Machine Learning Source Separation using Maximum A Posteriori Nonnegative Matrix Factorization

Bin Gao, *Member, IEEE,* W.L. Woo, *Senior Member, IEEE,* and Bingo W-K. Ling, *Senior Member, IEEE*

***Abstract* — A novel unsupervised machine learning algorithm for single channel source separation (SCSS) is presented. The proposed method is based on nonnegative matrix factorization which is optimized under the framework of maximum a posteriori (MAP) probability and Itakura-Saito (IS) divergence. The method enables a generalized criterion for variable sparseness to be imposed onto the solution and prior information to be explicitly incorporated through the basis vectors. In addition, the method is scale invariant where both low and high energy components of a signal are treated with equal importance. The proposed algorithm is a more complete and efficient approach for matrix factorization of signals that exhibit temporal dependency of the frequency patterns. Experimental tests have been conducted and compared with other algorithms to verify the efficiency of the proposed method.**

I. Introduction

NONNEGATIVE Matrix Factorization (NMF) is an emerging machine learning technique [1-5] for data mining, dimensionality reduction, pattern recognition, object detection, classification, and blind source separation (BSS) [6-9]. In recent times, single channel source separation (SCSS) is becoming more important especially using matrix factorization methods [10–28]. The SCSS problem can be treated with one observation and several unknown sources, namely:

$$y(t)=\sum_{i=1}^{I}x_i(t) \tag{1}$$

where $i=1,\ldots,I$ denotes the number of sources and $t=1,2,\ldots,T$ denotes time index and the goal is to estimate the sources $x_i(t)$ when only the observation signal $y(t)$ is available. NMF-based methods exploit an appropriate time-frequency (TF) analysis on the mono input recordings, yielding a TF representation. The decomposition is usually sought after through the minimization problem

$$\min_{\mathbf{D},\mathbf{H}} D\left(|\mathbf{Y}|^{.2}\middle|\mathbf{D},\mathbf{H}\right) \quad \text{subject to } \mathbf{D}\geq 0,\mathbf{H}\geq 0 \tag{2}$$

where $|\mathbf{Y}|^{.2}\in\Re_{+}^{F\times T_s}$ is the power TF representation of mixture $y(t)$ while $\mathbf{D}\in\Re_{+}^{F\times I}$ and $\mathbf{H}\in\Re_{+}^{I\times T_s}$ are two nonnegative matrices. F and T_s represent total frequency units and time slots, respectively in the TF domain. The matrix $\mathbf{D}$ can be compressed and reduced to its integral components such that it contains a set of spectral basis vectors, and $\mathbf{H}$ is a code matrix which describes the amplitude of each basis vector at each time point. The distance function $D\left(|\mathbf{Y}|^{.2}\middle|\mathbf{D},\mathbf{H}\right)$ is separable measure of fit. Commonly used cost functions for NMF are the generalized Kullback-Leibler (KL) divergence and Least Square (LS) distance [12]. NMF decomposition is not unique [14] and to overcome this limitation, a sparseness constraint [15, 16] can be added to the cost function. This can be achieved by regularization using the L_1-norm.

Over the few years, several types of prior over $\mathbf{D}$ and $\mathbf{H}$ have been proposed and maximum a-posteriori (MAP) criterion is used to optimise the spectral basis, code and prior parameters. These methods include the followings: NMF with Temporal Continuity and Sparseness Criteria [15] (NMF-TCS) 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. Automatic Relevance Determination NMF (NMF-ARD) [27, 28] exploits a hierarchical Bayesian framework sparse NMF that amounts to imposing an exponential prior for pruning and thereby enables estimation of the NMF model order. Bayesian NMF methods using Gamma distribution prior have also been proposed in [25]. Regardless of the cost function and different prior constraint being used, the standard NMF or MAP NMF models [27, 28, 31] are only satisfactory for solving source separation provided that the spectral frequencies of the audio signal do not change over time. However, this is not the case for many realistic audio signals. As a result, the spectral basis obtained via the NMF or MAP NMF decomposition is not adequate to capture the temporal dependency of the frequency patterns within the signal. In addition, most methods developed so far work only for music separation and have some important limitations that explicitly employ some prior knowledge about the sources. As a consequence, those methods are able to deal only with a very specific set of signals and situations.

In recent years, research has been undertaken to extend the sparse NMF to a two-dimensional convolution of $\mathbf{D}$ and $\mathbf{H}$ which culminated to the SNMF2D [16]. This allows the SNMF2D to capture both the temporal structure and the pitch change of a 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. SNMF2D imposes uniform sparsity on all temporal codes and 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 inevitably constrains all codes to be stationary. However, audio signals are

Bin Gao is with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China. W. L. Woo is with the School of Electrical and Electronic Engineering, Newcastle University, England, United Kingdom. Bingo W-K Ling is with the Faculty of Engineering, Guangdong University of Technology, China.
Correspondence author email address: w.l.woo@ncl.ac.uk

non-stationary even in the TF domain and have different temporal structure and sparsity. Hence, they cannot be realistically enforced by a fixed distribution. These characteristics are even more pronounced between different types of audio signals. Moreover, since the SNMF2D introduces many temporal shifts this will result in more temporal codes to deviate from the fixed distribution. Therefore, when SNMF2D imposes uniform sparsity on all the temporal codes, this will unavoidably result in 'under- or over-sparse' factorization which will subsequently lead to ambiguity in separating audio mixtures. Therefore, the above suggests that the current form of SNMF2D is still technically lacking and is not readily suited for SCSS especially mixtures involving different types of signals.

In this paper, a new matrix factorization algorithm is proposed for SCSS. Firstly, the proposed cost function is specially developed for factorization of non-stationary signals that exhibit temporal dependency of the frequency patterns. The proposed algorithm will overcome all the limitations associated with the SNMF2D as previously discussed above. The proposed model allows 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 proposed 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. In addition, 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. Secondly, the proposed factorization is based on IS divergence and has the property of scale invariant where lower energy components in the TF domain can be treated with equal importance as higher energy components. This is particularly relevant to audio sources since they are frequently characterized by large dynamic ranges. Finally, as each audio signal has its own temporal dependency of the frequency patterns, designing the appropriate spectral basis to match these features is imperative. If spectral bases share some degree of correlation, then this information should be captured to enable better matrix factorization Towards this end, we have developed a modified Gaussian prior on $\mathbf{D}$ to allow the proposed matrix factorization to capture the spectral basis efficiently.

The paper is organized as follows: In Section II, the new algorithm is derived and the proposed source separation framework is developed. Experimental results and comparison with other matrix factorization methods are presented in Section III. Finally, Section IV concludes the paper.

II. MAP REGULARIZED NMF2D WITH ITAKURA-SAITO DIVERGENCE

A. Itakura-Saito divergence

The IS divergence is a measure of the perceptual difference between an original spectrum $P(\omega)$ and an approximation $\hat{P}(\omega)$ of that spectrum [19]. Recently IS divergence has picked up renewed interest in NMF. The IS divergence leads to desirable statistical interpretations of the NMF problem [13]. Most significantly, NMF with IS divergence can provide scale invariant property which enables low energy components of $|\mathbf{Y}|^{.2}$ to bear the same relative importance as high energy ones. This is relevant to situations where the coefficients of $|\mathbf{Y}|^{.2}$ have a large dynamic range such as in audio short-term spectra. This property, in particular, can effectively separate the audio mixture when given only one channel recording. The IS divergence is formally defined as follows:

$$d_{IS}(a\,|\,b)=\frac{a}{b}-\log\frac{a}{b}-1 \tag{3}$$

The IS divergence is a limiting case of the β-divergence which is defined as

$$d_{\beta}(a\,|\,b)=\begin{cases}\dfrac{1}{\beta(\beta-1)}\left(a^{\beta}+(\beta-1)b^{\beta}-\beta a\,b^{\beta-1}\right) & \beta\in\Re\setminus\{0,1\}\\ a\left(\log a-\log b\right)+\left(b-a\right) & \beta=1\\ \dfrac{a}{b}-\log\dfrac{a}{b}-1 & \beta=0\end{cases} \tag{4}$$

It is interesting to note that for $\beta=2$ we obtain the Euclidean distance expressed by Frobenius norm and for $\beta=1$ we obtain the generalized Kullback-Leibler divergence as defined in (4). For $\beta=0$, this results in the IS divergence which is the unique to the β-divergence as it holds the property of scale invariance, namely:

$$\begin{cases}d_{\beta}(\gamma a\,|\,\gamma b)=\gamma^{\beta}d_{\beta}(a\,|\,b)\\ d_{IS}(\gamma a\,|\,\gamma b)=d_{IS}(a\,|\,b)\end{cases} \tag{5}$$

This shows that a good fit of the factorization for a lower energy a will cost as much as higher energy component b. On the other hand, factorizations by exploiting LS or KL divergence are highly dependent on the large amplitude coefficients but ignore the less precision in the estimation of the low-power components.

B. Proposed Variable Regularized IS-vRNMF2D

In this section, we derive a new *variable regularized* nonnegative matrix factorization algorithm. Considering the following generative model for observation $\mathbf{Y}\in\Re_{+}^{F\times T_s}$:

$$|\mathbf{Y}|^{.2}\approx\sum_{i}^{I}\sum_{\tau}^{\tau_{\max}}\sum_{\phi}^{\phi_{\max}}\overset{\downarrow\phi}{\mathbf{D}}_{i}^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}_{i}^{\phi} \tag{6}$$

In (3), "$\bullet$" is element-wise product, The vertical arrow in $\overset{\downarrow\phi}{\mathbf{D}}^{\tau}$ denotes downward shift which moves each element in the matrix $\mathbf{D}^{\tau}$ down by ϕ rows, and the horizontal arrow in $\overset{\rightarrow\tau}{\mathbf{H}}^{\phi}$ denotes right shift which moves each element in the matrix $\mathbf{H}^{\phi}$ to the right by τ columns. $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ are the i^{th} column of $\mathbf{D}^{\tau}$ and i^{th} row of $\mathbf{H}^{\phi}$, respectively. The matrix $\mathbf{D}^{\tau}=\left\{\mathbf{D}_{f,i}^{\tau}\,|\,f=1,...,F \text{ and } i=1,...,I\right\}$ denotes the τ^{th} slice of basis $\mathbf{D}$ and $\mathbf{H}^{\phi}=\left\{\mathbf{H}_{i,t_s}^{\phi}\,|\,i=1,...,I \text{ and } t_s=1,...,T_s\right\}$ denotes the ϕ^{th} slice of code $\mathbf{H}$. In source separation, $\mathbf{D}_i^{\tau}$ represents the spectral

basis of the i^{th} source while $\mathbf{H}_i^{\phi}$ represents the temporal code for each spectral basis element. To facilitate the decomposition in (6), we define $\mathbf{D}=\left[\mathbf{D}^1\mathbf{D}^2\cdots\mathbf{D}^{\tau_{\max}}\right]$, $\mathbf{\Lambda}=\left[\mathbf{\Lambda}^1\mathbf{\Lambda}^2\cdots\mathbf{\Lambda}^{\phi_{\max}}\right]$ and $\mathbf{H}=\left[\mathbf{H}^1\mathbf{H}^2\cdots\mathbf{H}^{\phi_{\max}}\right]$, and then choose a prior distribution $p(\mathbf{D},\mathbf{H})$ over the factors $\{\mathbf{D},\mathbf{H}\}$. $\mathbf{\Lambda}^{\phi}=\left\{\Lambda_{i,t_s}^{\phi}\mid i=1,...,I \text{ and } t_s=1,...,T_s\right\}$ denotes the ϕ^{th} slice of sparse parameter $\mathbf{\Lambda}$. The terms $\tau_{\max}$, $\phi_{\max}$ are the maximum number of τ shifts, ϕ shifts, respectively. The posterior can be found by using Bayes' theorem as:

$$p\left(\mathbf{D},\mathbf{H}|\mathbf{Y},\mathbf{\Lambda}\right)=\frac{p\left(\mathbf{Y}\mid\mathbf{D},\mathbf{H}\right)p\left(\mathbf{D}\right)p\left(\mathbf{H}\mid\mathbf{\Lambda}\right)}{P\left(\mathbf{Y}\right)} \tag{7}$$

where the denominator is a constant and it is assumed $\mathbf{D}$ and $\mathbf{H}$ are jointly independent so that the log-posterior is given by:

$$\log p\left(\mathbf{D},\mathbf{H}|\mathbf{Y},\mathbf{\Lambda}\right)=\log p\left(\mathbf{Y}\mid\mathbf{D},\mathbf{H}\right)+\log p\left(\mathbf{D}\right)+\log p\left(\mathbf{H}\mid\mathbf{\Lambda}\right)+\text{const} \tag{8}$$

where ' const ' denotes constant. Under the i.i.d. noise assumption, the negative log likelihood $-\log p\left(\mathbf{Y}\mid\mathbf{D},\mathbf{H}\right)$ is given by:

$$\begin{aligned}
-\log p\left(\mathbf{Y}\mid\mathbf{D},\mathbf{H}\right)&=-\sum_{t_s=1}^{Ts}\sum_{f=1}^{F}\log\xi\left(\left.\frac{|\mathbf{Y}|_{f,t_s}^2}{\sum_i\sum_\tau\sum_\phi\overset{\downarrow\phi}{\mathbf{D}}{}_{f,i}^{\tau}\overrightarrow{\mathbf{H}}{}_{i,t_s}^{\phi}}\right|\varepsilon/\upsilon\right)\Bigg/\sum_i\sum_\tau\sum_\phi\overset{\downarrow\phi}{\mathbf{D}}{}_{f,i}^{\tau}\overrightarrow{\mathbf{H}}{}_{i,t_s}^{\phi}\\
&=-\sum_{t_s=1}^{Ts}\sum_{f=1}^{F}\log\left(\frac{\upsilon^{\varepsilon}}{\Gamma(\varepsilon)}\left(\frac{|\mathbf{Y}|_{f,t_s}^2}{\sum_i\sum_\tau\sum_\phi\overset{\downarrow\phi}{\mathbf{D}}{}_{f,i}^{\tau}\overrightarrow{\mathbf{H}}{}_{i,t_s}^{\phi}}\right)^{\varepsilon-1}\exp\left(-\upsilon\frac{|\mathbf{Y}|_{f,t_s}^2}{\sum_i\sum_\tau\sum_\phi\overset{\downarrow\phi}{\mathbf{D}}{}_{f,i}^{\tau}\overrightarrow{\mathbf{H}}{}_{i,t_s}^{\phi}}\right)\right)\Bigg/\sum_i\sum_\tau\sum_\phi\overset{\downarrow\phi}{\mathbf{D}}{}_{f,i}^{\tau}\overrightarrow{\mathbf{H}}{}_{i,t_s}^{\phi}\\
&\doteq\upsilon\sum_{t_s=1}^{Ts}\sum_{f=1}^{F}\frac{|\mathbf{Y}|_{f,t_s}^2}{\sum_i\sum_\tau\sum_\phi\overset{\downarrow\phi}{\mathbf{D}}{}_{f,i}^{\tau}\overrightarrow{\mathbf{H}}{}_{i,t_s}^{\phi}}-\frac{\varepsilon}{\upsilon}\log\frac{|\mathbf{Y}|_{f,t_s}^2}{\sum_i\sum_\tau\sum_\phi\overset{\downarrow\phi}{\mathbf{D}}{}_{f,i}^{\tau}\overrightarrow{\mathbf{H}}{}_{i,t_s}^{\phi}}-1\\
&=d_{IS}\left(|\mathbf{Y}|^2\left|\sum_i\sum_\tau\sum_\phi\overset{\downarrow\phi}{\mathbf{D}}{}_{f,i}^{\tau}\overrightarrow{\mathbf{H}}{}_{i,t_s}^{\phi}\right.\right)
\end{aligned} \tag{9}$$

where ' $\doteq$ ' denotes equality up to a positive scale and a constant, and $d_{IS}(\cdot)$ is the IS divergence. The IS divergence [13] is formally defined as $d_{IS}(a\mid b)=\frac{a}{b}-\log\frac{a}{b}-1$. The ratio ε/υ is simply the mean of the Gamma distribution which by definition is equal to unity. Thus, the last line of (10) is obtained by setting $\varepsilon/\upsilon=1$. The IS divergence has the property of scale invariant where any low energy component in $|\mathbf{Y}|^2$ in (9) will bear the same relative importance as the high energy ones. This is very relevant to situations in which the coefficients of $|\mathbf{Y}|^2$ have large dynamic range such as in audio short-term spectra.

In our proposed model, the prior over $\mathbf{D}$ is a factorial model $p(\mathbf{D})=\prod_{\tau=0}^{\tau_{\max}}p(\mathbf{D}^{\tau})$ where the τ^{th} slice of $\mathbf{D}$ is assumed to be distributed as multivariate rectified Gaussian with covariance matrix $\mathbf{\Sigma}_{\tau}$. Considering the zero mean of the rectified Gaussian distribution [29] i.e. set $\bar{\mathbf{u}}^{\tau}=0$, as approximated as

$$p(\mathbf{D}^{\tau})\propto\begin{cases}\exp\left(-\frac{1}{2}\mathbf{d}^{\tau\mathbf{T}}\mathbf{\Sigma}_{\tau}^{-1}\mathbf{d}^{\tau}\right), & \mathbf{d}^{\tau}\geq 0\\ 0, & \mathbf{d}^{\tau}<0\end{cases} \tag{10}$$

where $\mathbf{\Sigma}_{\tau}=\begin{bmatrix}\mathbf{\Sigma}_{1,1,\tau} & \cdots & \mathbf{\Sigma}_{1,I,\tau}\\ \vdots & \ddots & \vdots\\ \mathbf{\Sigma}_{I,1,\tau} & \cdots & \mathbf{\Sigma}_{I,I,\tau}\end{bmatrix}$ is the covariance matrix of $vec\left(\mathbf{D}^{\tau}\right)$. In above, ' $\mathbf{T}$ ' denotes matrix transpose, $vec(\cdot)$ represents the column vectorisation and $\mathbf{\Sigma}_{i,j,\tau}=E\left[\mathbf{D}_i^{\tau}\mathbf{D}_j^{\tau\mathbf{T}}\right]$ is the cross-correlation between the basis vectors $\mathbf{D}_i^{\tau}$ and $\mathbf{D}_j^{\tau}$, ' $E[\bullet]$ ' denotes the expectation. The covariance matrix $\mathbf{\Sigma}_{\tau}$ can be partitioned as

$$\mathbf{\Sigma}_{\tau}=\underbrace{\begin{bmatrix}\mathbf{\Sigma}_{1,1,\tau} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}_{2,2,\tau} & \mathbf{0} & \cdots & \mathbf{0}\\ \vdots & \mathbf{0} & \ddots & \ddots & \vdots\\ \mathbf{0} & \vdots & \ddots & \ddots & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Sigma}_{I,I,\tau}\end{bmatrix}}_{\mathbf{\Sigma}_{diag,\tau}}+\underbrace{\begin{bmatrix}\mathbf{0} & \mathbf{\Sigma}_{1,2,\tau} & \cdots & \cdots & \mathbf{\Sigma}_{1,I,\tau}\\ \mathbf{\Sigma}_{2,1,\tau} & \mathbf{0} & \mathbf{\Sigma}_{2,3,\tau} & \cdots & \vdots\\ \vdots & \mathbf{\Sigma}_{3,2,\tau} & \ddots & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & \mathbf{\Sigma}_{I-1,I,\tau}\\ \mathbf{\Sigma}_{I,1,\tau} & \cdots & \cdots & \mathbf{\Sigma}_{I,I-1,\tau} & \mathbf{0}\end{bmatrix}}_{\mathbf{\Sigma}_{off,\tau}} \tag{11}$$

In (11), $\mathbf{0}$ is a $F\times F$ matrix with zero elements and $\mathbf{\Sigma}_{diag,\tau}^{-1}$ is the inverse covariance matrix of $\mathbf{\Sigma}_{diag,\tau}$. In deriving $\mathbf{\Sigma}_{\tau}^{-1}$, we have invoked the Woodbury matrix identity which states that the inverse of a rank-*k* correction of some matrix can be computed by doing a rank-*k* correction to the inverse of the original matrix. and thus, the inverse covariance matrix becomes

$$\begin{aligned}
\mathbf{\Sigma}_{\tau}^{-1}&=\left(\mathbf{\Sigma}_{diag,\tau}+\mathbf{\Sigma}_{off,\tau}\right)^{-1}\\
&\approx\mathbf{\Sigma}_{diag,\tau}^{-1}-\mathbf{\Sigma}_{diag,\tau}^{-1}\mathbf{\Sigma}_{off,\tau}\mathbf{\Sigma}_{diag,\tau}^{-1}\\
&=\Omega_{diag,\tau}-\Omega_{off,\tau}
\end{aligned} \tag{12}$$

where $\Omega_{diag,\tau}=\mathbf{\Sigma}_{diag,\tau}^{-1}$, $\Omega_{off,\tau}=\mathbf{\Sigma}_{diag,\tau}^{-1}\mathbf{\Sigma}_{off,\tau}\mathbf{\Sigma}_{diag,\tau}^{-1}$. The $(i,j)^{\text{th}}$ sub-matrix of $\Omega_{off,\tau}$ is given by

$$\Omega_{off,i,j,\tau}=\Sigma_{i,i,\tau}^{-1}\Sigma_{i,j,\tau}\Sigma_{j,j,\tau}^{-1} \tag{13}$$

Assuming that the elements within the same basis vector are uncorrelated, the above matrices simplify to $\Sigma_{i,i,\tau}=\sigma_{i,\tau}^2\mathbf{I}$, $\Sigma_{j,j,\tau}=\sigma_{j,\tau}^2\mathbf{I}$ and $\Sigma_{i,j,\tau}=c_{i,j,\tau}\mathbf{I}$ where $\sigma_{i,\tau}^2$ is the variance of the basis vector $\mathbf{D}_i^{\tau}$ and $c_{i,j,\tau}$ is the cross-covariance between $\mathbf{D}_i^{\tau}$ and $\mathbf{D}_j^{\tau}$. Thus, $\Omega_{off,i,j,\tau}$ can be expressed as

$$\Omega_{off,i,j,\tau}=\mu_{ij\tau}\mathbf{I} \tag{14}$$

where

$$\mu_{ij\tau}=\sigma_{i,\tau}^{-2}\,\sigma_{j,\tau}^{-2}\,c_{i,j,\tau} \tag{15}$$

Using above,

$$\begin{aligned}
-\log p(\mathbf{D}^{\tau})&\approx\frac{1}{2}vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}}\Omega_{diag,\tau}vec\left(\mathbf{D}^{\tau}\right)+\frac{1}{2}vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}}\Omega_{off,\tau}vec\left(\mathbf{D}^{\tau}\right)\\
&=\gamma+\frac{1}{2}vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}}\Omega_{off,\tau}vec\left(\mathbf{D}^{\tau}\right)
\end{aligned}$$

(16)

The first term $\gamma = \frac{1}{2} vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}} \Omega_{diag,\tau} vec\left(\mathbf{D}^{\tau}\right)$ relates only to the power of $\mathbf{D}^{\tau}$ while the second term $vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}} \Omega_{off,\tau} vec\left(\mathbf{D}^{\tau}\right) = \sum_{i,j,(i\neq j)} \mu_{dj\tau} \mathbf{D}_i^{\tau\mathrm{T}} \mathbf{D}_j^{\tau}$ measures the sum of weighted correlation between $\mathbf{D}_i^{\tau}$ and $\mathbf{D}_j^{\tau}$ for all $i, j \quad i \neq j$. Hence, the interesting information is actually contained in the second term which represents the prior information of the basis vectors. By including this term, the underlying correlation between the different basis vectors can be incorporated into the matrix factorization to yield results that reflect on this prior information. Therefore, with the factorial model in (16) the desired constraint assumes the following form:

$$f(\mathbf{D}) = -\sum_{\tau=0}^{\tau_{\max}} \log p_{D^{\tau}}(\mathbf{d}^{\tau}) \doteq \sum_{\substack{i,j \\ (i\neq j)}} \sum_{\tau} \mu_{ij\tau} \mathbf{D}_d^{\tau\mathrm{T}} \mathbf{D}_j^{\tau} \tag{17}$$

The prior on $\mathbf{H}$, this is constrained to an exponential distribution with independent decay parameters, namely,

$$\begin{aligned} p(\mathbf{H} \mid \boldsymbol{\Lambda}) &= \prod_{\phi=0}^{\phi_{\max}} p\left(\mathbf{H}^{\phi} \mid \boldsymbol{\Lambda}^{\phi}\right) \\ &= \prod_{\phi=0}^{\phi_{\max}} \left(\prod_{i=1}^{I} \prod_{t_s=1}^{T_s} p\left(\mathbf{H}_{i,t_s}^{\phi} \mid \Lambda_{i,t_s}^{\phi}\right) \right) \end{aligned} \tag{18}$$

where $p\left(\mathbf{H}_{i,t_s}^{\phi} \mid \Lambda_{i,t_s}^{\phi}\right) = \prod_{\phi} \prod_{i} \prod_{t_s} \Lambda_{i,t_s}^{\phi} \exp\left(-\Lambda_{i,t_s}^{\phi} \mathbf{H}_{i,t_s}^{\phi}\right)$. Following (18), the negative log prior on $\mathbf{H}$ is defined as $-\log p(\mathbf{H} \mid \boldsymbol{\Lambda}) = f(\mathbf{H}) = \sum_{\phi,i,t_s} \Lambda_{i,t_s}^{\phi} \mathbf{H}_{i,t_s}^{\phi} - \sum_{i,t_s,\phi} \log \Lambda_{i,t_s}^{\phi}$. In (18), it is worth pointing out that *each individual element* in $\mathbf{H}$ is constrained to a exponential distribution with independent decay parameter Λ_{i,t_s}^{ϕ} so that each element in $\mathbf{H}$ can be driven to be optimally sparse in the L_1-norm.

In this paper, the probabilistic framework is used for the purpose of developing a platform to incorporate the statistical correlation between $\mathbf{D}_i^{\tau}$ and $\mathbf{D}_j^{\tau}$ into the matrix factorization as part of the regularization. In feature extraction, such constraint is required in order to fully extract the basis especially in situation where the patterns contain overlapping features. Despite our proposed prior model for $\mathbf{D}$ stems from the rectified Gaussian distribution, it is a combination of constrained and unconstrained parameterization of the inverse covariance matrix. By substituting (17), (18) and (9) into (8), we may construct the following cost function:

$$\begin{aligned} L &= \sum_{f,t_s} \left[\frac{|\mathbf{Y}|_{f,t_s}^2}{\mathbf{Z}_{f,t_s}} - \log\left(\frac{|\mathbf{Y}|_{f,t_s}^2}{\mathbf{Z}_{f,t_s}} \right) - 1 \right] + f(\mathbf{H}) + f(\mathbf{D}) \\ &= \sum_{f,t_s} \left[\frac{|\mathbf{Y}|_{f,t_s}^2}{\mathbf{Z}_{f,t_s}} - \log\left(\frac{|\mathbf{Y}|_{f,t_s}^2}{\mathbf{Z}_{f,t_s}} \right) - 1 \right] + \sum_{\substack{i,j \\ (i\neq j)}} \sum_{\tau} \mu_{ij\tau} \mathbf{D}_i^{\tau\mathrm{T}} \mathbf{D}_j^{\tau} \\ &\quad + \sum_{\phi,i,t_s} \Lambda_{i,t_s}^{\phi} \mathbf{H}_{i,t_s}^{\phi} - \sum_{i,t_s,\phi} \log \Lambda_{i,t_s}^{\phi} \end{aligned} \tag{19}$$

where $\mathbf{Z} = \sum_i \sum_{\tau} \sum_{\phi} \mathbf{D}_i^{\tau} \mathbf{H}_i^{\phi}$. The sparsity term $f(\mathbf{H})$ forms the L_1-norm regularization to resolve the ambiguity by forcing all structure in $\mathbf{H}$ onto $\mathbf{D}$. Therefore, the sparseness of the solution is highly dependent on the regularization parameters Λ_{i,t_s}^{ϕ}.

1) Estimation of the spectral basis and temporal code:

Using (19), the derivatives corresponding to $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ are given by:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{D}_{f',i'}^{\tau'}} &= \sum_{f,t_s} \left(\left(\mathbf{Z}_{f,t_s} \right)^{-2} \left(\mathbf{Z}_{f,t_s} - |\mathbf{Y}|_{f,t_s}^2 \right) \right) \mathbf{H}_{i',t_s-\tau'}^{f-f'} + \sum_{j\neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'} \\ &= -\sum_{\phi,t_s} \left(\left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-2} \left(|\mathbf{Y}|_{f'+\phi,t_s}^2 - \mathbf{Z}_{f'+\phi,t_s} \right) \right) \mathbf{H}_{i',t_s-\tau'}^{\phi} + \sum_{j\neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'} \end{aligned} \tag{20}$$

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{H}_{i',t'}^{\phi'}} &= \sum_{f,t_s} \mathbf{D}_{f-\phi',i'}^{t_s-t_s'} \left(\left(\mathbf{Z}_{f,t_s} \right)^{-2} \left(\mathbf{Z}_{f,t_s} - |\mathbf{Y}|_{f,t_s}^2 \right) \right) + \Lambda_{i',t_s'}^{\phi'} \\ &= -\sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-2} \left(|\mathbf{Y}|_{f,t_s'+\tau}^2 - \mathbf{Z}_{f,t_s'+\tau} \right) \right) + \Lambda_{i',t_s'}^{\phi'} \end{aligned} \tag{21}$$

Using multiplicative gradient descent approach [10] as follows:

$$\theta \leftarrow \theta \cdot \left(\frac{\left[\nabla f(\theta)\right]_-}{\left[\nabla f(\theta)\right]_+} \right) \quad \text{where} \quad \nabla f(\theta) = \left[\nabla f(\theta)\right]_+ - \left[\nabla f(\theta)\right]_- \tag{22}$$

Therefore, we have:

$$\begin{aligned} \left[\nabla f(\theta)\right]_-^{\mathbf{D}} &= \sum_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-2} |\mathbf{Y}|_{f'+\phi,t_s}^2 \mathbf{H}_{i',t_s-\tau'}^{\phi} \\ \left[\nabla f(\theta)\right]_+^{\mathbf{D}} &= \sum_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-1} \mathbf{H}_{i',t_s-\tau'}^{\phi} + \sum_{j\neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'} \\ \left[\nabla f(\theta)\right]_-^{\mathbf{H}} &= \sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-2} \left(|\mathbf{Y}|_{f,t_s'+\tau}^2 \right) \right) \\ \left[\nabla f(\theta)\right]_+^{\mathbf{H}} &= \sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-1} + \Lambda_{i',t_s'}^{\phi'} \end{aligned} \tag{23}$$

Inserting (23) into (22) leads to the multiplicative update rules:

$$\mathbf{D}_{f',i'}^{\tau'} \leftarrow \mathbf{D}_{f',i'}^{\tau'} \frac{\sum_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-2} |\mathbf{Y}|_{f'+\phi,t_s}^2 \mathbf{H}_{i',t_s-\tau'}^{\phi}}{\sum_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-1} \mathbf{H}_{i',t_s-\tau'}^{\phi} + \sum_{j\neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'}} \tag{24}$$

$$\mathbf{H}_{i',t_s'}^{\phi'} \leftarrow \mathbf{H}_{i',t_s'}^{\phi'} \frac{\sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-2} \left(|\mathbf{Y}|_{f,t_s'+\tau}^2 \right) \right)}{\sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-1} + \Lambda_{i',t_s'}^{\phi'}} \tag{25}$$

The update of $\boldsymbol{\Lambda}$ follows by setting $\partial L / \partial \Lambda_{i',t_s'}^{\phi'} = 0$:

$$\frac{\partial L}{\partial \Lambda_{i',t_s'}^{\phi'}} = \mathbf{H}_{i',t_s'}^{\phi'} - \frac{1}{\Lambda_{i',t_s'}^{\phi'}} \quad , \quad \therefore \Lambda_{i',t_s'}^{\phi'} = \frac{1}{\mathbf{H}_{i',t_s'}^{\phi'}} \tag{26}$$

In (26), ' a/b ' represents element-wise divide. The multiplicative learning rules in (24) and (25) can be written in terms of matrix notation as:

$$\mathbf{D}^{\tau} \leftarrow \mathbf{D}^{\tau} \bullet \frac{\sum_{\phi} \left(\left(\overset{\uparrow\phi}{\mathbf{Z}} \right)^{-2} \bullet |\overset{\uparrow\phi}{\mathbf{Y}}|^2 \right) \overset{\rightarrow\tau^{\mathrm{T}}}{\mathbf{H}^{\phi}}}{\sum_{\phi} \left(\overset{\uparrow\phi}{\mathbf{Z}} \right)^{-1} \overset{\rightarrow\tau^{\mathrm{T}}}{\mathbf{H}^{\phi}} + \mathbf{D}^{\tau} \boldsymbol{\Xi}^{\tau\mathbf{T}}} \qquad \mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow\phi^{\mathrm{T}}}{\mathbf{D}^{\tau}} \left(\left(\overset{\leftarrow\tau}{\mathbf{Z}} \right)^{-2} \bullet |\overset{\leftarrow\tau}{\mathbf{Y}}|^2 \right)}{\sum_{\tau} \overset{\downarrow\phi^{\mathrm{T}}}{\mathbf{D}^{\tau}} \left(\overset{\leftarrow\tau}{\mathbf{Z}} \right)^{-1} + \boldsymbol{\Lambda}^{\phi}} \tag{27}$$

where " $\cdot$ " denotes the element wise operation and $\boldsymbol{\Xi}^{\tau}$ is a $I \times I$ matrix whose $(i,j)^{\text{th}}$ element is given by $\mu_{ij\tau}$ except the diagonal elements being zeros. In (27), $\boldsymbol{\Lambda}^{\phi}$ is the matrix representation of $\Lambda_{i',t_s'}^{\phi'}$ which is adaptive and the parameter $\mu_{ij\tau}$ in $\boldsymbol{\Xi}^{\tau}$ is non-adaptive which can be selected manually depending on applications. We term the above algorithm as the *variable regularized nonnegative matrix 2-D factorization with IS divergence* (IS-vRNMF2D) and have been summarized the

proposed algorithm in Table I. where $\psi = 10^{-6}$ is the threshold for ascertaining the convergence.

Table I: Pseudo codes for IS-vRNMF2D algorithms

IS-vRNMF2D algorithm

Input: $|\mathbf{Y}|^{.2}$, random nonnegative matrix $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$, ϕ, τ $\mu_{ij\tau}$ **Output:** $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$

Procedure:

Compute initialize cost value $Cost(1)$ using

$$L = \sum_{f,t_s}\left[\frac{|\mathbf{Y}|^2_{f,t_s}}{\mathbf{Z}_{f,t_s}} - \log\left(\frac{|\mathbf{Y}|^2_{f,t_s}}{\mathbf{Z}_{f,t_s}}\right) - 1\right] + \sum_{\substack{i,j \\ (i\neq j)}}\sum_{\tau}\mu_{ij\tau}\mathbf{D}_i^{\tau\mathrm{T}}\mathbf{D}_j^{\tau} + \sum_{\phi,i,t_s}\Lambda^{\phi}_{i,t_s}\mathbf{H}^{\phi}_{i,t_s} - \sum_{i,t_s,\phi}\log\Lambda^{\phi}_{i,t_s}$$

for *n=1: maximum number of iterations*

Compute $\mathbf{Z} = \sum_i\sum_{\tau}\sum_{\phi}\mathbf{D}_i^{\tau\,\downarrow\phi}\mathbf{H}_i^{\phi\,\rightarrow\tau}$.

- Update $\mathbf{D}^{\tau'}_{f',i'}$ using
- $$\mathbf{D}^{\tau} \leftarrow \mathbf{D}^{\tau} \bullet \frac{\sum_{\phi}\left(\left(\mathbf{Z}^{\uparrow\phi}\right)^{.-2} \bullet |\mathbf{Y}|^{\uparrow\phi.2}\right)^{\rightarrow\tau^{\mathrm{T}}}\mathbf{H}^{\phi}}{\sum_{\phi}\left(\mathbf{Z}^{\uparrow\phi}\right)^{.-1\,\rightarrow\tau^{\mathrm{T}}}\mathbf{H}^{\phi} + \mathbf{D}^{\tau}\boldsymbol{\Xi}^{\tau\mathbf{T}}}$$ for all ϕ, τ.

Re-compute $\mathbf{Z} = \sum_i\sum_{\tau}\sum_{\phi}\mathbf{D}_i^{\tau\,\downarrow\phi}\mathbf{H}_i^{\phi\,\rightarrow\tau}$ using the updated $\mathbf{D}^{\tau'}_{f',i'}$..

- Update $\mathbf{H}^{\phi'}_{i',t'_s}$ using
- $$\mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\sum_{\tau}\mathbf{D}^{\tau\,\downarrow\phi^{\mathrm{T}}}\left(\left(\mathbf{Z}^{\leftarrow\tau}\right)^{.-2} \bullet |\mathbf{Y}|^{\leftarrow\tau.2}\right)}{\sum_{\tau}\mathbf{D}^{\tau\,\downarrow\phi^{\mathrm{T}}}\left(\mathbf{Z}^{\leftarrow\tau}\right)^{.-1} + \boldsymbol{\Lambda}^{\phi}}$$ for all ϕ, τ.

- Update $\Lambda^{\phi}_{i,t_s} = 1/\mathbf{H}^{\phi}_{i,t_s}$

Re-compute the cost value $Cost(n)$ using the updated parameters $\mathbf{D}^{\tau'}_{f',i'}$ and $\mathbf{H}^{\phi'}_{i',t'_s}$.

end

Stopping criterion: $\frac{Cost(n-1) - Cost(n)}{Cost(n)} < \psi$.

We can convert the proposed method to SNMF2D, the cost function with sparse penalty is $L = \sum_{f,t_s}\left[\frac{|\mathbf{Y}|^2_{f,t_s}}{\tilde{\mathbf{Z}}_{f,t_s}} - \log\left(\frac{|\mathbf{Y}|^2_{f,t_s}}{\tilde{\mathbf{Z}}_{f,t_s}}\right) - 1\right] + \Lambda f(\mathbf{H})$ subject to $\mathbf{H}^{\phi} \sim p\left(\mathbf{H}^{\phi} \mid \Lambda\right) = \prod_{i=1}^{I}\prod_{t_s=1}^{T_s}\Lambda\exp\left(-\Lambda\mathbf{H}^{\phi}_{i,t_s}\right)$ where Λ is a constant and can be set manually, $\tilde{\mathbf{Z}} = \sum_{\tau,\phi}\tilde{\mathbf{D}}^{\tau\,\downarrow\phi}\mathbf{H}^{\phi\,\rightarrow\tau}$, $\tilde{\mathbf{D}}^{\tau}_{f,i} = \mathbf{D}^{\tau}_{f,i} \Big/ \sqrt{\sum_{\tau,f}(\mathbf{D}^{\tau}_{f,i})^2}$ and the negative log prior on $\mathbf{H}$ is defined as $f(\mathbf{H})$ can be $L_1 - norm$ given by $f(\mathbf{H}) = \|\mathbf{H}\|_1 = \sum_{\phi,i,t_s}\mathbf{H}^{\phi}_{i,t_s}$. In the proposed method, prior distributions on both $\mathbf{D}$ and $\mathbf{H}$ have been incorporated into the cost function, and it can be simplified to IS-SNMF2D model by setting the independent decay parameter $\Lambda^{\phi}_{i,t_s} = \Lambda$ and letting $\mu_{ij\tau} = 0$ for all elements. This explicitly constrains a uniform regularization across all element in $\mathbf{H}^{\phi}$. However, unlike the standard SNMF2D in [16], the above SNMF2D is optimized using the IS divergence and we term this algorithm as IS-SNMF2D. The IS-SNMF2D method will be compared with our proposed IS-vRNMF2D algorithm

III. Experimental Results And Analysis

A. TF Representation of SCSS

The TF representation of the mixture in (1) is given by $Y(f,t_s) = \sum_{i=1}^{I}X_i(f,t_s)$ where $Y(f,t_s)$, $X_i(f,t_s)$ denote the TF components which are obtained by applying the short time Fourier transform (STFT) e.g. $Y(f,t_s) = STFT\left(y(t)\right)$. The time slots are given by $t_s = 1,2,\ldots,T_s$ while frequencies by $f = 1,2,\ldots,F$, we represent this as $\mathbf{Y} = \left[Y(f,t_s)\right]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$ and $\mathbf{X}_i = \left[X_i(f,t_s)\right]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$. The power spectrogram is defined as the squared magnitude STFT and hence, its matrix representation is given by $|\mathbf{Y}|^{.2} \approx \sum_{i=1}^{I}|\mathbf{X}_i|^{.2}$. The matrices we seek to determine are $\left\{|\mathbf{X}_i|^{.2}\right\}_{i=1}^{I}$ and this will be obtained by using our proposed matrix factorization as $\left|\tilde{\mathbf{X}}_i\right|^{.2} = \sum_{\tau}\sum_{\phi}\mathbf{D}_i^{\tau\,\downarrow\phi}\mathbf{H}_i^{\phi\,\rightarrow\tau}$ with $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ estimated using (27). Once these matrices are estimated, we form the i^{th} binary mask according to $\mathbf{mask}_i(f,t_s) = 1$ if $\left|\tilde{X}_i(f,t_s)\right|^{.2} > \left|\tilde{X}_j(f,t_s)\right|^{.2}$ and zero otherwise. Finally, the estimated time-domain sources are obtained as $\tilde{\mathbf{x}}_i = STFT^{-1}\left(\mathbf{mask}_i \bullet \mathbf{Y}\right)$ for $i = 1,\ldots,I$ where $\tilde{\mathbf{x}}_i = \left[\tilde{x}_i(1),\ldots,\tilde{x}_i(T)\right]^{\mathbf{T}}$ denotes the i^{th} estimated audio sources in time-domain.

B. Experiment Set-up

The proposed monaural source separation algorithm is tested on recorded audio signals. All simulations are conducted using a PC with Intel Core 2 CPU 6600 @ 2.4GHz and 2GB RAM. The experiments consist of four audio sources (i.e. male speech, female speech, jazz and piano music), two mixture types (i.e. mixture of music and speech signals; mixture between different type of music signals) and each mixture is generated by adding two sources. All mixtures are sampled at 16kHz sampling rate. The TF representation is computed by applying the STFT with 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 50Hz to 8kHz with 24 bins per octave. As for our proposed algorithm, the convolutive components are selected as

follows: (i) For jazz and speech mixture, $\tau_{\max}=2$, $\phi_{\max}=2$, and $\mu_{ij\tau}=1.5$ for $\forall i,j,\tau$. (ii) For jazz and piano mixture, $\tau_{\max}=6$, $\phi_{\max}=9$, and $\mu_{ij\tau}=2.5$ for $\forall i,j,\tau$. (iii) For piano and speech mixture, $\tau_{\max}=6$, $\phi_{\max}=9$, and $\mu_{ij\tau}=2$ for $\forall i,j,\tau$. The separation performance is measured by the Signal-to-Distortion Ratio (SDR) and the routines for computing this is obtained from the SiSEC'08 webpage [30].

C. Impact of Regularization on Matrix Factorization and Source Separation

In this section, we will investigate the impact of regularization. We will show that when the sparse constraints are not controlled, the matrix factorization will be either under- or over-sparse and this will result in ambiguity in the estimation of recovered sources. We first show the TF domain of the original audio signals (male speech and jazz music) and its mixture in Figure 1. Figures 2-4 show the factorization results based on the IS-SNMF2D and our proposed method. The temporal codes in Figure 2 show that the resulting factorization is under-sparse when Λ^{ϕ}_{i,t_s} is fixed to a small value whereas Figure 3 shows the case of Λ^{ϕ}_{i,t_s} fixed to a large value that resulted in over-sparse factorization. On the other hand, Figure 4 shows the factorization that is just sparse enough by using our proposed adaptive sparsity parameters. This is an important factor in SCASS that will crucially affect the separation performance and we will shortly demonstrate this effect in Figures 5 and 6. In general source separation problem, the performance depends on how distinguishable the two spectral bases $\mathbf{D}_1^{\tau}$ and $\mathbf{D}_2^{\tau}$ are from each other [20]. When $\mathbf{D}_1^{\tau}$ and $\mathbf{D}_2^{\tau}$ are distinguishable and since $\{\mathbf{H}_i^{\phi}\}_{i=1}^{2}$ are sparse, then the mixing ambiguity between $|\mathbf{X}_1|^{\cdot 2}$ and $|\mathbf{X}_2|^{\cdot 2}$ which constitutes the magnitude of interference in the TF domain will be small. Now the requirement that $\mathbf{D}_1^{\tau}$ and $\mathbf{D}_2^{\tau}$ to be distinguishable has been made possible by our proposed algorithm since we have explicitly incorporated the modified Gaussian prior information onto these spectral bases so that the spectral overlap between any two bases is as small as possible. As a direct result and by exploiting the sparse property of $\{\mathbf{H}_i^{\phi}\}_{i=1}^{2}$, it is now possible to determine $|\mathbf{X}_1|^{\cdot 2}$ and $|\mathbf{X}_2|^{\cdot 2}$ from $|\mathbf{Y}|^{\cdot 2}$. Figures 2 to 4 show the results of $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ when the factorization is obtained by using the IS-SNMF2D and our proposed method. In comparison, the estimation of $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ based on the IS-SNMF2D are very coarse when the sparse regularization is uncontrolled. Hence, this results in poorer estimation of the recovered sources as shown in Figure 5.

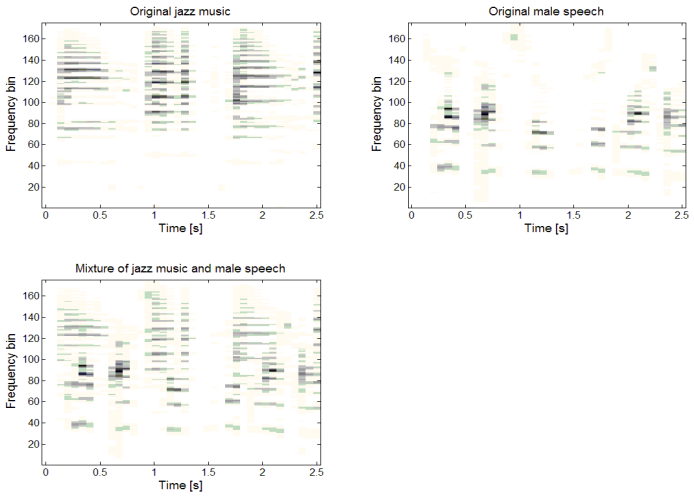

Figure 1: The spectrogram of jazz music, male speech (top panels) and mixed signal (bottom panels)

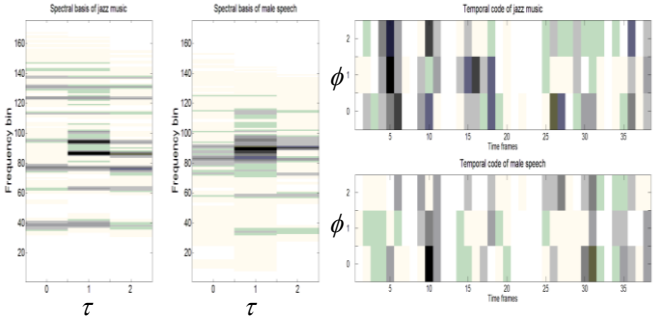

Figure 2: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ using the IS-SNMF2D by setting $\mu_{ij\tau}=0$ and $\Lambda^{\phi}_{i,t_s}=c=0.01$

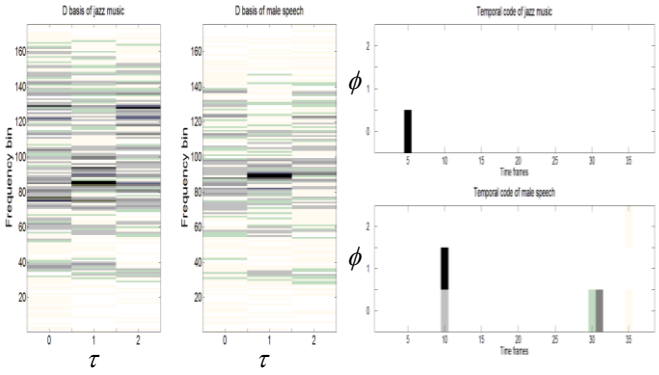

Figure 3: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ using the IS-SNMF2D by setting $\mu_{ij\tau}=0$ and $\Lambda^{\phi}_{i,t_s}=c=100$

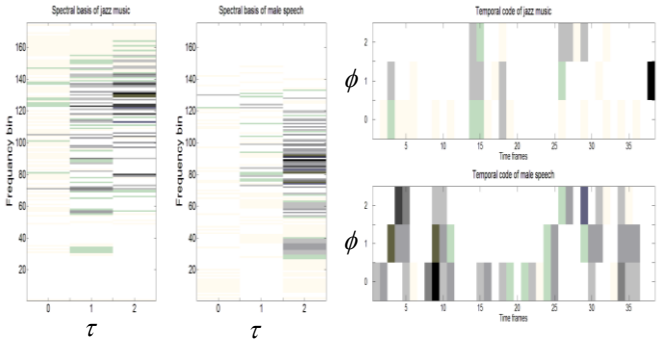

Figure 4: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ using the IS-vRNMF2D

In Figure 5, the top and middle panels clearly reveal that good separation performance require suitably controlled sparse regularization. In the case of under-sparse factorization, the spectral basis of the source repeats too frequently in the spectrogram and this results in redundant information which still retains the mixed components as noted in the top panels (indicated by the red box marked area). In the case of over-sparse factorization, the spectral basis of the source occurs too rarely in the spectrogram and this results in less 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 IS-vRNMF2D, it assigns a regularization parameter to each temporal code which is individually and adaptively tuned to yield the optimal number of times the spectral basis of a source recurs in the spectrogram. This is noted in the bottom panels which clearly show the optimal separation result. To investigate the effects of $\mu_{ij\tau}$ and Λ_{i,t_s}^{ϕ} on the separation performance, three cases are conducted:

Case (i): No sparseness $\Lambda_{i,t_s}^{\phi}=0$ and $\mu_{ij\tau}$ is varied as $\mu_{ij\tau}=0,0.5,1.0,\ldots,5$.

Case (ii): Uniform and constant sparseness $\Lambda_{i,t_s}^{\phi}=c$ and $\mu_{ij\tau}$ is varied as $\mu_{ij\tau}=0,0.5,1.0,\ldots,5$.

Case (iii): Adaptive sparseness according to (32) and $\mu_{ij\tau}$ is varied as $\mu_{ij\tau}=0,0.5,1.0,\ldots,5$.

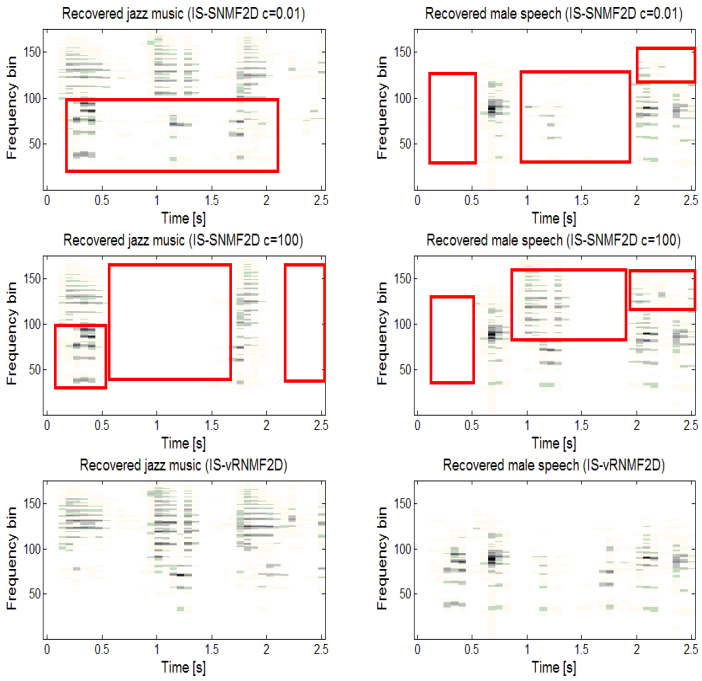

Figure 5: Separation results (IS-SNMF2D and IS-vRNMF2D)

The separation results in terms of the SDR are given in Figure 6. Case (iii) renders the best performance and the average improvement can be summarized as follows: (i) For music mixture, the average SDR improvement is 0.7dB per source and (ii) for mixtures of music and speech (first two panels of Figure 6), the improvement is 1dB per source. As expected, incorporating sparseness into the factorization improves the separation performance as noted in all panels of Figure 6. The results have also clearly indicated that there are certain values of $\mu_{ij\tau}$ where the algorithm performs the best. In the case of music and speech mixtures, the best performance is obtained when $\mu_{ij\tau}$ ranges from 1.5 to 2.5. As for music mixture, the best performance is obtained when $\mu_{ij\tau}$ ranges from 2 to 3. However, when $\mu_{ij\tau}$ is set to be either too low or high, the performance will degrade. It is also worth pointing out that the separation results are rather coarse when the factorization is non-regularized (i.e. without prior pdf on $\mathbf{D}$ and $\mathbf{H}$). Here, we see that for music mixture, the SDR is only 2.7dB and for mixtures of music (jazz or piano) and speech, the average SDR is only 3.3dB. However, by incorporating regularization (i.e. through $\mu_{ij\tau}$ and Λ_{i,t_s}^{ϕ}), the performance has significantly increased over twice for both types of audio mixture. This is clearly evident in the case of jazz and speech mixture when Λ_{i,t_s}^{ϕ} is adaptive while $\mu_{ij\tau}$ is set to 1.5, the SDR result is 7.7dB (≈5.89 in linear scale) whereas for the case of without regularization the SDR result is only 4dB (≈2.51 in linear scale). This amounts to slightly above twice better performance using the proposed regularization than that without regularization.

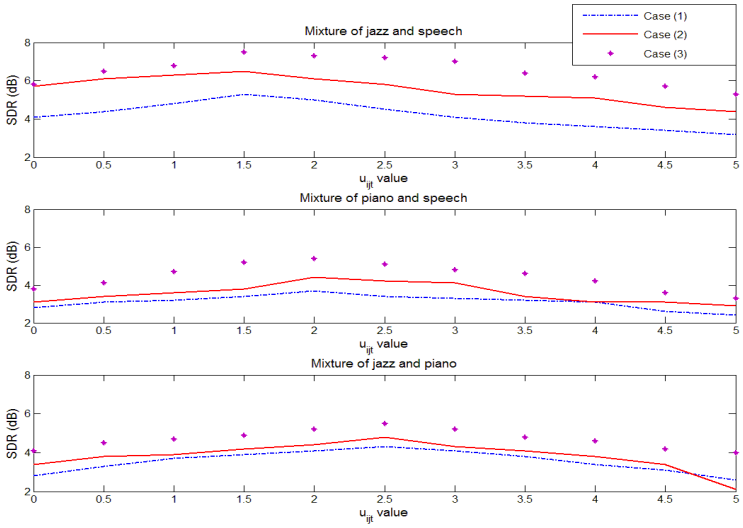

Figure 6: SDR results as a function of $\mu_{ij\tau}$ and sparseness

We also added Figure 7 which shows the convergence trajectory results. In particular, the algorithm converges very quickly to the steady-state solution in no more than twenty iterations.

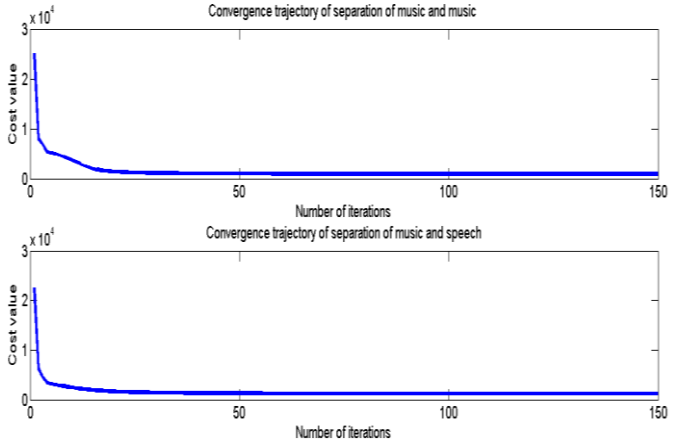

Figure 7: convergence trajectory for the proposed algorithm

D. Comparison with other NMF methods

In Section III C, analysis has been carried out to investigate effects on regularization parameters in source separation. In this evaluation, we compare the proposed method with latest MAP based NMF source separation methods. These consist of the followings:

- Automatic Relevance Determination NMF (NMF-ARD) proposed in [27] 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, $p(\mathbf{H}\mid\lambda)=\prod_i \lambda_i^{T_s}\exp-\left(\lambda_i\sum_{t_s}\mathbf{H}_{i,t_s}\right)$ and uses ARD approach to determine the desirable number of components in $\mathbf{D}$. The initialization number of components in $\mathbf{D}$ is 10.
- NMF-ARD proposed in [28] exploits a Bayesian framework that amounts to imposing Gamma distribution priors with tied precision parameter ψ_i for pruning and thereby enables estimation of the NMF model order. Each precision parameter ψ_i is given by a Gamma distribution, namely $p(\psi_i\mid\varsigma_i,\eta_i)=\frac{\eta_i^{\varsigma_i}}{\Gamma(\varsigma_i)}\psi_i^{\varsigma_i-1}\exp(-\psi_i\eta_i),\ \psi_i\geq 0$ and uses ARD approach to determine the desirable number of components in $\mathbf{D}$. The hyperparameters setting are $\varsigma=\eta=1$ and the initialization number of components in $\mathbf{D}$ is 10.
- NMF with Temporal Continuity and Sparseness Criteria [15] (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 [15], the temporal continuity α is chosen as [0,1,10,100,1000], sparseness weight λ is chosen as [0,1,10,100,1000] and the initialization number of components in $\mathbf{D}$ is tested from 2 to 10. The best separation result is retained for comparison.

The experiments are based on separating two audio sources from a single channel mixture. For all NMF-ARD methods, the final components will be clustered with respect to each source when the number of components exceeds than number of sources. Since more than two components are used and the tested methods are blind, there is no information to tell which component belongs to which source. Thus, we utilize the clustering method proposed in [15] where the original sources are used as reference to create component clusters for each source. The following figures show an example of separating mixture of jazz music and female speech. The TF domain of the original audio signals and its mixture are shown in Figure 8.

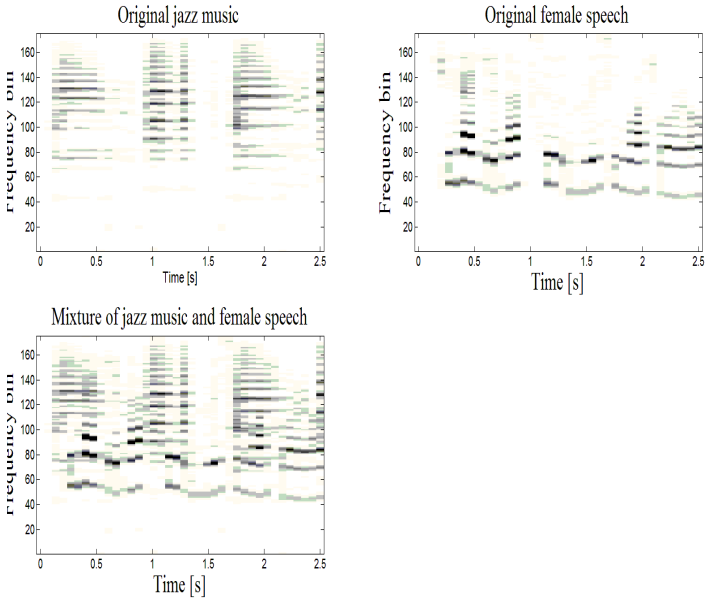

Figure 8: The spectrogram of jazz music, female speech (top panels) and mixed signal (bottom panels).

Figures 9 show the factorization results based on the proposed method, NMF-ARD in [27], NMF-ARD in [28] with Gamma priors and NMF-TCS method [15], respectively.

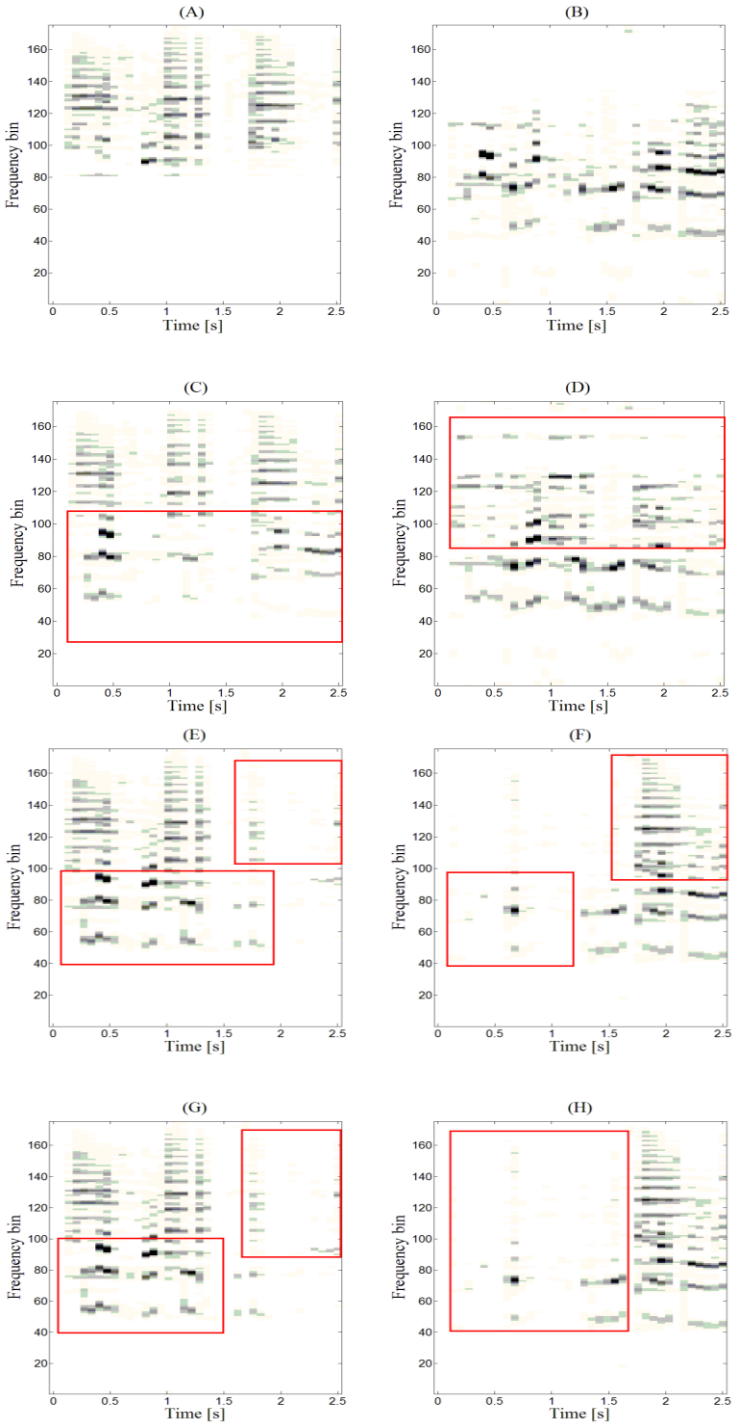

Figure 9: Separated signals in TF domain. (A)-(B): Proposed method. (C)-(D): NMF-ARD with exponential prior. (E)-(F): NMF-ARD with Gamma priors. (G)-(H): NMF-TCS.

In Figure 9, panels (A)-(B) show that the proposed method has successfully recovered both jazz music and female speech. On the other hand, panels (C)-(H) show that the compared NMF methods are less successful in separating the 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. The overall results are summarized in Table III and Figure 10.

Table III: SDR results in dB using different NMF methods

Mixtures	**Separation methods**	**Average SDR (dB) with standard deviation**
Music and music	Proposed method	6.6±0.6
	NMF-ARD [27]	2.3±0.25
	NMF-ARD [28]	1.4±0.4
	NMFTCS	3.5±0.3
Music and speech	Proposed method	7.2±0.35
	NMF-ARD [27]	2.6±0.2
	NMF-ARD [28]	1.1±0.15
	NMFTCS	3.8±0.5

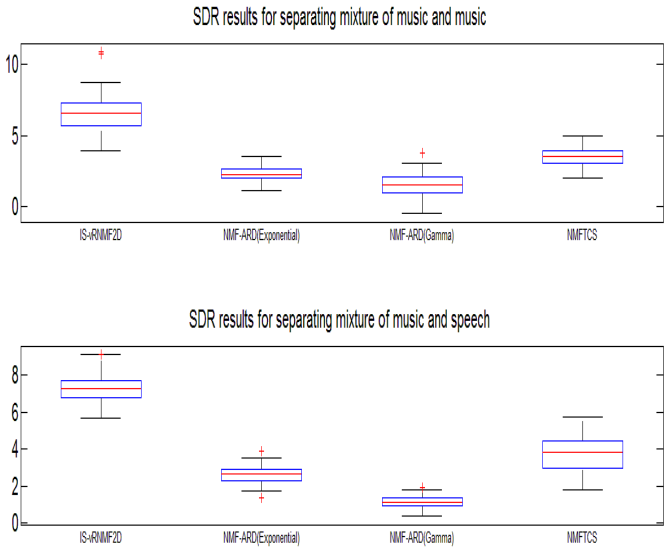

Figure 10: boxplot of SDR results with standard deviation for using different NMF methods

Analyzing the results, we may summarize the average improvement of our method over the other NMF related methods as follows: (i) For music mixture, the average improvement is 4.5dB per source. (ii) For mixture of music and speech, the improvement is 5dB per source. In percentage, this translates to an average improvement of 125% for mixture of music sources and 150% for mixture of music and speech sources. NMF-TCS leads higher performance than NMF-ARD because this method considers temporal information of the code. However, it does not capture the spectral dependency of the frequency patterns within the signal. As a result, multiple notes in an audio signal will be characterized as one note and this leads to substitution of error in the form of interference when the spectrogram is reconstructed from the obtained spectral basis and temporal code. On the contrary, our proposed algorithm renders a more optimal part-based decomposition for audio source separation. The decomposition is more unique than the

above methods under certain conditions e.g. variable sparseness and prior pdf on spectral basis leading to more robust separation results.

In the final experiment, the proposed method is tested on professionally produced music recordings of well-known song namely "You raise me up" by Kenny G. The music consists of two excerpts of length approximately 20s on mono channel and resampled to 16 kHz. The song is an instrumental music consist of saxophone and piano sound. The factors of τ and ϕ shifts are set to have $\tau_{\max} = 8$ and $\phi_{\max} = 32$ while $\mu_{ij\tau}$ is set to 2.5. Since the original source spatial images are not available for this experiment, the separation performance is assessed perceptually and informally by analyzing the log-frequency spectrogram of the estimated source images and listening to the separated sound. This task was a tough task since the instruments play many different notes in the recording. Figure 11 shows the separation results of the saxophone and piano sound. The high pitch of continuous saxophone sound is shown in the Figure 11(B) while the notes of the piano are evidently present in Figure 11(C). In the overall, our proposed method has successfully separated the professionally produced music recordings and gives a perceptually pleasant listening experience.

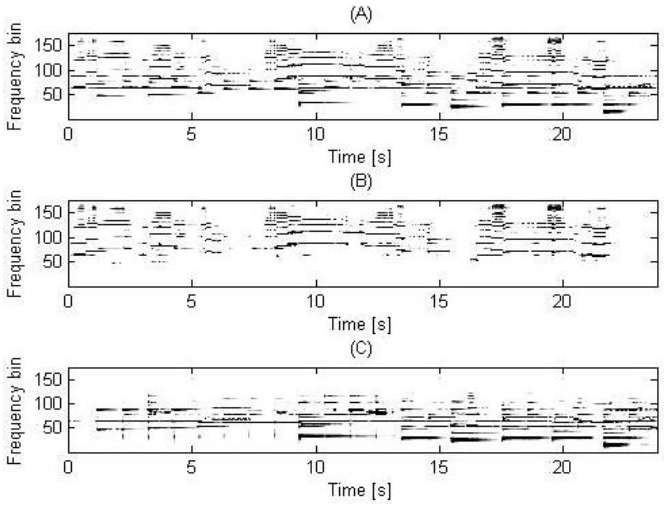

Figure 11: Separation result for song "You raised me up" by Kenny G. (A) Recorded music. (B) Separated saxophone sound. (C) Separated piano sound.

IV. Conclusion

The paper presents a new variable regularized nonnegative matrix two-dimensional factorization with Itakura-Saito divergence. The impetus behind the proposed work is that sparseness achieved by the conventional SNMF is not effective enough; in source separation it is necessary to yield control over the degree of sparseness explicitly for each temporal code. The proposed method enjoys at least three significant advantages: Firstly, it avoids strong constraints of separating mixture without training knowledge where only single channel is provided. Secondly, the sparse regularisation term is adaptively tuned using a *maximum a posteriori* approach to yield the desired sparse decomposition. Finally, the modified Gaussian prior is formulated to express the basis vectors more effectively; thus enabling the spectral and temporal features of the sources to be extracted more efficiently.

References

[1] N. Tengtrairat, Bin Gao, W.L. Woo and S.S. Dlay, "Single Channel Blind Separation using Pseudo-Stereo Mixture and Complex 2D Histogram," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 24, no. 11, pp. 1-14, 2013.

[2] Bin Gao, W.L. Woo and S.S. Dlay, "Unsupervised Single Channel Separation of Non-Stationary Signals using Gammatone Filterbank and Itakura-Saito Nonnegative Matrix Two-Dimensional Factorizations," *IEEE Trans. on Circuits and Systems I*, vol. 60, no. 3, pp. 662-675, 2013.

[3] Bin Gao, W.L. Woo and S.S. Dlay, "Variational Regularized Two-Dimensional Nonnegative Matrix Factorization," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 23, no.5, pp. 703-716, 2012.

[4] Bin Gao, W.L. Woo and S.S. Dlay, "Adaptive Sparsity Nonnegative Matrix Factorization for Single Channel Source Separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 989-1001, 2011.

[5] Bin Gao, W.L. Woo and S.S. Dlay, "Single Channel Blind Source Separation Using EMD-Subband Variable Regularized Sparse Features," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 961-976, 2011.

[6] Jingyi Zhang, W.L. Woo, and S.S. Dlay, "Blind Source Separation of Post-Nonlinear Convolutive Mixture," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2311-2330, 2007.

[7] P. Gao, W.L. Woo and S.S. Dlay, "Nonlinear Signal Separation for Multi-Nonlinearity Constrained Mixing Model", *IEEE Trans. on Neural Networks*, vol. 17, no. 3, pp. 796-802, May 2006.

[8] W.L. Woo and S.S. Dlay, "Neural Network Approach to Blind Signal Separation of Mono-nonlinearly Mixed Signals", *IEEE Trans. on Circuits and System I*, vol. 52, no. 2, pp. 1236-1247, June 2005.

[9] Jingyi Zhang, W.L. Woo, and S.S. Dlay, "An Expectation-Maximisation Approach to Blind Source Separation of Nonlinear Convolutive Mixture," *IET Signal Processing*, vol. 1, no. 2, pp. 51-65, 2007.

[10] R.C. Zhi, M. Flierl, Q. Ruan, and W.B. Kleijn, "Graph-Preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 41, pp. 38–52, 2011.

[11] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '03)*, New Paltz, NY, USA, October 2003, pp. 177–180.

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

[13] C. Fevotte, N. Bertin and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no 3, pp. 793-830, 2009.

[14] S. Rickard and A. Cichocki, "When is non-negative matrix decomposition unique?" in *Proc. of Information Sciences and Systems, CISS 2008, 42nd Annual Conference on*. March 2008, pp. 1091 - 1092

[15] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria." *IEEE Transactions on Audio, Speech, and Language Processing*, vol 15, no. 3, pp. 1066–1074, March 2007

[16] M. Morup and M. N. Schmidt, "Sparse non-negative matrix factor 2-D deconvolution." Tech. Rep Technical University of Denmark, Copenhagen, Denmark, 2006.

[17] J. Zhang, L.Wei, X.Feng and Y. Wang, "Pattern expression nonnegative matrix factorization: algorithm and applications to blind source separation" *Computational Intelligence and Neuroscience*, vol. 2008, Article ID 168769, 10 pages, 2008

[18] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models" *Computational Intelligence and Neuroscience*. doi: 10.1155/2009/785152. 2009.

[19] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc of 6th Intl. Congress on Acoustics*, Tokyo, Japan, Aug. 1968. pp. 17–20.

[20] A. Cichocki and R. Zdunek, "Multilayer nonnegative matrix factorization," *Electronics Letters*, vol. 42, no. 16, pp. 947-948, 2006.

[21] G. Zhou, Z. Yang, S. Xie, J. Yang, "Online blind source separation using incremental nonnegative matrix factorization with volume constraint," *IEEE Trans. on Neural Networks*, vol. 22, no. 4, pp. 550–560, Nov. 2011

[22] 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

[23] D. Donoho and V. Stodden, "When does non-negative matrix factorisation give a correct decomposition into parts?" *in Proceeding of Advances in Neural Information Processing Systems*, vol. 17, 2003.

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

[25] R. Zdunek, A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization," *International Journal of Signal Processing*, vol. 87, no. 8, pp. 1904-1916, 2007.

[26] X. Li, W.K Cheung, J.M. Liu, "Improving POMDP tractability via belief compression and clustering," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, pp. 125–136, 2010.

[27] M. Mørup and K.L. Hansen "Tuning pruning in sparse non-negative matrix factorization," in *Proc. of 17th European Signal Processing Conference (EUSIPCO'09)*, Glasgow, Scotland, 2009.

[28] V.Y.F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the beta-divergence," in *Proc of NIPS workshop on Sparse Representation and Low-rank Approximation*, 2011.

[29] M. Harva and A. Kaban, "Variational learning for rectified factor analysis," *Signal Processing*, vol. 87, no. 3, 2007.

[30] "Signal Separation Evaluation Campaign (SiSEC 2008)," 2008. [Online]. Available: http://sisec.wiki.irisa.fr

[31] Quanquan Gu, Jie Zhou, Two Dimensional Nonnegative Matrix Factorization, The 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, pp.2069-2072, 2009

Bin Gao received his B.S. degree in communications and signal processing from Southwest Jiao Tong University in Chengdu (2001-2005), China and MSc degree with Distinction in communications and signal processing from Newcastle University, UK (2006-2007). He obtained PhD degree (2007-2011) from Newcastle University and his research topic was single channel blind source separation. Upon graduation, he worked as a Research Associate with the same university on wearable acoustic sensor technology. Currently, he is an Associate Professor with the School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. His research interests include sensor signal processing, machine learning, structural health monitoring, and nondestructive testing and evaluation where he actively publishes in these areas. He is a very reviewer for many international journals and long standing conferences. Dr Gao is a Member of the Institution of Electrical and Electronic Engineering (IEEE).

Wai Lok Woo was born in Malaysia. He received the BEng degree (1st Class Hons.) in Electrical and Electronics Engineering and the PhD degree from the Newcastle University, UK. He was awarded the IEE Prize and the British Scholarship to continue his research work. He is currently a Senior Lecturer with the School of Electrical and Electronic Engineering. His major research is in the mathematical theory and algorithms for nonlinear signal and image processing. This includes areas of machine learning for signal processing, blind source separation, multidimensional signal processing, signal/image deconvolution and restoration. He has an extensive portfolio of relevant research supported by a variety of funding agencies. He has published over 250 papers on these topics on various journals and international conference proceedings. Currently, he is Associate Editor of several international journals and has served as lead-guest editor of journals' special issues. He actively participate in international conferences and workshops, and serves on their organizing and technical committees. In addition, he is a consultant to a number of industrial companies that involve the use of statistical signal and image processing techniques. Dr Woo is a Senior Member of the IEEE and a Member of the Institution Engineering Technology (IET).

Bingo W-K. Ling is a fellow of IET and a senior member of IEEE. He received the B.Eng. (Hons) and M.Phil. degrees from the department of Electrical and Electronic Engineering, the Hong Kong University of Science and Technology, in 1997 and 2000, respectively, and the Ph.D. degree from the department of Electronic and Information Engineering from the Hong Kong Polytechnic University in 2003. He joined the King's College London and the University of Lincoln as a Lecturer and a Principal Lecturer in 2004 and 2010, respectively. Then, he promoted to a Reader in 2011. Finally, he joined the Guangdong University of Technology as a Hundred-People-Plan Distinguished Professor in 2012. He obtained the National Young-Thousand-People-Plan Award in 2013 as well as "Outstanding Reviewers" Award from the IEEE Instrumentation and Measurement Society in 2008 and 2012. He serves as a Chairman of the IET Guangzhou Local Network and a member of several technical committees of the IEEE Circuits and Systems Society. He is associate editors of several international journals and served as editor-in-chiefs of several special issues of international journals. He has delivered about 20 seminars and published about 5 book chapters, 80 international journal papers and 55 international conference papers.