Adaptive Sparsity Non-Negative Matrix Factorization for Single-Channel Source Separation

Bin Gao, W. L. Woo, *Member, IEEE*, and S. S. Dlay

***Abstract*—A novel method for adaptive sparsity non-negative matrix factorization is proposed. The proposed factorization decomposes an information-bearing matrix into two-dimensional convolution of factor matrices that represent the spectral dictionary and temporal codes. We derive a variational Bayesian approach to compute the sparsity parameters for optimizing the matrix factorization. The method is demonstrated on separating audio mixtures recorded from a single channel. In addition, we have proven that the extraction of the spectral dictionary and temporal codes is significantly more efficient with adaptive sparsity which subsequently leads to better source separation performance. Experimental tests and comparisons with other sparse factorization methods have been conducted to verify the efficacy of the proposed method.**

***Index Terms*—Audio processing, non-negative matrix factorization (NMF), single-channel source separation, sparse features.**

I. Introduction

In recent years, many algorithms have been developed for matrix factorization. These consist of principal component analysis (PCA), independent component analysis (ICA) [1]–[4] and non-negative matrix factorization (NMF) [5]–[8]. Comparing to PCA and ICA, NMF gives a more part-based decomposition [7] and the decomposition is unique under certain conditions [8], making it unnecessary to impose the constraints in the form of orthogonality and independence. These properties have led to a significant interest in NMF lately, e.g., blind source separation (BSS) [9]–[11], data classification [12], [13], data mining [14], pattern recognition [15], object detection [16] and dimensionality reduction [17]. In this paper, we propose a new NMF method for solving BSS problem. In a conventional NMF, given a data matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L] \in \Re_+^{K \times L}$ with $\mathbf{Y}_{k,l} > 0$, NMF factorizes this matrix into a product of two non-negative matrices

$$\mathbf{Y} \approx \mathbf{DH} \tag{1}$$

where $\mathbf{D} \in \Re_+^{K \times \ell}$ and $\mathbf{H} \in \Re_+^{\ell \times L}$ where K and L represent the total number of rows and columns in matrix $\mathbf{Y}$, respectively. If ℓ is chosen to be $\ell = L$, no benefit is achieved at all. Thus the idea is to determine $\ell < L$ so that the matrix $\mathbf{D}$ can be compressed and reduced to its integral components such as $\mathbf{D}_{K \times \ell}$ is a matrix containing a set of dictionary vectors, and $\mathbf{H}_{\ell \times L}$ is an encoding matrix that describes the amplitude of each dictionary vector at each time point. A popular approach to solve the NMF optimization problem is the multiplicative update algorithm by Lee and Seung [7]. Multiplicative update-based families of parameterized cost functions such as the Beta divergence [18], and Csiszar's divergences [19] have also been presented as well. A sparseness constraint [20], [21] can be added to the cost function, and this can be achieved by regularization using the L_1-norm. Here, "sparseness" refers to a representational scheme where only a few units (out of a large population) are effectively used to represent typical data vectors [20]. In effect, this implies most units taking values close to zero while only few take significantly nonzero values. Several other types of prior over $\mathbf{D}$ and $\mathbf{H}$ can be defined, e.g., in [22]–[25], it is assumed that the prior of $\mathbf{D}$ and $\mathbf{H}$ satisfy the exponential density and the prior for the noise variance is chosen as an inverse gamma density. In [26], Gaussian distributions are chosen for both $\mathbf{D}$ and $\mathbf{H}$. The model parameters and hyperparameters are adapted by using the Markov chain Monte Carlo (MCMC) [27]–[29]. In all cases, a fully Bayesian treatment is applied to approximate inference for both model parameters and hyperparameters. While these approaches increase the accuracy of matrix factorization, it only works efficient when large sample dataset is available. Moreover, it consumes significantly high computational complexity at each iteration to adapt the parameters and its hyperparameters. Regardless of the cost function and sparseness constraint being used, the standard NMF or SNMF models [30] are only satisfactory for solving source separation provided that the spectral frequencies of the analyzed audio signal do not change over time. However, this is not the case for many realistic audio signals. As a result, the spectral dictionary obtained via the NMF or SNMF decomposition is not adequate to capture the temporal dependency of the frequency patterns within the signal. The recently developed two-dimensional sparse NMF deconvolution (SNMF2D) model [30]–[32] extends the NMF model to be a two-dimensional convolution of $\mathbf{D}$ and $\mathbf{H}$ where the spectral dictionary and temporal code are optimized using the following cost functions with sparse penalty:

$$C_{LS} : \frac{1}{2} \sum_{k,l} \left(\mathbf{Y}_{k,l} - \tilde{\mathbf{Z}}_{k,l} \right)^2 + \lambda f(\mathbf{H}) \tag{2}$$

$$C_{KL} : \sum_{k,l} \mathbf{Y}_{k,l} \log \frac{\mathbf{Y}_{k,l}}{\tilde{\mathbf{Z}}_{k,l}} - \mathbf{Y}_{k,l} + \tilde{\mathbf{Z}}_{k,l} + \lambda f(\mathbf{H}) \tag{3}$$

Manuscript received September 15, 2010; revised February 22, 2011; accepted June 07, 2011. Date of publication June 27, 2011; date of current version August 17, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Michael Elad.

The authors are with the School of Electrical, Electronic, and Computer Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K. (e-mail: bin.gao@ncl.ac.uk; w.l.woo@ncl.ac.uk; s.s.dlay@ncl.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2011.2160840

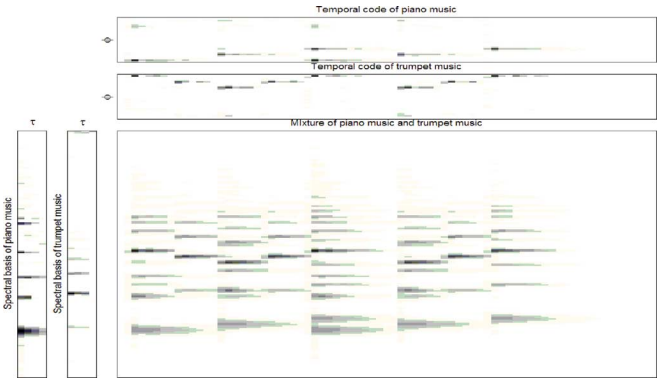

Fig. 1. Estimated spectral dictionary and temporal code of piano and trumpet mixture using SNMF2D.

for $\forall k \in K, \forall l \in L$ where $\tilde{\mathbf{Z}} = \sum_{\tau,\phi} \overset{\downarrow\phi}{\tilde{\mathbf{D}}}{}^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}$, $\tilde{\mathbf{D}}^{\tau}_{k,d} = \mathbf{D}^{\tau}_{k,d} / \sqrt{\sum_{\tau,k} \left(\mathbf{D}^{\tau}_{k,d}\right)^2}$ and $f(\mathbf{H})$ can be any function with positive derivative such as $L_\alpha - norm$ $(\alpha > 0)$ given by $f(\mathbf{H}) = \|\mathbf{H}\|_\alpha = \left(\sum_{\phi,d,l} \left|\mathbf{H}^{\phi}_{d,l}\right|^{\alpha}\right)^{1/\alpha}$. Here $\overset{\downarrow\phi}{\tilde{\mathbf{D}}}{}^{\tau}$ denotes the downward shift which moves each element in the matrix down by ϕ rows, and $\overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}$ denotes the right shift which moves each element in the matrix to the right by τ columns. The SNMF2D separates music mixture based on log-frequency spectrogram. The classic spectrogram decomposes signals to components of linearly spaced frequencies. However, in western music, the typically used frequencies are geometrically spaced. Thus, obtaining an acceptable low-frequency resolution is absolutely necessary, while a resolution that is geometrically related to the frequency is desirable, although not critical. The constant Q transform as introduced in [33], tries to solve both issues. In general, the twelve-tone equal tempered scale which forms the basis of modern western music divides each octave into twelve half notes where the frequency ratio between each successive half note is equal [31]. The fundamental frequency of the note which is k_Q half note above can be expressed as $f^{Q}_{k_Q} = f_{\text{fund}} \cdot 2^{k_Q/24}$. Taking the logarithmic, this gives $\log f^{Q}_{k_Q} = \log f_{\text{fund}} + \frac{k_Q}{24} \log 2$. Thus, in a log-frequency representation the notes are linearly spaced. In our method, the frequency axis of the obtained spectrogram is logarithmically scaled and grouped into 175 frequency bins in the range of 50 Hz to 8 kHz (given f_s =16 kHz) with 24 bins per octave and the bandwidth follows the constant-Q rule. Fig. 1 shows an example of the estimated spectral dictionary $\mathbf{D}$ and temporal code $\mathbf{H}$ based on SNMF2D method on the log-frequency spectrogram.

The $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ matrices can be derived using the cost functions (2) or (3). The SNMF2D is effective in single-channel audio source separation (SCASS) because it is able to capture both the temporal structure and the pitch change of an audio source. However, the drawbacks of SNMF2D originate from its lack of a generalized criterion for controlling the sparsity of $\mathbf{H}$. In practice, the sparsity parameter is set manually. When SNMF2D imposes uniform sparsity on all temporal codes, this is equivalent to enforcing each temporal code to be identical to a fixed distribution according to the selected sparsity parameter. In addition, by assigning the fixed distribution onto each individual code, this is equivalent to constraining all codes to be stationary. However, audio signals are nonstationary in the TF domain and have different temporal structure and sparsity. Hence, they cannot be realistically enforced by a fixed probability distribution. These characteristics are even more pronounced between different types of audio signals. In addition, since the SNMF2D introduces many temporal shifts, this will result in more temporal codes to deviate from the fixed distribution. In such situation, the obtained factorization will invariably suffer from either under- or over-sparseness which subsequently lead to ambiguity in separating the audio mixture. Thus, the above suggests that the present form of SNMF2D is still technically lacking and is not readily suited for SCASS especially mixtures involving different types of audio signals

In this paper, a novel adaptive sparsity two-dimensional non-negative matrix factorization is proposed. Our proposed model allows the following: 1) overcomplete representation by allowing many spectral and temporal shifts which are not inherent in the NMF and SNMF models. Thus, imposing sparseness is necessary to give unique and realistic representations of the non-stationary audio signals. Unlike the SNMF2D, our model imposes sparseness on $\mathbf{H}$ 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. 2) 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 SNMF2D. The proposed method is tested on the application of single channel music separation and the results show that our proposed method can give superior separation performance.

The paper is organized as follows. In Section II, the new model is derived. Experimental results coupled with a series of performance comparison with other NMF techniques are presented in Section III. Finally, Section IV concludes the paper.

II. Proposed Method

In this paper, we derive a new factorization method termed as the *adaptive sparsity* two-dimensional non-negative matrix factorization. The model is given by

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

$$\text{where } \mathbf{H}^{\phi} \sim p\left(\mathbf{H}^{\phi} | \boldsymbol{\lambda}^{\phi}\right) = \prod_{d=1}^{d_{\max}} \prod_{l=1}^{l_{\max}} \lambda^{\phi}_{d,l} \exp\left(-\lambda^{\phi}_{d,l} \mathbf{H}^{\phi}_{d,l}\right). \quad (4)$$

In (4), it is worth pointing out that *each individual element* in $\mathbf{H}^{\phi}$ is constrained to an exponential distribution 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 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 Section I. In Section III, we will present the details of the sparsity analysis for source separation and evaluate its performance against with other existing methods.

A. Formulation of the Proposed Adaptive Sparsity NMF2D

To facilitate such spectral dictionaries with adaptive sparse coding, we first define $\mathbf{D} = [\mathbf{D}^0 \quad \mathbf{D}^1 \quad \cdots \quad \mathbf{D}^{\tau_{\max}}]$, $\mathbf{H} = [\mathbf{H}^0 \quad \mathbf{H}^1 \quad \cdots \quad \mathbf{H}^{\phi_{\max}}]$, and $\boldsymbol{\lambda} = \left[\boldsymbol{\lambda}^1 \boldsymbol{\lambda}^2 \cdots \boldsymbol{\lambda}^{\phi_{\max}}\right]$, and then choose a prior distribution $p(\mathbf{D}, \mathbf{H})$ over the factors $\{\mathbf{D}, \mathbf{H}\}$ in the analysis equation. The posterior can be found by using Bayes' theorem as

$$p\left(\mathbf{D}, \mathbf{H} \middle| \mathbf{Y}, \sigma^2, \boldsymbol{\lambda}\right) = \frac{p\left(\mathbf{Y} \middle| \mathbf{D}, \mathbf{H}, \sigma^2\right) p\left(\mathbf{D}, \mathbf{H} \middle| \boldsymbol{\lambda}\right)}{P(\mathbf{Y})} \tag{5}$$

where the denominator is constant and therefore, the log-posterior can be expressed as

$$\begin{aligned} &\log p\left(\mathbf{D}, \mathbf{H} \middle| \mathbf{Y}, \sigma^2, \boldsymbol{\lambda}\right) \\ &\quad = \log p\left(\mathbf{Y} | \mathbf{D}, \mathbf{H}, \sigma^2\right) + \log p\left(\mathbf{D}, \mathbf{H} \middle| \boldsymbol{\lambda}\right) + \text{const.} \end{aligned} \tag{6}$$

Thus, the likelihood of the factors $\mathbf{D}$ and $\mathbf{H}$ can be written[1] as

$$p\left(\mathbf{Y} | \mathbf{D}, \mathbf{H}, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\left\|\mathbf{Y} - \sum_d \sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{D}}{}_d^\tau \overset{\rightarrow\tau}{\mathbf{H}}{}_d^\phi\right\|_F^2}{2\sigma^2}\right] \tag{7}$$

where $\|.\|_F$ denotes the Frobenius norm. The second term in (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 exponential distributed with independent decay parameters, namely,

$$\begin{aligned} p(\mathbf{H}|\boldsymbol{\lambda}) &= \prod_\phi \prod_d \prod_l \lambda_{d,l}^\phi \exp\left(-\lambda_{d,l}^\phi \mathbf{H}_{d,l}^\phi\right) \\ \text{so that } f(\mathbf{H}) &= \sum_{\phi,d,l} \lambda_{d,l}^\phi \mathbf{H}_{d,l}^\phi. \end{aligned} \tag{8}$$

We constraint $\|\mathbf{D}_d\|_F = 1$ which corresponds to the improper prior

$$p_{\mathbf{D}}(\mathbf{D}) \propto \prod_d \delta\left(\|\mathbf{D}_d\|_F - 1\right). \tag{9}$$

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

$$\begin{aligned} L &\propto \frac{1}{2\sigma^2}\left\|\mathbf{Y} - \sum_d \sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{D}}{}_d^\tau \overset{\rightarrow\tau}{\mathbf{H}}{}_d^\phi\right\|_F^2 + f(\mathbf{H}) - \log p_{\mathbf{D}}(\mathbf{D}) \\ &= \frac{1}{2\sigma^2}\left\|\mathbf{Y} - \sum_d \sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{D}}{}_d^\tau \overset{\rightarrow\tau}{\mathbf{H}}{}_d^\phi\right\|_F^2 + \sum_{\phi,d,l} \lambda_{d,l}^\phi \mathbf{H}_{d,l}^\phi \\ &\quad - \sum_d \log\delta\left(\|\mathbf{D}_d\|_F - 1\right). \end{aligned} \tag{10}$$

The sparsity term $f(\mathbf{H})$ 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 sparseness of the solution in (8) is highly dependent on the regularization parameter $\lambda_{d,l}^\phi$.

1) Estimation of the Dictionary and Temporal Code: In (10), the last term constrains each spectral dictionary to unit length. This can be easily satisfied by normalizing each spectral dictionary according to $\tilde{\mathbf{D}}_{k,d}^\tau = \mathbf{D}_{k,d}^\tau / \sqrt{\sum_{\tau,k} \left(\mathbf{D}_{k,d}^\tau\right)^2}$ 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 (10) corresponding to $\mathbf{D}^\tau$ and $\mathbf{H}^\phi$ of the adaptive sparsity factorization model are given by

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

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{H}_{d',l'}^{\phi'}} &= \frac{\partial}{\partial \mathbf{H}_{d',l'}^{\phi'}}\left(\frac{1}{2\sigma^2}\sum_{k,l}\left(\mathbf{Y}_{k,l} - \tilde{\mathbf{Z}}_{k,l}\right)^2 + f(\mathbf{H})\right) \\ &= -\frac{1}{\sigma^2}\sum_{\tau,k}\tilde{\mathbf{D}}_{k-\phi',d'}^\tau\left(\mathbf{Y}_{k,l'+\tau} - \tilde{\mathbf{Z}}_{k,l'+\tau}\right) + \frac{\partial f(\mathbf{H})}{\partial \mathbf{H}_{d',l'}^{\phi'}}. \end{aligned} \tag{12}$$

Thus, by following the approach of Lee and Seung [5], in matrix notation, the multiplicative learning rules become

$$\mathbf{H}^\phi \leftarrow \mathbf{H}^\phi \bullet \frac{\sum_\tau \overset{\downarrow\phi^{\mathbf{T}}}{\tilde{\mathbf{D}}^\tau} \overset{\leftarrow\tau}{\mathbf{Y}}}{\sum_\tau \overset{\downarrow\phi^{\mathbf{T}}}{\tilde{\mathbf{D}}^\tau} \overset{\leftarrow\tau}{\tilde{\mathbf{Z}}} + \boldsymbol{\lambda}^\phi} \quad \text{here } f(\mathbf{H}) = \sum_{\phi,d,l} \lambda_{d,l}^\phi \mathbf{H}_{d,l}^\phi \tag{13}$$

$$\mathbf{D}^\tau \leftarrow \tilde{\mathbf{D}}^\tau \bullet \frac{\sum_\phi \overset{\uparrow\phi}{\mathbf{Y}} \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^\phi} + \tilde{\mathbf{D}}^\tau diag\left(\sum_\tau \mathbf{1}\left(\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^\phi}\right) \bullet \tilde{\mathbf{D}}^\tau\right)\right)}{\sum_\phi \overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^\phi} + \tilde{\mathbf{D}}^\tau diag\left(\sum_\tau \mathbf{1}\left(\left(\overset{\uparrow\phi}{\mathbf{Y}} \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^\phi}\right) \bullet \tilde{\mathbf{D}}^\tau\right)\right)} \tag{14}$$

where $\tilde{\mathbf{D}}_{k,d}^\tau = \mathbf{D}_{k,d}^\tau / \sqrt{\sum_{\tau,k}\left(\mathbf{D}_{k,d}^\tau\right)^2}$. In (14), superscript "$\mathbf{T}$" denotes matrix transpose, "$\bullet$" is the element wise product, and $diag(\cdot)$ denotes a matrix with the argument on the diagonal. The column vectors of $\mathbf{D}^\tau$ will be factor-wise normalized to unit length.

2) Estimation of the Adaptive Sparsity Parameter: Since $\overset{\rightarrow\tau}{\mathbf{H}}{}^\phi$ is obtained directly from the original sparse code matrix $\overset{\rightarrow 0}{\mathbf{H}}{}^\phi$, it

[1]To avoid cluttering the notation, we shall remove the upper limits from the summation terms. The upper limits can be inferred from (4).

suffices to compute just for the regularization parameters associated with $\overrightarrow{\mathbf{H}}^{\phi}$. Therefore, we can set the cost function in (10) with $\tau_{\max}=0$ as

$$F(\mathbf{H})=\frac{1}{2\sigma^2}\left\|Vec(\mathbf{Y})-\sum_{\phi=0}^{\phi_{\max}}\left(\mathbf{I}\otimes\overrightarrow{\mathbf{D}}^{\downarrow\phi}\right)Vec\left(\mathbf{H}^{\phi}\right)\right\|_F^2+\sum_{\phi=0}^{\phi_{\max}}\left(\underline{\boldsymbol{\lambda}}^{\phi}\right)^{\mathbf{T}}Vec\left(\mathbf{H}^{\phi}\right) \quad (15)$$

with $Vec(\cdot)$ represents the column vectorization, "$\otimes$" is the Kronecker product, and $\mathbf{I}$ is the identity matrix. Defining the following terms:

$$\underline{\mathbf{y}}=Vec(\mathbf{Y}),\quad \overline{\mathbf{D}}=\left[\mathbf{I}\otimes\mathbf{D}^{\downarrow 0}\vdots\mathbf{I}\otimes\mathbf{D}^{\downarrow 1}\vdots\cdots\vdots\mathbf{I}\otimes\mathbf{D}^{\downarrow\phi_{\max}}\right]$$

$$\underline{\mathbf{h}}=\begin{bmatrix}Vec(\mathbf{H}^0)\\ \cdots\\ Vec(\mathbf{H}^1)\\ \cdots\\ \vdots\\ \cdots\\ Vec(\mathbf{H}^{\phi_{\max}})\end{bmatrix},\ \underline{\boldsymbol{\lambda}}=\begin{bmatrix}\underline{\boldsymbol{\lambda}}^0\\ \cdots\\ \underline{\boldsymbol{\lambda}}^1\\ \cdots\\ \vdots\\ \cdots\\ \underline{\boldsymbol{\lambda}}^{\phi_{\max}}\end{bmatrix},\ \underline{\boldsymbol{\lambda}}^{\phi}=\begin{bmatrix}\lambda^{\phi}_{1,1}\\ \lambda^{\phi}_{2,1}\\ \vdots\\ \lambda^{\phi}_{d_{\max},l_{\max}}\end{bmatrix}. \quad (16)$$

Thus, (15) can be rewritten in terms of $\underline{\mathbf{h}}$ as

$$F(\underline{\mathbf{h}})=\frac{1}{2\sigma^2}\left\|\underline{\mathbf{y}}-\overline{\mathbf{D}}\underline{\mathbf{h}}\right\|_F^2+\underline{\boldsymbol{\lambda}}^{\mathbf{T}}\underline{\mathbf{h}}. \quad (17)$$

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 the Jensen's inequality, it can be shown that for any distribution $Q(\underline{\mathbf{h}})$, the log-likelihood function satisfies the following:

$$\ln p\left(\underline{\mathbf{y}}|\underline{\boldsymbol{\lambda}},\overline{\mathbf{D}},\sigma^2\right)\geq\int Q\left(\underline{\mathbf{h}}\right)\ln\left(\frac{p\left(\underline{\mathbf{y}},\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}},\overline{\mathbf{D}},\sigma^2\right)}{Q\left(\underline{\mathbf{h}}\right)}\right)d\underline{\mathbf{h}}. \quad (18)$$

One can easily check that the distribution that maximizes the right-hand side of (18) is given by $Q\left(\underline{\mathbf{h}}\right)=p\left(\underline{\mathbf{h}}|\underline{\mathbf{y}},\underline{\boldsymbol{\lambda}},\overline{\mathbf{D}},\sigma^2\right)$ which is the posterior distribution of $\underline{\mathbf{h}}$. In this paper, we represent the posterior distribution in the form of Gibbs distribution: as follows:

$$Q\left(\underline{\mathbf{h}}\right)=\frac{1}{Z_h}\exp\left[-F\left(\underline{\mathbf{h}}\right)\right]\ \text{ where }\ Z_h=\int\exp\left[-F\left(\underline{\mathbf{h}}\right)\right]d\underline{\mathbf{h}}. \quad (19)$$

The functional form of the Gibbs distribution in (19) is expressed in terms of $F\left(\underline{\mathbf{h}}\right)$ and this is crucial as it will enable us to simplify the variational optimization of $\underline{\boldsymbol{\lambda}}$. The maximum-likelihood estimation of $\underline{\boldsymbol{\lambda}}$ can be expressed by

$$\begin{aligned}\underline{\boldsymbol{\lambda}}^{ML}&=\arg\max_{\underline{\boldsymbol{\lambda}}}\ \ln p\left(\underline{\mathbf{y}}|\underline{\boldsymbol{\lambda}},\overline{\mathbf{D}},\sigma^2\right)\\&=\arg\max_{\underline{\boldsymbol{\lambda}}}\int Q\left(\underline{\mathbf{h}}\right)\ln p\left(\underline{\mathbf{y}},\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}},\overline{\mathbf{D}},\sigma^2\right)d\underline{\mathbf{h}}\\&=\arg\max_{\underline{\boldsymbol{\lambda}}}\int Q\left(\underline{\mathbf{h}}\right)\left(\ln p\left(\underline{\mathbf{y}}|\underline{\mathbf{h}},\sigma^2,\overline{\mathbf{D}}\right)+\ln p\left(\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}}\right)\right)d\underline{\mathbf{h}}\\&=\arg\max_{\underline{\boldsymbol{\lambda}}}\int Q\left(\underline{\mathbf{h}}\right)\ln p\left(\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}}\right)d\underline{\mathbf{h}}. \quad (20)\end{aligned}$$

Similarly,

$$\begin{aligned}\sigma^2_{ML}&=\arg\max_{\sigma^2}\int Q\left(\underline{\mathbf{h}}\right)\left(\ln p\left(\underline{\mathbf{y}}|\underline{\mathbf{h}},\sigma^2,\overline{\mathbf{D}}\right)+\ln p\left(\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}}\right)\right)d\underline{\mathbf{h}}\\&=\arg\max_{\sigma^2}\int Q\left(\underline{\mathbf{h}}\right)\ln p\left(\underline{\mathbf{y}}|\underline{\mathbf{h}},\sigma^2,\overline{\mathbf{D}}\right)d\underline{\mathbf{h}}. \quad (21)\end{aligned}$$

Since each element of $\mathbf{H}$ is constrained to be exponential distributed with independent decay parameters, this gives $p\left(\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}}\right)=\prod_p\lambda_p\exp\left(-\lambda_p h_p\right)$ and therefore, (20) becomes

$$\underline{\boldsymbol{\lambda}}^{ML}=\arg\max_{\underline{\boldsymbol{\lambda}}}\int Q\left(\underline{\mathbf{h}}\right)\left(\ln\lambda_p-\lambda_p h_p\right)d\underline{\mathbf{h}}. \quad (22)$$

The Gibbs distribution $Q\left(\underline{\mathbf{h}}\right)$ 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 (22) is obtained by differentiating the terms within the integral with respect to λ_p and the end result is given by

$$\lambda_p=\frac{1}{\int h_p Q\left(\underline{\mathbf{h}}\right)d\underline{\mathbf{h}}}\quad\text{for }\ p=1,2,\ldots,R \quad (23)$$

where λ_p is the pth element of $\underline{\boldsymbol{\lambda}}$. Since $p\left(\underline{\mathbf{y}}|\underline{\mathbf{h}},\overline{\mathbf{D}},\sigma^2\right)=\left(1/(2\pi\sigma^2)^{N_0/2}\right)\exp(-(1/2\sigma^2)\left\|\underline{\mathbf{y}}-\overline{\mathbf{D}}\underline{\mathbf{h}}\right\|^2)$ where $N_o=K\times L$, the iterative update rule for σ^2_{ML} is given by

$$\begin{aligned}&\sigma^2_{ML}\\&=\arg\max_{\sigma^2}\int Q(\underline{\mathbf{h}})\left(-\frac{N_0}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\left\|\underline{\mathbf{y}}-\overline{\mathbf{D}}\underline{\mathbf{h}}\right\|^2\right)d\underline{\mathbf{h}}\\&=\frac{1}{N_0}\int Q\left(\underline{\mathbf{h}}\right)\left(\left\|\underline{\mathbf{y}}-\overline{\mathbf{D}}\underline{\mathbf{h}}\right\|^2\right)d\underline{\mathbf{h}}. \quad (24)\end{aligned}$$

Despite the simple form of (23) and (24), the integral is difficult to compute analytically and therefore, we seek an approximation to $Q\left(\underline{\mathbf{h}}\right)$. We note that the solution $\underline{\mathbf{h}}$ naturally partition 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, the $F(\underline{\mathbf{h}})$ can be expressed as follows:

$$\begin{aligned}F(\underline{\mathbf{h}})&=\frac{1}{2\sigma^2}\left\|\underline{\mathbf{y}}-\overline{\mathbf{D}}_P\underline{\mathbf{h}}_P-\overline{\mathbf{D}}_M\underline{\mathbf{h}}_M\right\|_F^2+\underline{\boldsymbol{\lambda}}_P^{\mathbf{T}}\underline{\mathbf{h}}_P+\underline{\boldsymbol{\lambda}}_M^{\mathbf{T}}\underline{\mathbf{h}}_M\\&=\underbrace{\frac{1}{2\sigma^2}\left\|\underline{\mathbf{y}}-\overline{\mathbf{D}}_M\underline{\mathbf{h}}_M\right\|_F^2+\underline{\boldsymbol{\lambda}}_M^{\mathbf{T}}\underline{\mathbf{h}}_M}_{F(\underline{\mathbf{h}}_M)}\\&\quad+\underbrace{\frac{1}{2\sigma^2}\left\|\underline{\mathbf{y}}-\overline{\mathbf{D}}_P\underline{\mathbf{h}}_P\right\|_F^2+\underline{\boldsymbol{\lambda}}_P^{\mathbf{T}}\underline{\mathbf{h}}_P}_{F(\underline{\mathbf{h}}_P)}\\&\quad+\underbrace{\frac{1}{2\sigma^2}\left[2\left(\overline{\mathbf{D}}_M\underline{\mathbf{h}}_M\right)^{\mathbf{T}}\left(\overline{\mathbf{D}}_P\underline{\mathbf{h}}_P\right)-\left\|\underline{\mathbf{y}}\right\|^2\right]}_{G}\\&=F(\underline{\mathbf{h}}_M)+F(\underline{\mathbf{h}}_P)+G. \quad (25)\end{aligned}$$

In (25), the term $\left\|\underline{\mathbf{y}}\right\|^2$ in G is a constant and the cross-term $\left(\overline{\mathbf{D}}_M\underline{\mathbf{h}}_M\right)^{\mathbf{T}}\left(\overline{\mathbf{D}}_P\underline{\mathbf{h}}_P\right)$ measures the orthogonality between $\overline{\mathbf{D}}_M\underline{\mathbf{h}}_M$ and $\overline{\mathbf{D}}_P\underline{\mathbf{h}}_P$, where $\overline{\mathbf{D}}_P$ is the sub-matrix of $\overline{\mathbf{D}}$ that corresponds to $\underline{\mathbf{h}}_P$, $\overline{\mathbf{D}}_M$ is the sub-matrix of $\overline{\mathbf{D}}$ that corresponds to $\underline{\mathbf{h}}_M$. In this paper, we intend to simply the expression in (25)

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

$$\begin{aligned} Q(\underline{\mathbf{h}}) &= \frac{1}{Z_h}\exp\left[-F(\underline{\mathbf{h}})\right] \approx \frac{1}{Z_h}\exp\left[-\left(F(\underline{\mathbf{h}}_P)+F(\underline{\mathbf{h}}_M)\right)\right] \\ &= \frac{1}{Z_h}\exp\left[-F(\underline{\mathbf{h}}_P)\right]\exp\left[-F(\underline{\mathbf{h}}_M)\right] \\ &= \frac{1}{Z_P}\exp\left[-F(\underline{\mathbf{h}}_P)\right]\frac{1}{Z_M}\exp\left[-F(\underline{\mathbf{h}}_M)\right] \\ &= Q_P(\underline{\mathbf{h}}_P)Q_M(\underline{\mathbf{h}}_M) \end{aligned} \tag{26}$$

with $Z_P = \int \exp\left[-F\left(\underline{\mathbf{h}}_P\right)\right] d\underline{\mathbf{h}}_P$ and $Z_M = \int \exp\left[-F\left(\underline{\mathbf{h}}_M\right)\right] d\underline{\mathbf{h}}_M$. 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, $\underline{\mathbf{h}}^{MAP}$:

$$\begin{aligned} &Q_P\left(\underline{\mathbf{h}}_P \geq 0\right) \\ &\propto \exp\left\{-\left[\left(\frac{\partial F}{\partial \underline{\mathbf{h}}}\right)\Big|_{\underline{\mathbf{h}}^{MAP}}\right]_P^{\mathbf{T}} \underline{\mathbf{h}}_P - \frac{1}{2}\underline{\mathbf{h}}_P^{\mathbf{T}}\overline{\mathbf{\Lambda}}_P\underline{\mathbf{h}}_P\right\} \\ &= \exp\left[-\left(\overline{\mathbf{\Lambda}}\underline{\mathbf{h}}^{MAP} - \frac{1}{\sigma^2}\overline{\mathbf{D}}^{\mathbf{T}}\underline{\mathbf{y}} + \underline{\boldsymbol{\lambda}}\right)_P^{\mathbf{T}}\underline{\mathbf{h}}_P - \frac{1}{2}\underline{\mathbf{h}}_P^{\mathbf{T}}\overline{\mathbf{\Lambda}}_P\underline{\mathbf{h}}_P\right] \end{aligned} \tag{27}$$

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

$$\hat{Q}_P\left(\underline{\mathbf{h}}_P \geq 0\right) = \prod_{p\in P}\frac{1}{u_p}\exp\left(-\frac{h_p}{u_p}\right). \tag{28}$$

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 $\hat{Q}_P$

$$\begin{aligned} \underline{\mathbf{u}} &= \arg\min_{\underline{\mathbf{u}}} \int \hat{Q}_P\left(\underline{\mathbf{h}}_P\right)\ln\frac{\hat{Q}_P\left(\underline{\mathbf{h}}_P\right)}{Q_P\left(\underline{\mathbf{h}}_P\right)}d\underline{\mathbf{h}}_P \\ &= \arg\min_{\underline{\mathbf{u}}}\int \hat{Q}_P(\underline{\mathbf{h}}_P)[\ln\hat{Q}_P(\underline{\mathbf{h}}_P) - \ln Q_P(\underline{\mathbf{h}}_P)]d\underline{\mathbf{h}}_P \end{aligned} \tag{29}$$

which leads to

$$\min_{u_p}\hat{\underline{\mathbf{b}}}_P^{\mathbf{T}}\underline{\mathbf{u}} + \frac{1}{2}\underline{\mathbf{u}}^{\mathbf{T}}\hat{\mathbf{\Lambda}}\underline{\mathbf{u}} - \sum_{p\in P}\ln u_p \tag{30}$$

where $\hat{\underline{\mathbf{b}}}_P = \left(\overline{\mathbf{\Lambda}}\underline{\mathbf{h}}^{MAP} - (1/\sigma^2)\overline{\mathbf{D}}^{\mathbf{T}}\underline{\mathbf{y}} + \underline{\boldsymbol{\lambda}}\right)_P$ and $\hat{\mathbf{\Lambda}} = \overline{\mathbf{\Lambda}}_P + diag\left(\overline{\mathbf{\Lambda}}_P\right)$. The optimization of (30) can be accomplished by using the non-negative quadratic programming method [34] or Gaussian–Newton multiplicative updates [35]. As for components $\underline{\mathbf{h}}_M$, since none of the non-negative constraints are active, we approximate $Q_M\left(\underline{\mathbf{h}}_M\right)$ as unconstrained Gaussian with mean $\underline{\mathbf{h}}_M^{MAP}$. Thus, using the factorized approximation $Q\left(\underline{\mathbf{h}}\right) = \hat{Q}_P\left(\underline{\mathbf{h}}_P\right)Q_M\left(\underline{\mathbf{h}}_M\right)$ in (23), we obtain the following:

$$\lambda_p = \begin{cases} \frac{1}{h_p^{MAP}}, & \text{if } p \in M \\ \frac{1}{u_p}, & \text{if } p \in P \end{cases} \tag{31}$$

for $p = 1, 2, \ldots, R$ and h_p^{MAP} is the pth element of sparse code $\underline{\mathbf{h}}_P$ computed from (13) and its covariance $\mathbf{C}$ is given by

$$C_{pm} = \begin{cases} \left(\overline{\mathbf{\Lambda}}_P^{-1}\right)_{pm}, & \text{if } p, m \in M \\ u_p^2\delta_{pm}, & \text{Otherwise.} \end{cases} \tag{32}$$

Thus, the update rule for σ^2 computed from (24) can be obtained as

$$\sigma^2 = \frac{1}{N_0}\left[\left(\underline{\mathbf{y}} - \overline{\mathbf{D}}\underline{\hat{\mathbf{h}}}\right)^{\mathbf{T}}\left(\underline{\mathbf{y}} - \overline{\mathbf{D}}\underline{\hat{\mathbf{h}}}\right) + \mathrm{Tr}\left(\overline{\mathbf{D}}^{\mathbf{T}}\overline{\mathbf{D}}\mathbf{C}\right)\right] \tag{33}$$

where $\hat{h}_p = \begin{cases} h_p^{MAP} & \text{if } p \in M \\ u_p & \text{if } p \in P \end{cases}$. 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 source separation in the following section. The specific steps of the proposed method can be summarized as: 1) initialize $\mathbf{D}^\tau$ and $\underline{\mathbf{H}}^\phi$ with nonnegative values; 2) normalize $\tilde{\mathbf{D}}_{k,d}^\tau = \mathbf{D}_{k,d}^\tau / \sqrt{\sum_{\tau,k}\left(\mathbf{D}_{k,d}^\tau\right)^2}$ and compute $\tilde{\mathbf{Z}} = \sum_d\sum_\tau\sum_\phi \overset{\downarrow\phi}{\tilde{\mathbf{D}}}_d^\tau \overset{\rightarrow\tau}{\mathbf{H}}_d^\phi$. 3). Minimize (30) with respect to u_p; 4) calculate λ_p and σ^2 using (31) and (33); 5). update $\mathbf{H}^\phi$ using (13) and re-compute $\tilde{\mathbf{Z}} = \sum_d\sum_\tau\sum_\phi \overset{\downarrow\phi}{\tilde{\mathbf{D}}}_d^\tau \overset{\rightarrow\tau}{\mathbf{H}}_d^\phi$; and 6) update $\mathbf{D}^\tau$ using (14).

III. Single-Channel Source Separation

A. Time-Frequency Representation

The SCASS problem can be treated with one observation and several unknown sources, namely $y(t) = \sum_{d=1}^{d_{\max}} x_d(t)$, where $d = 1, \ldots, d_{\max}$ denotes the sources number and $t = 1, 2, \ldots, T$ denotes the time index. The goal is to estimate the sources $x_d(t)$ when only the observation signal $y(t)$ is available. The time–frequency (TF) representation of the mixture $y(t)$ is given by $Y(f, t_s) = \sum_{d=1}^{d_{\max}} X_d(f, t_s)$, where $Y(f, t_s)$ and $X_d(f, t_s)$ denote the TF components obtained by applying the short time Fourier transform (STFT) on $y(t)$ and $x_d(t)$, respectively, e.g., $Y\left(f, t_s\right) = STFT\left(y\left(t\right)\right)$. The time slots are given by $t_s = 1, 2, \ldots, T_s$ while frequency bins by $f = 1, 2, \ldots, F$. Since each component is a function of t_s and f, we represent this as $\mathbf{Y} = [Y(f, t_s)]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$ and $\mathbf{X}_i = [X_i(f, t_s)]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$. The power spectrogram is defined as the squared magnitude STFT and hence, its matrix representation is given by $|\mathbf{Y}|^{\cdot 2} \approx \sum_{d=1}^{d_{\max}} |\mathbf{X}_d|^{\cdot 2}$ where the superscript "$\cdot$" represents element wise operation. The matrices we seek to determine are $\left\{|\mathbf{X}_d|^{\cdot 2}\right\}_{d=1}^{d_{\max}}$ which will be obtained by using our proposed matrix factorization as $\left|\tilde{\mathbf{X}}_d\right|^{\cdot 2} = \sum_\tau\sum_\phi \overset{\downarrow\phi}{\mathbf{D}}_d^\tau \overset{\rightarrow\tau}{\mathbf{H}}_d^\phi$ with $\mathbf{D}_d^\tau$ and $\mathbf{H}_d^\phi$ estimated using (13) and (14). Once these matrices are estimated, we form the dth binary mask according to $W_d(f, t_s) = 1$ if $\left|\tilde{X}_d(f, t_s)\right|^{\cdot 2} > \left|\tilde{X}_j(f, t_s)\right|^{\cdot 2}$ $d \neq j$ and zero otherwise. Finally, the estimated time-domain sources are obtained as $\tilde{\mathbf{x}}_d = STFT^{-1}\left(\mathbf{W}_d \bullet \mathbf{Y}\right)$, where $\tilde{\mathbf{x}}_d = [\tilde{x}_d(1), \ldots, \tilde{x}_d(T)]^{\mathbf{T}}$ denotes the dth estimated audio sources in the time-domain.

B. Efficiency of Source Extraction in TF Domain

In this subsection, we will analyze how different sparsity factorization methods impact on the source extraction performance in TF domain for SCASS. 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 have no overlap in the TF domain, an optimum mask $W_d^{\text{opt}}(f,t_s)$ (optimal source extractor) exists which allows one to extract the dth original source from the mixture as

$$X_d(f,t_s) = W_d^{\text{opt}}(f,t_s)Y(f,t_s). \tag{34}$$

Given any TF mask $W_d(f,t_s)$ (source extractor) such that $0 \leq W_d(f,t_s) \leq 1$ for all (f,t_s), we define the efficiency of source extraction (ESE) in the TF domain for target source $x_d(t)$ in the presence of the interfering sources $\beta_d(t) = \sum_{j=1,j\neq d}^{d_{\max}} x_j(t)$ as

$$\psi(W_d) \triangleq \frac{\|W_d(f,t_s)X_d(f,t_s)\|_F^2}{\|X_d(f,t_s)\|_F^2} - \frac{\|W_d(f,t_s)B_d(f,t_s)\|_F^2}{\|X_d(f,t_s)\|_F^2} \tag{35}$$

where $X_d(f,t_s)$ and $B_d(f,t_s)$ are the TF representations of $x_d(t)$ and $\beta_d(t)$, respectively. The above represents the normalized energy difference between the extracted source and interferences. We also define the ESE of the mixture with respect to all the $d_{\max}$ sources as

$$\Omega = \frac{1}{d_{\max}} \sum_{d=1}^{d_{\max}} \psi(W_i). \tag{36}$$

Equation (35) is equivalent to measuring the ability of extracting the dth source $X_d(f,t_s)$ from the mixture $Y(f,t_s)$ given the TF mask $W_d(f,t_s)$. Equation (36) measures the ability of extracting all the $d_{\max}$ sources simultaneously from the mixture. To further study the ESE, we use the following two criteria [36]: 1) preserved signal ratio (PSR) which determines how well the mask preserves the source of interest and 2) signal-to-interference ratio (SIR) which indicates how well the mask suppresses the interfering sources:

$$\begin{aligned} PSR_{W_d}^{X_d} &\triangleq \frac{\|W_d(f,t_s)X_d(f,t_s)\|_F^2}{\|X_d(f,t_s)\|_F^2} \\ SIR_{W_d}^{X_d} &\triangleq \frac{\|W_d(f,t_s)X_d(f,t_s)\|_F^2}{\|W_d(f,t_s)B_d(f,t_s)\|_F^2} \quad . \end{aligned} \tag{37}$$

Using (37), (35) can be expressed as $\psi(W_d) = PSR_{W_d}^{X_d} - PSR_{W_d}^{X_d}/SIR_{W_d}^{X_d}$. Analyzing the terms in (34), we have

$$\begin{aligned} PSR_{W_d}^{X_d} &:= \begin{cases} 1, & \text{if supp } W_d^{\text{opt}} = \text{supp } W_d \\ < 1, & \text{if supp } W_d^{\text{opt}} \subset \text{supp } W_d \end{cases} \\ SIR_{W_d}^{X_d} &:= \begin{cases} \infty, & \text{if supp}\,[W_d X_d] \cap \text{supp}\, B_d = \emptyset \\ \text{finite}, & \text{if supp}\,[W_d X_d] \cap \text{supp}\, B_d \neq \emptyset \end{cases} \end{aligned} \tag{38}$$

where "supp" denotes the support. When $\psi(W_d) = 1$ (i.e., $PSR_{W_d}^{X_d} = 1$ and $SIR_{W_d}^{X_d} = \infty$), this indicates that the mixture $y(t)$ is separable with respect to the dth source $x_d(t)$. In other words, $X_d(f,t_s)$ does not overlap with $B_d(f,t_s)$ and the TF mask $W_d(f,t_s)$ has perfectly separated the dth source $X_d(f,t_s)$ from the mixture $Y(f,t_s)$. This corresponds to $W_d(f,t_s) = W_d^{\text{opt}}(f,t_s)$ in (34). Hence, this is the maximum attainable $\psi(W_d)$ value. For other cases of $PSR_{W_d}^{X_d}$ and $SIR_{W_d}^{X_d}$, we have $\psi(W_d) < 1$. Using the above concept, we can extend the analysis for the case of separating $d_{\max}$ sources. A mixture $y(t)$ is fully separable to all the N sources if and only if $\Omega = 1$ in (36). For the case $\Omega < 1$, this implies that some of the sources overlap with each other in the TF domain and therefore, they cannot be fully separated. Thus, Ω provides the quantitative performance measure to evaluate how separable the mixture is in the TF domain. In the following, we show the analysis of how different sparsity factorization methods affect the ESE of the mixture

IV. Results and Analysis

A. Experiment Setup

The proposed method is tested by separating music sources. Several experimental simulations under different conditions have been designed to investigate the efficacy of the proposed method. All simulations and analyses are performed using a PC with Intel Core 2 CPU 6600 at 2.4 GHz and 2 GB RAM. MATLAB is used as the programming platform. To generate mixed signal, we have analyzed a 4-s polyphonic music containing trumpet and piano. The mixed signal is sampled at 16–kHz sampling rate. In addition to above polyphonic music mixture, we have also tested the proposed method in the wider types of music mixtures. Thirty music signals including ten jazz, ten piano, and ten trumpet signals are selected from the RWC [37] database. Three types of mixture have been generated: 1) jazz mixed with piano; 2) jazz mixed with trumpet; and 3) piano mixed with trumpet. The sources are randomly chosen from the database and the mixed signal is generated by adding the chosen sources. In all cases, the sources are mixed with equal average power over the duration of the signals. The TF representation is computed by normalizing the time-domain signal to unit power and computing the STFT using 2048 point Hanning window FFT with 50% overlap. The frequency axis of the obtained spectrogram is then logarithmically scaled and grouped into 175 frequency bins in the range of 50 Hz to 8 kHz with 24 bins per octave. This corresponds to twice the resolution of the equal tempered musical scale. For the proposed adaptive sparsity factorization model, the convolutive components in time and frequency are selected to be 1) for piano and trumpet mixture $\tau = \{0,\ldots,3\}$ and $\phi = \{0,\ldots,31\}$, respectively; 2) for piano and jazz mixture $\tau = \{0,\ldots,6\}$ and $\phi = \{0,\ldots,9\}$, respectively; 3) for trumpet and jazz mixture $\tau = \{0,\ldots,6\}$ and $\phi = \{0,\ldots,9\}$, respectively. The corresponding sparse factor was determined by (31). We have evaluated our separation performance in terms of the signal-to-distortion ratio (SDR) which is one form of perceptual measure. This is a global measure that unifies source-to-interference ratio (SIR), source-to-artifacts ratio (SAR), and source-to-noise ratio (SNR). MATLAB routines for computing these criteria are obtained from the SiSEC'08 webpage [38], [39].

B. Impact of Adaptive and Fixed Sparsity

In this implementation, we have conducted several experiments to compare the performance of the proposed method with

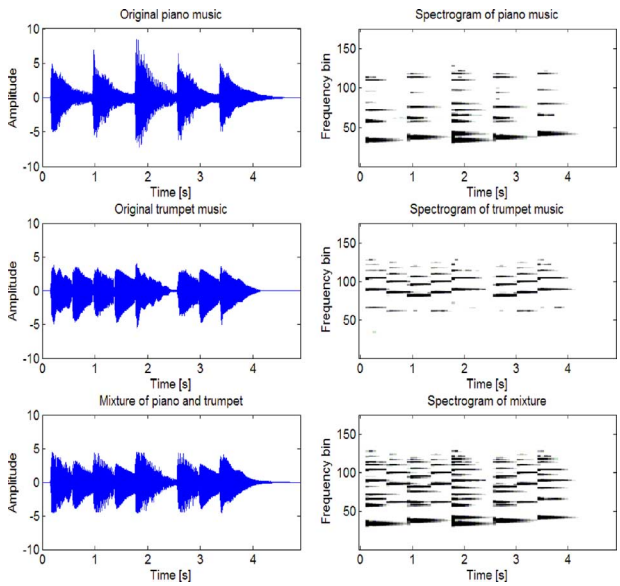

Fig. 2. Time-domain representation and spectrogram of the piano music (top panels), trumpet music (middle panels), and mixed signal (bottom panels).

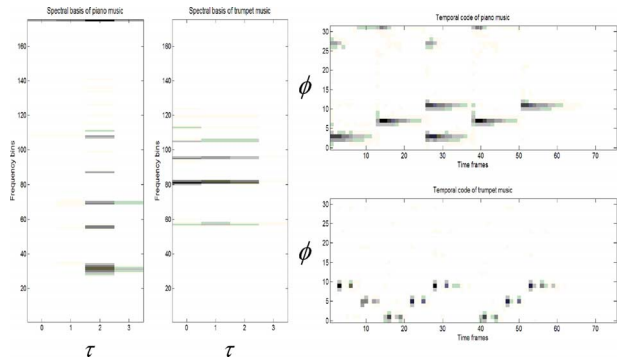

Fig. 3. Estimated $\mathbf{D}_d^\tau$ and $\mathbf{H}_d^\phi$ for Case 1).

SNMF2D under different sparsity regularization. To investigate the impact of sparsity regularization on source separation performance, three cases[2] are conducted:

Case 1) Uniform constant sparsity with low sparseness, $\lambda_{d,l}^\phi = \lambda = 0.01$ for all d, l, ϕ.

Case 2) Uniform constant sparsity with high sparseness, $\lambda_{d,l}^\phi = \lambda = 100$ for all d, l, ϕ.

Case 3) Proposed adaptive sparsity according to (31).

Fig. 2 shows the time and TF domains of the original trumpet, piano music and its mixture. The trumpet and the piano play a different short melodic passage each consisting of three distinct notes. However, both trumpet and piano overlap in time, and the piano notes are interspersed in frequency with the trumpet notes. Hence, this is a challenging task for single-channel separation which will test the impact of sparsity for matrix factorization.

1) Estimated Spectral Dictionary and Temporal Codes: Figs. 3–5 show the matrix factorization results in terms of the spectral dictionary $\mathbf{D}_d^\tau$ and temporal codes $\mathbf{H}_d^\phi$ for Cases 1)–3), respectively. Fig. 3 shows the case of "under-sparse" factorization which is clearly evident by the spreading of the estimated

[2]Cases 1) and 2) correspond to the two-dimensional sparse non-negative matrix deconvolution (SNMF2D) [30], [31]. This section therefore presents the comparison of our proposed method with the SNMF2D with uniform constant sparsity.

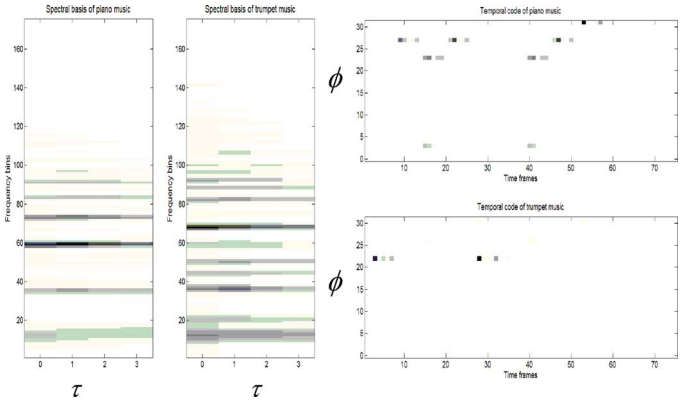

Fig. 4. Estimated $\mathbf{D}_d^\tau$ and $\mathbf{H}_d^\phi$ for Case 2).

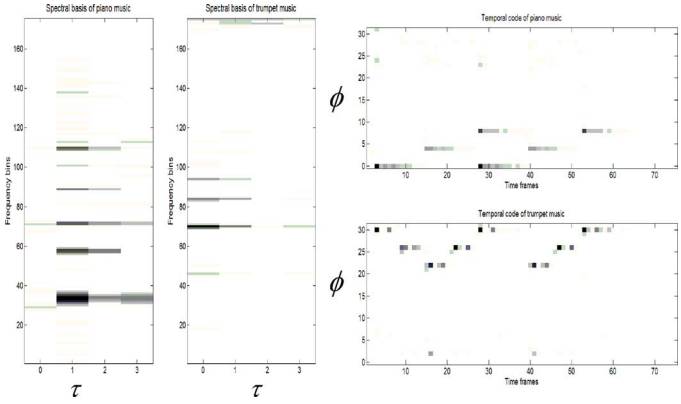

Fig. 5. Estimated $\mathbf{D}_d^\tau$ and $\mathbf{H}_d^\phi$ for Case 3).

temporal codes. Fig. 4 shows the case of "over-sparse" factorization where some of the temporal codes have been discarded. On the other hand, Fig. 5 shows the case of "optimally-sparse" factorization based on the proposed adaptive tuning of the sparsity parameter.

2) Audio Source Separation Results: In above, the analysis of the sparsity factorization was presented in terms of $\mathbf{D}_d^\tau$ and $\mathbf{H}_d^\phi$. In the following, the audio source separation results for each case are shown. In particular, Figs. 6 and 7 show the separated sources in terms of spectrogram and time-domain representation, respectively. Panels (C)–(H) in both Figs. 6 and 7 clearly show that better source separation results require careful selection of the sparsity regularization. In the case of "under-sparse" factorization [e.g., (C)-(D)], the factorization still contains the mixed components (as indicated by the red box marked area) in each separated source. In the case of over-sparse factorization [e.g., (E)-(F)], the spectral dictionary of the source occurs too rarely in the spectrogram and this results in lesser information which do not fully recover the original source as noted in the middle panels (indicated by the red box marked area). In the case of the proposed method [e.g., (G)-(H)], 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 a source 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. In the conventional SNMF2D 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

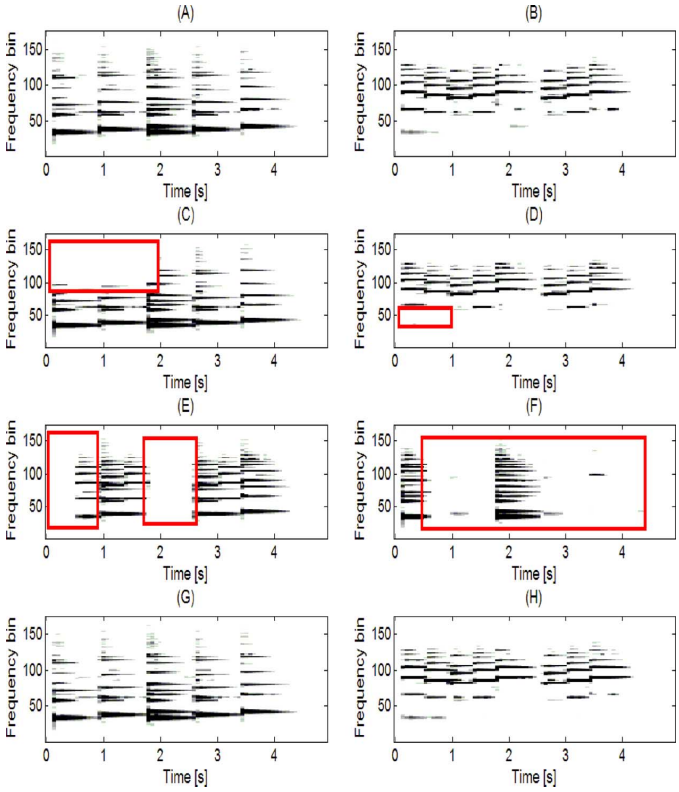

Fig. 6. Separated signals in spectrogram. (A)–(B): original piano and trumpet music. (C)–(D): piano and trumpet music for Case 1). (E)–(F): piano and trumpet music for Case 2). (G)–(H): piano and trumpet music for Case 3).

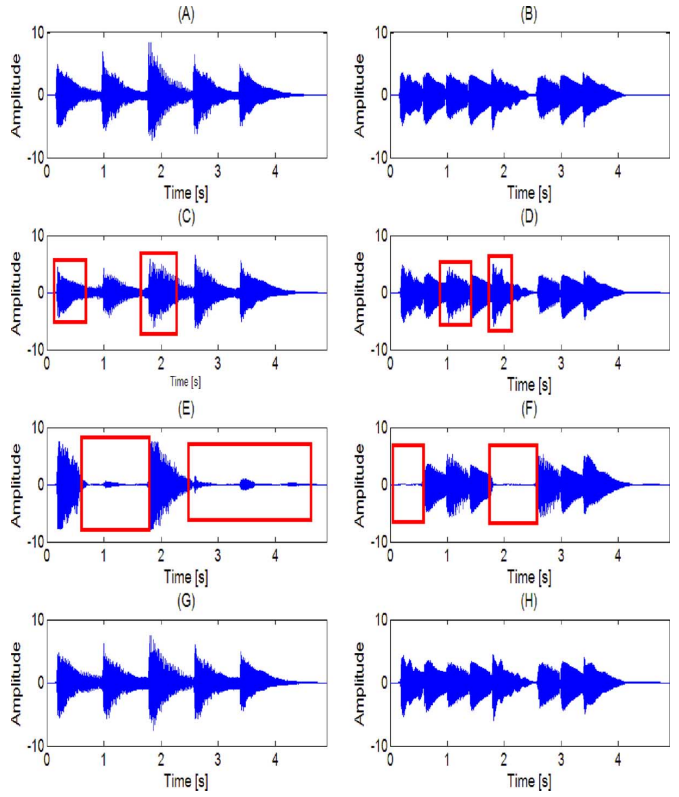

Fig. 7. Separated signals in time-domain. (A)–(B): piano and trumpet music for Case 1). (C)–(D): piano and trumpet music for Case 2). (E)–(F): piano and trumpet music for Case 3).

"over-sparse" factorization which eventually results in inferior

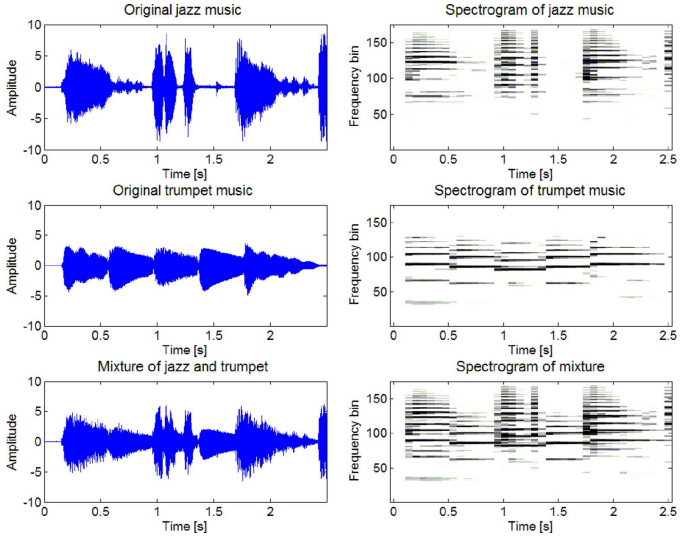

Fig. 8. Time-domain representation and spectrogram of the jazz music (top panels), trumpet music (middle panels), and mixed signal (bottom panels).

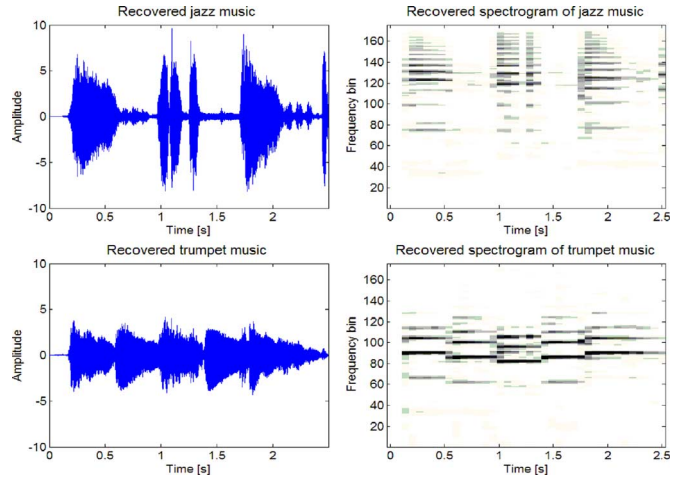

Fig. 9. Separated signals in time and TF domain.

separation performance. Figs. 8 and 9 shows another example of separating jazz and trumpet mixture based on the proposed method.

In Fig. 8, it is shown that both trumpet and jazz music overlap in time, and the jazz notes are cross interspersed in frequency with the trumpet notes (e.g., both notes mixed together between 0 to 0.5 s, 0.8 to 1.4 s, and 1.7 to 2.3 s). Fig. 9 shows the separated sources in terms of the spectrogram and time-domain representation, respectively.

In Fig. 9, it is clearly shown that all cross interspersed notes associated with each source have been successfully separated by using the proposed adaptive method. The overall comparison results between the adaptive and uniform sparsity methods have been summarized in Table I. According to the table, SNMF2D with adaptive sparsity tends to yield better result than the uniform sparsity-based methods. We may summarize the average performance improvement of our method against the uniform constant sparsity method: 1) For the piano and trumpet music, the improvement per source in terms of the SDR is 2 dB, SAR 1.8 dB, and SIR 2.2 dB. 2) For the piano and jazz music, the improvement per source in terms of SDR is 1.3 dB, SAR 1.1 dB,

TABLE I
PERFORMANCE COMPARISON BETWEEN ADAPTIVE AND UNIFORM SPARSITY METHODS

Mixture	Methods	SDR	SAR	SIR
Piano and trumpet music	Proposed adaptive sparsity	10.1	11.3	11.6
	(Best) Uniform sparsity	8.1	9.5	9.4
Piano and jazz music	Proposed adaptive sparsity	8.5	9.2	9.5
	(Best) Uniform sparsity	7.2	8.1	7.4
Trumpet and jazz music	Proposed adaptive sparsity	9.7	10.3	9.7
	(Best) Uniform sparsity	8.6	8.8	7.9

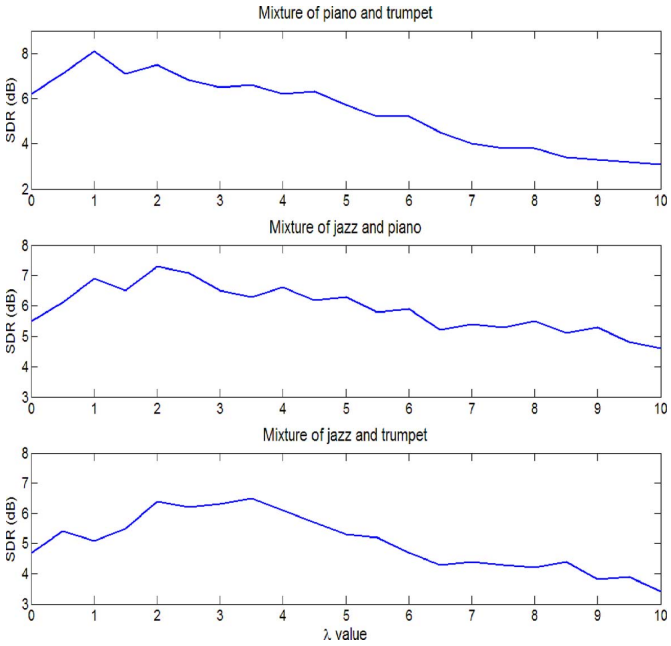

Fig. 10. Separation results of SNMF2D by using different uniform regularization.

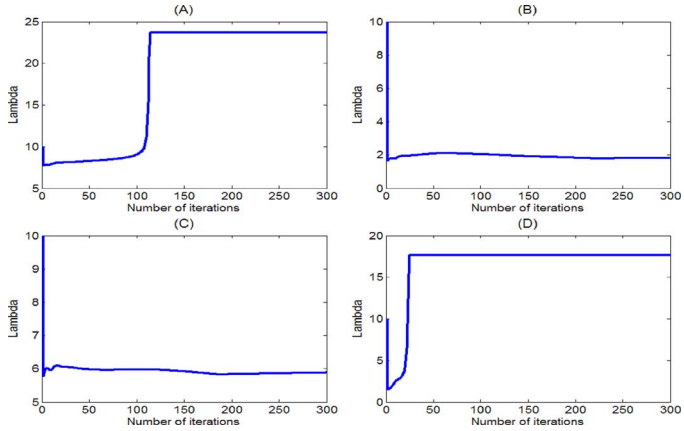

Fig. 11. Convergence trajectory of the sparsity: (A) $\lambda_{1,1}^{\phi=0}$, (B) $\lambda_{1,5}^{\phi=0}$, (C) $\lambda_{1,10}^{\phi=0}$, (D) $\lambda_{1,15}^{\phi=0}$.

and SIR 1.7 dB. 3) For the trumpet and jazz music, the improvement per source in terms of SDR is 1.1 dB, SAR 1.5 dB, and SIR 1.8 dB.

On a point of interest, the analyses for Cases 1) and 2) in Figs. 6 and 7 are based on the single fixed uniform sparsity parameter where $\lambda_{d,l}^{\phi} = \lambda$ is set to be either too high and too low, respectively. From these results, it could be argued that such settings of uniform sparsity parameter are unrealistic for source separation. To investigate this further, 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 plotted in Fig. 10. 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 Table I.

In Fig. 10, the results have clearly indicated that there are certain values of λ where the SNMF2D performs with exceptionally good results. In the case of piano and trumpet mixtures, the best performance is obtained when λ ranges from 0.5 to 2 where the highest SDR is 8.1 dB. As for jazz and piano mixtures, the best performance is obtained when λ ranges from 1.0 to 2.5 where the highest SDR is 7.2 dB and for jazz and trumpet mixtures, the best performance is obtained when λ ranges from 2 to 3.5 where the highest SDR is 8.6 dB. On the contrary, when λ is set too high, the separation performance tends to degrade. It is also worth pointing out that the separation results are coarse when the factorization is non-regularized Here, we see that 1) for piano and trumpet mixtures, the SDR is only 6.2 dB, 2) for jazz and piano mixtures, the SDR is only 5.6 dB, and 3) for jazz and trumpet mixtures, the SDR is only 4.7 dB. From above, it is evident that 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 sparseness necessary for matrix factorization to resolve the ambiguity between the sources in the TF domain.

3) Adaptive Behavior of Sparsity Parameter: In this subsection, 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. Fig. 11 shows the convergence trajectory of four adaptive sparsity 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 $\lambda_{d,l}^{\phi} = 10$ for all d, l, ϕ and are subsequently adapted according to (31). After 300 iterations, the above sparsity parameters converge to their steady-states. By examining Fig. 11, it is noted that the converged steady-state values are significantly different for each sparsity parameter e.g., $\lambda_{1,1}^{\phi=0} = 24.4$, $\lambda_{1,5}^{\phi=0} = 2.1$, $\lambda_{1,10}^{\phi=0} = 5.9$, and $\lambda_{1,15}^{\phi=0} = 17.4$ even though they started at the same initial condition. This shows that each element code has its own sparseness. In addition, it is worth pointing out that in the case of piano and trumpet mixture the SDR result rises to 10 dB when $\lambda_{d,l}^{\phi}$ is adaptive. This represents a 2 dB per source improvement over the case of uniform constant sparsity (which is only 8.1 dB in Table I). On the separate

TABLE II
OVERALL ESE PERFORMANCE

Mixture	Methods	PSR	SIR	ESE
Piano and trumpet music	IBM	0.98	231	0.98
	Uniform sparsity	0.76	103	0.75
	Adaptive sparsity	0.9	191	0.89
Piano and jazz music	IBM	0.98	214	0.98
	Uniform sparsity	0.68	84	0.67
	Adaptive sparsity	0.86	175	0.86
Trumpet and jazz music	IBM	0.97	228	0.97
	Uniform sparsity	0.73	98	0.72
	Adaptive sparsity	0.88	181	0.88

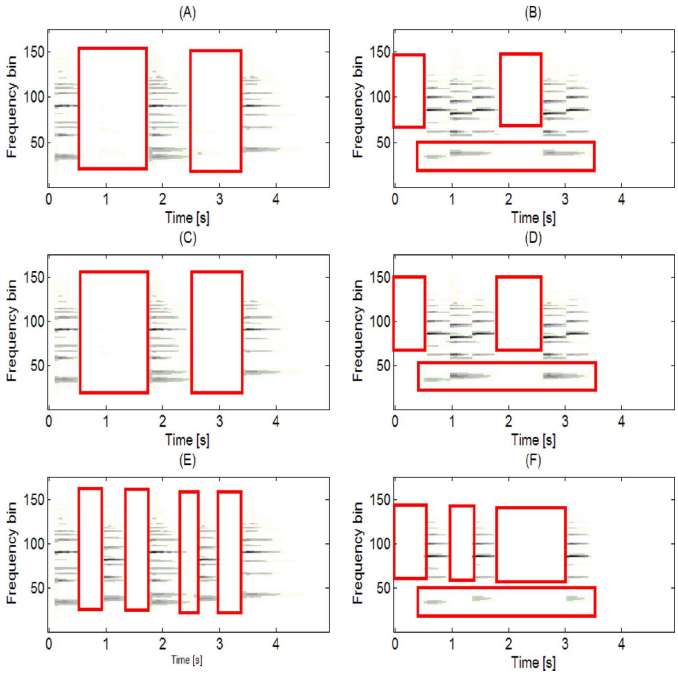

Fig. 12. Separated signals in spectrogram. (A)-(B): piano and trumpet music using SNMF. (C)-(D): piano and trumpet music using NMF-ARD. (E)-(F): piano and trumpet music using NMF-TCS.

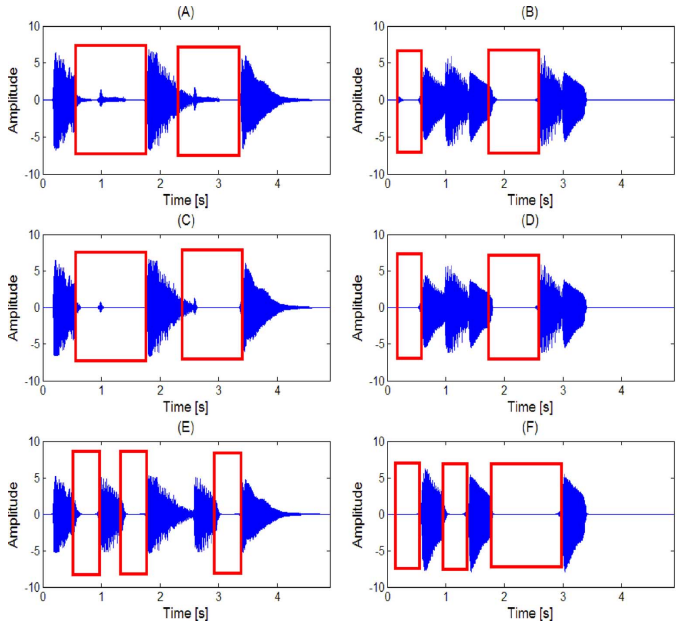

Fig. 13. Separated signals in time-domain. (A)-(B): piano and trumpet music using SNMF. (C)-(D): piano and trumpet music using NMF-ARD. (E)-(F): piano and trumpet music using NMF-TCS.

hand, when no sparsity is imposed onto the codes the SDR result immediately deteriorates to approximately 6 dB. This represents a 4 dB per source depreciation compared with the proposed adaptive sparsity method. From above, the results are ready to suggest that the performances of source separation have 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 source separation performance as indicated in Table I.

4) Efficiency of Source Extraction in TF Domain: In this subsection, we will analyze the efficiency of source extraction based on the three cases previously enunciated at the beginning of Section IV-B. Binary masks are constructed using the approach discussed in Section III-A for each of the three cases. To ensure fair comparison, we generate the ideal binary mask (IBM) [40] from the original source which is used as a reference for comparison. The IBM for a target source is found for each TF unit by comparing the energy of the target source to the energy of all the interfering sources. Hence, the ideal binary mask produces the optimal signal-to-distortion ratio (SDR) gain of all binary masks and thus, it can be considered as an optimal source extractor in TF domain. The comparison results between IBM, uniform sparsity and proposed adaptive sparsity are tabulated in Table II.

In Table II, the results of PSR, SIR, and ESE for each mixture type are obtained by averaging over 100 realizations. From listening performance test, any $\psi(W_d) > 0.8$ indicates acceptable quality of source extraction performance in TF domain. Therefore, it is noted from the results in Table II that both IBM and the proposed method satisfy this condition. In addition, the proposed method yields better ESE improvement against the uniform sparsity method. The average improvement results have been summarized as follows: 1) for the piano and trumpet music,

TABLE III
PERFORMANCE COMPARISON BETWEEN OTHER NMF BASED SCASS METHODS AND PROPOSED METHOD

Mixtures	Methods	SDR	SAR	SIR
Piano and trumpet music	Proposed method	10.1	11.3	11.6
	NMF-TCS	4.5	7.2	7.1
	SNMF	3.2	5.1	5.8
	NMF-ARD	3.5	5.7	6.4
Piano and jazz music	Proposed method	8.5	9.2	9.5
	NMF-TCS	4.2	6.6	5.5
	SNMF	2.9	4.7	4.5
	NMF-ARD	3.1	5.8	5.2
Trumpet and jazz music	Proposed method	9.7	10.3	9.7
	NMF-TCS	5.1	6.5	5.6
	SNMF	3.2	5.8	5.3
	NMF-ARD	3.7	6.1	5.5

TABLE IV
ESE COMPARISON BETWEEN OTHER NMF-BASED SCASS METHODS AND THE PROPOSED METHOD

Mixtures	Methods	PSR	SIR	ESE
Piano and trumpet music	Proposed method	0.9	191	0.89
	NMF-TCS	0.51	32	0.49
	SNMF	0.38	18	0.36
	NMF-ARD	0.42	20	0.40
Piano and jazz music	Proposed method	0.86	175	0.86
	NMF-TCS	0.45	27	0.43
	SNMF	0.31	19	0.29
	NMF-ARD	0.36	21	0.34
Trumpet and jazz music	Proposed method	0.88	181	0.88
	NMF-TCS	0.47	31	0.45
	SNMF	0.33	17	0.31
	NMF-ARD	0.35	20	0.33

18.4%; 2) for the piano and jazz music 26.5%; 3) for the trumpet and jazz music, 20.6%. In addition, the average SIR of the proposed method exhibits much a higher value than the uniform sparsity SNMF2D. This clearly shows that the amount of interference between any two sources is lesser for the proposed method. Therefore, the above results unanimously indicate that the proposed adaptive sparsity method leads to higher ESE results than the uniform constant sparsity method.

C. Impact of Adaptive and Fixed Sparsity

In Section IV-B, analysis has been carried out to investigate effects between adaptive sparsity and uniform constant sparsity on source separation. In this evaluation, we compare the proposed method with other sparse NMF-based source separation methods. These consist of the following:

- SNMF (a multiplicative update algorithm by Lee and Seung [5]). The uniform constant sparsity parameter is progressively varied from 0 to 10 with every increment of 0.1 (i.e., $\lambda = 0, 0.1, 0.2, \ldots, 10$) and the best result is retained for comparison.
- Automatic relevance determination NMF (NMF-ARD) [41] exploits a hierarchical Bayesian framework SNMF that amounts to imposing an exponential prior for pruning and thereby enables estimation of the NMF model order. The NMF-ARD assumes prior on $\mathbf{H}$, namely, $p(\mathbf{H}|\lambda) = \prod_d \lambda_d^{l_{\max}} \exp -(\lambda_d \sum_l \mathbf{H}_{d,l})$ and uses automatic relevance determination (ARD) approach to determine the desirable number of components in $\mathbf{D}$. The initialization number of components in $\mathbf{D}$ is 10.
- NMF with temporal continuity and sparseness criteria [42] (NMF-TCS) is based on factorizing the magnitude spectrogram of the mixed signal into a sum of components, which include the temporal continuity and sparseness criteria into the separation framework. In [42], the temporal continuity α is chosen as $[0, 1, 10, 100, 1000]$, sparseness weight β is chosen as $[0, 1, 10, 100, 1000]$. The best separation result is retained for comparison.

In Figs. 12 and 13, panels (A)-(F) show that the above methods did not fully separate the music mixture. Many spectral and temporal components are missing from the recovered sources and these have been highlighted (marked red box) in all panels. The above methods fail to take into account the relative position of each spectrum and thereby discarding the temporal information. Better separation results will require a proper model that can represent both temporal structure and the pitch change which occurs when an instrument plays different notes simultaneously. If the temporal structure and the pitch change are not considered in the model, the mixing ambiguity is still contained in each separated source. Table III further gives the SDR, SAR, and SIR comparison results between our proposed method and the above three sparse NMF methods.

The improvement of our method compared with NMF-TCS, SNMF and NMF-ARD can be summarized as follows: 1) for the piano and trumpet music, the average improvement per source in terms of the SDR is 6.3 dB, SAR 4.8 dB, and SIR 5.1 dB;

2) for the piano and jazz music, the average improvement per source in terms of SDR is 5 dB, SAR 3.9 dB, and SIR 4.7 dB; 3) for the trumpet and jazz music, the average improvement per source in terms of SDR is 5.4 dB, SAR 4.2 dB, and SIR 4.3 dB. In the case of ESE (see Table IV), the proposed method exhibits much better average ESE of approximately 106.9%, 138.8% and 114.6% improvement with NMF-TCS, SNMF and NMF-ARD, respectively. Analyzing the separation results and ESE performance, the proposed method leads to the best separation performance for both recovered sources. The SNMF method performs with poorer results whereas the separation performance by the NMF-TCS method is slightly better than the NMF-ARD and SNMF methods. Our proposed method gives significantly better performance than the NMF-TCS, SNMF and NMF-ARD methods. The spectral dictionary obtained via NMF-TCS, SNMF and NMF-ARD methods are not adequate to capture the temporal dependency of the frequency patterns within the audio signal. In addition, the NMF-TCS, SNMF and NMF-ARD do not model notes but rather unique events only. Thus, if two notes are always played simultaneously they will be modeled as one component. Also, some components might not correspond to notes but rather to the model, e.g., background noise.

V. Conclusion

The paper presents a new adaptive sparsity non-negative matrix factorization. The impetus behind this work is that the sparsity achieved by SNMF and SNMF2D is not enough; in such situations it might be useful to control the degree of sparseness explicitly. In the proposed method, the regularization term is adaptively tuned using a variational Bayesian approach to yield desired sparse decomposition, thus enabling the spectral dictionary and temporal codes of nonstationary audio signals to be estimated more efficiently. This has been verified concretely based on our simulation results. In addition, the proposed method has yielded significant improvements in single channel music separation when compared with other sparse NMF-based source separation methods.

References

[1] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, 1994.

[2] S. C. Alvarez, A. Cichocki, and L. C. Ribas, "An Iterative inversion approach to blind source separation," *IEEE Trans. Neural Netw.*, vol. 11, no. 6, pp. 1423–1437, Nov. 2000.

[3] P. Gao, W. L. Woo, and S. S. Dlay, "Nonlinear signal separation for multi-nonlinearity constrained mixing model," *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp. 796–802, May 2006.

[4] J. Zhang, W. L. Woo, and S. S. Dlay, "Blind source separation of post-nonlinear convolutive mixture," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2311–2330, Nov. 2007.

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

[6] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[7] D. Lee and H. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[8] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Proc. NIPS*, 2003.

[9] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 538–5493, Mar. 2010.

[10] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[11] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2003, pp. 177–180.

[12] Y. C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classification," *Pattern Recognit. Lett.*, vol. 26, pp. 1327–1336, 2005.

[13] D. Guillamet and J. Vitri'a, "Introducing a weighted nonnegative matrix factorization for image classification," *Pattern Recognit. Lett.*, vol. 24, pp. 2447–2454, 2004.

[14] R. Zdunek and A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization," *Signal Process.*, vol. 87, no. 8, pp. 1904–1916, Aug. 2007.

[15] P. Sajda, S. Du, T. Brown, R. Stoyanova, D. Shungu, X. Mao, and L. Parra, "Non-negative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain," *IEEE Trans. Med. Imag.*, vol. 23, no. 12, pp. 1453–1465, 2004.

[16] F. J. Theis and G. A. García, "On the use of sparse signal decomposition in the analysis of multi-channel surface electromyograms," *Signal Process.*, vol. 86, no. 3, pp. 603–623, Mar. 2006.

[17] O. Okun and H. Priisalu, "Unsupervised data reduction," *Signal Process.*, vol. 87, no. 9, pp. 2260–2267, Sep. 2007.

[18] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 3, pp. 780–791, 2007.

[19] A. Cichocki, R. Zdunek, and S. I. Amari, "Csiszár's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Signal Separat. (ICABSS'06)*, Charleston, SC, Mar. 2006, vol. 3889, pp. 32–39.

[20] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[21] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[22] E. Vincent, "Musical source separation using time–frequency source priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 91–98, Jan. 2006.

[23] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, 2009, pp. 3137–3140.

[24] G. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Proc. 9th Int. Conf. Latent Variable Anal. Signal Separat. (LCA/ICA)*, 2010.

[25] M. Nakano *et al.*, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying in music spectrograms," in *Proc. 9th Int. Conf. Latent Variable Anal. Signal Separat. (LCA/ICA)*, 2010.

[26] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorization models," *Comput. Intell. Neurosci.*, no. Doi: 10.1155/2009/785152, 2009.

[27] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4133–4145, Nov. 2006.

[28] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorisation," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separat.*, 2009.

[29] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 880–887.

[30] M. Morup and M. N. Schmidt, *Sparse Non-Negative Matrix Factor 2-D Deconvolution*. Copenhagen, Denmark: Technical Univ. of Denmark, 2006.

[31] M. N. Schmidt and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proc.Int. Conf. Ind. Compon. Anal. Blind Signal Separat. (ICABSS'06)*, Charleston, SC, Mar. 2006, vol. 3889, pp. 700–707.

[32] 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.*, vol. 19, no. 4, pp. 961–976, May 2011.

[33] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, 1991.
[34] L. Yuanqing, "l_1-Norm sparse Bayesian learning: theory and applications," Ph.D. dissertation, Univ. of Pennsylvania, Philadelphia, 2008.
[35] F. Sha, L. K. Saul, and D. D. Lee, "Multiplicative updates for nonnegative quadratic programming in support vector machines," in *Proc. Adv. Neural Information Process. Syst.*, 2002, vol. 15, pp. 1041–1048.
[36] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.
[37] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. Int. Symp. Music Inf. Retrieval (ISMIR)*, Baltimore, MD, Oct. 2003, pp. 229–230.
[38] *Signal Separation Evaluation Campaign (SiSEC 2008),* 2008. [Online]. Available: http://sisec.wiki.irisa.fr
[39] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2005.
[40] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.
[41] M. Mørup and K. L. Hansen, "Tuning pruning in sparse non-negative matrix factorization," in *Proc. 17th Eur. Signal Process. Conf. (EUSIPCO'09)*, Glasgow, U.K., 2009.
[42] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

Bin Gao received the B.S. degree in communications and signal processing from Southwest Jiao Tong University, Chengdu, China in 2005, the M.Sc. degree (with distinction) in communications and signal processing from Newcastle University, Newcastle upon Tyne, U.K., in 2007, and the Ph.D. degree from Newcastle University in 2007 and his research topic was single-channel blind source separation under the supervision of Dr. Woo and Prof. Dlay.

Currently, He is a Research Associate at Newcastle University. His research interests include audio and image processing, machine learning, structured probabilistic modeling on audio applications such as audio source separation, feature extraction, and denoising.

W. L. Woo (M'11) was born in Malaysia. He received the B.Eng. degree (first class honors) in electrical and electronics engineering and the Ph.D. degree from the Newcastle University, Newcastle upon Tyne, U.K.

He is currently a Senior Lecturer with the School of Electrical, Electronics, and Computer Engineering, Newcastle University. His major research is in the mathematical theory and algorithms for nonlinear signal and image processing. This includes areas of blind source separation, machine learning, multidimensional signal processing, signal/image deconvolution, and restoration. He has an extensive portfolio of relevant research supported by a variety of funding agencies. Prior to joining the school, he worked on source separation techniques supported by QinetiQ on signal processing-based applications. He has published over 250 papers on these topics on various journals and international conference proceedings. Currently, he serves on the editorial board of the many international signal processing journals.

Dr. Woo was awarded the IEE Prize and the British Scholarship in 1998 to continue his research work. He actively participate in international conferences and workshops, and serves on their organizing and technical committees. In addition, he acts as a consultant to a number of industrial companies that involve the use of statistical signal and image processing techniques. He is also a member of the Institution Engineering Technology (IET)

S. S Dlay received the B.Sc. (honors) degree in electrical and electronic engineering and the Ph.D. in VLSI design from Newcastle University, Newcastle upon Tyne, U.K., in 1979 and 1983, respectively.

In 1984, he was appointed as a Post-Doctoral Research Associate at Newcastle University and helped to establish an Integrated Circuit Design Centre, funded by the EPSRC. In November 1984, he was appointed as a Lecturer in the Department of Electronic Systems Engineering at the University of Essex. In 1986 he rejoined Newcastle University as a Lecturer in the School of Electrical, Electronic, and Computer Engineering, then in 2001 he was promoted to Senior Lecturer. In recognition of his major achievements he has been appointed to a Personal Chair in Signal Processing Analysis. He is currently Head of the Signal Processing theme. He has published over 250 research papers and his research interests lie in the mathematical advancement and application of modern signal processing theory to biometrics and security, biomedical signal processing and implementation of signal processing architectures.

Prof. Dlay held a Scholarship from the Engineering and Physical Science Research Council (EPSRC) and the Charles Hertzmann Award. He serves on many editorial boards and has played an active role in numerous international conferences in terms of serving on technical and advisory committees as well as organizing special sessions. He is a College Member of the EPSRC.