorithm of the female utterance as shown in Fig. 13(b). Panels (c)–(d) denote the ideal and estimated binary masks using the proposed algorithm in the spectrogram of the same female utterance as shown in Fig. 13(e).



	J and M	J and F	P and M	P and F	M and F
□ EMD method	6.9	7.1	6.1	6.1	7.1
🟥 Proposed method	7.9	7.8	6.6	6.7	5.7

FIG. 16. (Color online) Comparison results between EMD method and the proposed method.

Adaptive Sparsity NMF2D with cochleagram has led to comparable better separation performance for most types of mixture except for the mixture of male speech and female speech. The reasons for using EMD as a preprocessing tool for SCBSS has been concluded in our previous paper: Audio signals are mostly non-stationary and the EMD decomposes the mixed signal into a collection of oscillatory basis components termed as intrinsic mode functions (IMFs) which contain the basic properties of the original source (e.g., amplitude and frequency). The impetus behind this is that the degree of mixing of the sources in the IMF domain is now less ambiguous and thus, the dominating source in the mixture is more easily detected. For the mixture of male and female speech, this poses a significant challenge since the fundamental pitches of both signals are too similar for the proposed Adaptive Sparsity NMF2D to separate. In using EMD, the spectral and temporal patterns (i.e., the spectral bases and temporal codes, respectively) associated with each IMF have been decomposed into simpler and sparser patterns than that of the mixed signal. As such, the separation process has been made easier, and thus the EMD method shows better separation performance. For other types of mixture, the spectral-temporal patterns are more distinguishable in the cochleagram domain and thus, the proposed Adaptive Sparsity NMF2D method is able to track the spectral-temporal patterns more effectively. In addition, as the EMD method uses the log-frequency spectrogram for separation which is not an optimal TF representation (this is previously proved in the separability analysis section) for audio mixture, the performance is inferior to the proposed method.



	J and M	J and F	P and M	P and F	M and F
□ EMD method	123.4	118.5	107.2	104.5	107.2
🟥 Proposed method	12.3	12.2	11.4	10.6	10.2

FIG. 17. (Color online) Comparison of computational time between EMD method and the proposed method.

We have also investigated the computational time to run the algorithms. The result is shown in Fig. 17. In comparison, it can be seen that the proposed method is substantially less computationally demanding (approximately 10 times faster) than the EMD method. This is because the EMD method is a three stage process which consists of EMD preprocessing, SNMF2D pattern separation and KLd based K-means classification. However, the proposed method uses only the Adaptive Sparsity NMF2D to separate the cochleagram mixture which saves a lot of computational time.

V. CONCLUSION

In this paper, a novel method for solving the single channel audio pattern separation from a pattern recognition framework has been proposed. The motivation behind this work is that the sparsity achieved by conventional matrix factorization is not enough; in such situations it is crucial to control the degree of sparsity explicitly. In the proposed method, the sparsity term is learned automatically using a variational approach to yield the desired sparse pattern estimation, thus enabling the spectral dictionary and temporal codes of non-stationary audio signals to be estimated more efficiently. Coupled with the theoretical support of signal separability in the TF domain, the separation system using the gammatone filterbank has shown to yield considerable success. This has been concretely verified based on the obtained results and analysis.

[1] P. Hursky, W. S. Hodgkiss, and W. A. Kuperman, "Matched field processing with data-derived modes," J. Acoust. Soc. Am. **109**, 1355–1366 (2001).

[2] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," J. Acoust. Soc. Am. **93**, 510–524 (1993).

[3] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979 (1953).

[4] N. Roman, S. Srinivasan, and D. Wang, "Binaural segregation in multisource reverberant environments," J. Acoust. Soc. Am. **120**, 4040–4051 (2006).

[5] P. Hursky, M. Siderius, M. B. Porter, and V. K. McDonald, "High-frequency (8−16 kHz) model-based source localization," J. Acoust. Soc. Am. **115**, 3021–3032 (2004).

[6] K. Han and D. Wang, "A classification based approach to speech segregation," J. Acoust. Soc. Am. **132**, 3475–3483 (2012).

[7] M. B. Priestley, "Evolutionary spectra and non-stationary processes," J. R. Statist. Soc. Ser. B (Methodological) **27**(2), 204–237 (1965).

[8] D. Lee and H. Seung, "Learning the parts of objects by nonnegative matrix factorisation," Nature **401**(6755), 788–791 (1999).

[9] S. Bucak and B. Gunsel, "Incremental subspace learning via non-negative matrix factorization," Pattern Recognit. **42**(5), 788–797 (2009).

[10] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied topolyphonic music transcription," IEEE Trans. Audio, Speech, Lang. Process. **18**(3), 538–5493 (2010).

[11] W. Liu and N. Zheng, "Non-negative matrix factorization based methods for object recognition," Pattern. Recognit. Lett.**25**(8), 893–897 (2004).

[12] S. Rickard and A. Cichocki, "When is non-negative matrix decomposition unique?," *42nd Annual Conference on Information Sciences and Systems, CISS2008* (March 2008), pp. 1091–1092.

[13] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proceedings of the 5th Conference on Music Information Retrieval (ISMIR'04)*, Spain (October 2004), pp. 318–325.

[14] R. M. Parry and I. Essa, "Incorporating phase information for source separation via spectrogram factorization," in *Proceedings. of the Conference*

on Acoustics, Speech and Signal Processing (ICASSP'07), Hawaii (April 2007), pp. 661–664.

[15] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," Neural Comput. **19**(3), 780–791 (2007).

[16] A. Cichocki, R. Zdunek, and S. I. Amari, "Csisz'ar's divergences for non-negative matrix factorization: family of new algorithms," in *Proceedings of the 6th International Conference on Independent Component Analysis and Signal Separation (ICA'06)*, Charleston, SC (March 2006), pp. 32–39.

[17] T. Virtanen, "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] M. H. Radfa and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," IEEE Trans. Audio, Speech Language Process. **15**(6), 2299–2310 (2007).

[19] S. Roweis, "One microphone source separation," Adv. Neural Inf. Process. Syst. **13**, 793–799 (2000).

[20] M. Morup and M. N. Schmidt, "Sparse non-negative matrix factor 2-D deconvolution," Technical Report, Denmark, 2006.

[21] M. N. Schmidt and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proceedings of the 6th International Conference on Independent Component Analysis and Signal Separation (ICA'06)*, Charleston, SC (March 2006), pp. 700–707.

[22] K. Gröchenig, *Foundations of Time-Frequency Analysis*, 1st ed. (Birkhäuser, Boston, 2001), Chap. 2.

[23] J. C. Brown, "Calculation of a constant Q spectral transform," J. Acoust. Soc. Am. **89**(1), 425–434 (1991).

[24] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Networks **15**(5), 1135–1150 (2004).

[25] R. Curtis, *The Computer Music Tutorial* (MIT Press, Cambridge, MA, 1996), Chap. 7.

[26] S. Schulz and T. Herfet, "Binaural source separation in non-ideal reverberant environments," in *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, Bordeaux, France (September 2007), pp. 10–15.

[27] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer, Norwell, MA, 2005), pp. 181–197.

[28] Y. Q. Lin, "l1-norm sparse Bayesian learning: theory and applications," Ph.D. thesis, University of Pennsylvania, 2008.

[29] F. Sha, L. K. Saul, and D. D. Lee, "Multiplicative updates for nonnegative quadratic programming in support vector machines," Proc Adv. Neural Inf. Process. Systems, **15**, 1041–1048 (2002).

[30] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Baltimore, Maryland (October 2003), pp. 229–230.

[31] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," Signal Process. **92**, 1928–1936 (2012).

[32] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004).

[33] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE **77**(2), 257–286 (1989).

[34] B. Gao, W. L. Woo, and S. S. Dlay, "Single channel source separation using EMD-subband variable regularized sparse features," IEEE Trans. Audio, Speech Lang. Process. **19**, 961–976 (2011).