Single-Channel Signal Separation using Spectral Basis Correlation with Sparse Nonnegative Tensor Factorization

P. Parathai[1], N. Tengtrairat[2], W. L. Woo[3], and Bin Gao[4]

Abstract -- **A novel approach for solving the single-channel signal separation (SCSS) is presented the proposed sparse nonnegative tensor factorization under the framework of maximum *a posteriori* probability and adaptively fine-tuned using the hierarchical Bayesian approach with a new mixing mixture model. The mixing mixture is an analogy of a stereo signal concept given by one real and the other virtual microphones. An "imitated-stereo" mixture model is thus developed by weighting and time-shifting the original single-channel mixture. This leads to an artificial mixing system of dual channels which gives rise to a new form of spectral basis correlation diversity of the sources. Underlying all factorization algorithms is the principal difficulty in estimating the adequate number of latent components for each signal. This paper addresses these issues by developing a framework for pruning unnecessary components and incorporating a modified multivariate rectified Gaussian prior information into the spectral basis features. The parameters of the imitated stereo model are estimated via the proposed sparse nonnegative tensor factorization with Itakura-Saito divergence. In addition, the separability conditions of the proposed mixture model are derived and demonstrated that the proposed method can separate real-time captured mixtures. Experimental testing on real-audio sources has been conducted to verify the capability of the proposed method.**

Keywords — **Blind source separation, underdetermined mixture, tensor factorization, unsupervised learning, multiplicative updates, source modeling.**

[1] School of Software Engineering, Payap University, Chiang Mai, Thailand: phetcharat@payap.ac.th
[2] School of Software Engineering, Payap University, Chiang Mai, Thailand: naruephorn_t@payap.ac.th
[3] Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, England, United Kingdom: wai.l.woo@northumbia.ac.uk
[4] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China: bin_gao@uestc.edu.cn

1 INTRODUCTION

Blind source separation (BSS) [29, 47] is the process of separating individual source signals without using the training information of the sources. BSS is flourishing in numerous fields, including underwater signal processing [31], communication [27], speech enhancement [37], biomedical [14] and audio signal recognitions [42]. One classical problem of BSS is the so-called "cocktail party problem" [4] is psychoacoustic phenomenon that indicates to the significant human capability to attend and recognize the speaker from the interference environment. An extreme case of BSS is termed as single channel blind source separation (SCBSS). The SCBSS aims to discover individual source signals from a single mixture recording without any a priori information of the sources. Since the number of source signals $\{x_n(t)\}_{n=1,2,\dots,N}$ is greater than the number of the observed mixture $y(t)$, this is known as the underdetermined SCBSS problem [2, 12, 20, 30, 33, 44]. Many algorithms have been successfully developed for SCBSS. The conventional ICA method [19] was adapted to the case of SCBSS which is known as single-channel ICA (SCICA). In [1, 21, 28, 40], a SCICA method is proposed which maps an observed single-channel mixture into a multi-channel model by breaking the observed vector into a sequence of contiguous blocks. These blocks are treated as a matrix where the standard ICA can then be employed to estimate the underlying sources. Generally, it has two major drawbacks of the SCICA method: first, the algorithm assumes stationary sources; and second, the sources are assumed to be disjoint in the frequency domain. These assumptions however do not always hold in applications. In the SCICA method, the sources are modeled as sparse combination of a set of time-domain basis functions which are initially derived using the standard ICA. This method renders optimal separation when the ICA basis functions corresponding to each source have minimal time-domain overlap. In the case where the basis functions have significant overlap with each other e.g. mixture of two speech sources or the basis functions of two sources are very similar, the method performs poorly. In [46], a single-channel mixture was applied multi-component radar or signal-dependent transforms [10, 32] to generate a multi-channel mixture. The generated multi-mixtures are subsequently separated by ICA. Another approach is decomposing a signal of interest into different sources is nonnegative matrix factorization (NMF) approach [24]. The NMF has been used for sound source separation of single-channel mixtures using the multiplicative update (MU) algorithm to solve its parametrical optimization based on the least square distance and Kullback-Leibler divergence as cost function in [25, 34, 35]. Later, other families of cost functions were continuously proposed for example the Beta divergence [22], Csiszár's divergences [5], and Itakura-Saito divergence [7]. Popular method in this category is the sparse non-negative matrix factorization (SNMF) [15] where sparseness constraints can be included into the cost function. The two-dimensional sparse NMF deconvolution (SNMF2D) [3, 11] uses a double convolution to model both spreading of spectral basis and variation of temporal structure inherent in the signals. In [23], sources are assumed to be non-stationary and nonnegative. The canonical tensor and least squares method is used to estimate the mixing model. The source is then discovered by a minimum mean-squared error beamformer approach without any hypothetical limitation. On a parallel development, NTF under a parallel factor analysis (PARAFAC) structure where the channel spectrograms are jointly modeled by a 3-valence tensor have been introduced in [8, 36]. Clustering of the spatial cues to group the NTF components (cNTF) is developed in [6] for multichannel audio source separation. In most applications, if the number of components (K) is too small, the data does not fit the model well. Conversely, if K is too large, then overfitting occurs. Choosing the right model is in particular challenging in the PARAFAC model as the number of components is specified for each modality separately. While these approaches increase the accuracy of matrix factorization, it only works when large sample dataset is available. However, the sparsity parameter is manually determined. This will then cause over or under sparsity that effect to separation performance. To find an elegant solution for this dichotomy between data fidelity and overfitting, it is crucial that the "right" model order of components is selected.

In this paper, a new framework for single-channel blind source separation (SCBSS) is proposed. The proposed solution separates sources from a single-channel without relying on training information about the original sources. The advantages of the proposed method are: 1) Analogous to the stereo signal concept given by one microphone. We create an imitated-stereo mixture from a single-channel mixture signal. From this stereo mixture the proposed algorithm can be employed to separate individual source from the mixtures. 2) Overcoming the limitations associated with the above NTF problems. Unlike the NTF, our model assigns a probability distribution to each element of unknown non-negative matrix $\boldsymbol{H} = \{h_{kt_s}\}$, where h, k, and t_s are an activation coefficient, audio components, time slots, respectively, and a sparsity parameter associated with each probability distribution. This sets up a platform to enable the sparsity parameter to be *individually optimized for each element code*. 3) Automatically detecting the optimal number of components K of the individual source (i.e. $K_j, j = 1, \dots, J$ where J is the maximum number of sources). It designates a prior distribution on $\mathbf{H}$ and determines the desirable K_j in an unknown basis $\mathbf{D}$ by pruning the irrelevant K_j from $\mathbf{D}$. The term $\mathbf{D}$ with the proper K_j is used for estimating the source which renders the better separation performance than $\mathbf{D}$ without the proper K_j. 4) Incorporating prior information of the basis vectors using the modified multivariate rectified Gaussian. This benefits the overall algorithm in terms of better estimation accuracy and more meaningful feature extraction that pertain to the data. Since each pattern in $\mathbf{Y}$ has its own features, designing the appropriate basis to match these features is imperative.

If these features share some degree of correlation, then this information should be captured to enable better part-based representations of each feature. Toward this end, we develop a modified Gaussian prior distribution on $\mathbf{D}$ to allow the proposed matrix factorization to capture the features of these patterns more efficiently. As our proposed method assigns a regularization parameter to each temporal code (which is individually optimized and adaptively tuned to yield the optimal sparse factorization) this Bayesian regularization improves the accuracy in resolving the spectral bases and the temporal codes which were previously not possible by using cNTF alone. This takes the advantage of the combination of the automatic detection of the optimal K_j through both the pruning technique and the prior information on $\mathbf{D}$. This results in the separation performance that surpasses the conventional cNTF.

The paper is organized as follows. Section 2 introduces the "imitated-stereo" mixture model along with the assumptions of the proposed method. The proposed demixing method and the formulation of the NTF algorithm are presented in Section 3. The separability of the mixture model is presented in Section 4. Experimental source separation results on musical data coupled with a series of performance comparison with other SCBSS techniques using the datasets from Real World Computing (RWC) [13] music database and the 2016 Signal Separation Evaluation Campaign (SiSEC) [39] are presented in Section 5. We finally conclude the paper in Section 6.

2 Single Channel Mixing Model

A. Imitated-Stereo Mixture Model

The single-channel blind source separation problem can be expressed as

$$y_1(t) = x_1(t) + x_2(t) + \cdots + x_{N_s}(t) \tag{1}$$

where $y_1(t)$ is the single channel observed mixture, $x_j(t)$ denotes the jth source signal, N_s, is the total number of source signals and $t = 1,2,\dots,T$ denotes the time index. To discover the original signals $x_j(t)$ given only by the sole observed mixture $y_1(t)$, we compose another mixture based on the autoregressive (AR) process of the sources. Most of audio signals can be modeled by the AR process. This enables us to propose the imitated mixture by time-shifting and weighting the observed mixture as

$$\begin{aligned} y_2(t) &= \tfrac{1}{1+|\beta|}\Big(y_1(t) \ + \ \beta\big(y_1(t-\delta)\big)\Big) \\ &= \tfrac{1}{1+|\beta|}\Big(x_1(t) + x_2(t) + \beta\big(x_1(t-\delta) + x_2(t-\delta)\big)\Big) \end{aligned} \tag{2}$$

where $\beta \in \Re$ is the weight parameter, and δ is the time-delay. The AR process of the signal can be expressed [43] as

$$x_j(t) = -\sum_{z=1}^{M_j} c_{x_j}(z;t)x_j(t-z) + e_j(t) \tag{3}$$

where M_j is the maximum AR order, z is the number of AR order, $c_{x_j}(z;t)$ denotes the zth order AR coefficient of the jth source signal at time t and $e_j(t)$ is an independent identically distributed (i.i.d.) random signal with variance σ^2 and zero mean. We term the mixing model in (2) and (3) as 'imitated -stereo' since the mixing model resembles a stereo signal where the attenuation of the sources differs but an only identical time delay; due to the fact that sources are at one location. By using the AR process in (3), the imitated mixture can be rewritten in terms of the sources, its coefficients and time-delay as

$$\begin{aligned} y_2(t) &= \frac{1}{1+|\beta|}\Big(-\sum_{z=1}^{M_1} c_{x_1}(z;t)x_1(t-z) + e_1(t) + \beta x_1(t-\delta) \\ &\quad -\sum_{z=1}^{M_2} c_{x_2}(z;t)x_2(t-z) + e_2(t) + \beta x_2(t-\delta)\Big) \\ &= \frac{\left(-c_{x_1}(\delta)+\beta\right) x_1(t-\delta)}{1+|\beta|} + \frac{\left(-c_{x_2}(\delta)+\beta\right) x_2(t-\delta)}{1+|\beta|} + \frac{-\sum_{\substack{z=1 \\ z\neq\delta}}^{M_1} c_{x_1}(z;t)x_1(t-z)+e_1(t)}{1+|\beta|} + \\ &\quad \frac{-\sum_{\substack{z=1 \\ z\neq\delta}}^{M_2} c_{x_2}(z;t)x_2(t-z)+e_2(t)}{1+|\beta|} \end{aligned} \tag{4}$$

The proposed mixing model in terms of the sources can now concisely be expressed in time representation as

$$\begin{aligned} y_1(t) &= \sum_{j=1}^{N_s} x_j(t) \\ y_2(t) &= \sum_{j=1}^{N_s} a_j x_j(t-\delta) + r_j(t) \end{aligned} \tag{5}$$

where $a_j(t;\delta,\beta)$ and $r_j(t;\delta,\beta)$ represent the mixing attenuation and residue of the j^{th} source, respectively.

$$a_j(t) \equiv a_j(t;\delta,\beta) = \frac{-c_{x_j}(\delta;t)+\beta}{1+|\beta|} \tag{6}$$

$$r_j(t) \equiv r_j(t;\delta,\beta) = \frac{-\sum_{\substack{z=1 \\ z\neq\delta}}^{M_j} c_{x_j}(z;t)x_j(t-z)+e_j(t)}{1+|\beta|} \tag{7}$$

Note that the parameterization of $a_j(t)$ and $r_j(t)$ depends on δ and β although this is not shown explicitly. For the time-frequency (TF) representation of $y_1(t)$ and $y_2(t)$, the mixing model can be expressed for $\forall(f,t_s)$ as

$$\begin{aligned} Y_1(f,t_s) &= \sum_{j=1}^{N_s} X_j(f,t_s) \\ Y_2(f,t_s) &= \sum_{j=1}^{N_s} \left(a_j(t_s)e^{-i2\pi f\delta}X_j(f,t_s-\delta) - R_j(f,t_s)\right) \end{aligned} \tag{8}$$

In (8), we use the fact that $e_j(t) \ll x_j(t)$, hence the TF of $r_j(t)$ in (7) becomes

$$R_j(f,t_s) \approx -\sum_{\substack{z=1 \\ z\neq\delta}}^{M_j} \frac{c_{x_j}(z;\tau)e^{-i2\pi fz}}{1+|\beta|} X_j(f,t_s-z) \tag{9}$$

From (8), it can be seen that the imitated-stereo mixture comprises of $a_j(t_s)e^{-i2\pi f\delta}$ and $X_j(f,t_s)$. A careful analysis of (5) will reveal that even if $X_j(f,t_s)$ is unknown, the signature of each source can be extracted directly from $Y_1(f,t_s)$ using only information of $a_j(t_s)e^{-i2\pi f\delta}$. Care must be exercised in selecting the time-delay δ in the imitated-stereo (2). The factor $e^{-i2\pi f\delta}$ is only uniquely specified if $|2\pi f\delta| < \pi$, to avoid phase-wrap. In order to avoid phase ambiguity, the time-delay δ must satisfy the following condition

$$|2\pi f_{max}\delta_{max}/f_s| < \pi \tag{10}$$

where δ_{max} is the maximum time delay, f_{max} is the maximum frequency present in the sources and f_s is the sampling frequency. Hence, a term δ_{max} can be determined from (10) according to

$$\delta_{max} < \frac{f_s}{2f_{max}} \tag{11}$$

As long as the delay parameter is less than δ_{max}, there will not be any phase ambiguity. This condition will be used to determine the range of δ in formulating the imitated-stereo mixture.

In the proposed framework, the excitation signal for each source is filtered by a different AR filter. By comparing with the observed mixture $y_1(t)$, the imitated - stereo mixture $y_2(t)$ has extra information of the sources i.e. $a_j(t)$, δ, and $r_j(t)$. This results in a form of temporal correlation diversity of the sources in terms of the AR coefficients. It is noted in (7) and (8) that the second channel ($y_2(t)$ or equivalently $Y_2(f,t_s)$) is a mixture of the original sources and weighted by the source's temporal correlation. Thus, our method in constructing the model enables this diversity to be manifested in the pair of imitated - stereo mixture as noted in $y_1(t)$ and $y_2(t)$. In addition, the residue $r_j(t)$ can be minimized by selecting the appropriate β and δ. As far as the authors are concerned, this is the first-time temporal correlation diversity is proposed for solving the SCBSS problem. The imitated - stereo mixture pave away of the SCBSS problem to enable applying tensor estimation.

Our novelty of the artificial-stereo mixture has been the emergence of a new diversity in the form of sources' temporal correlation within the context of SCBSS. Furthermore, the concept of temporal correlation admits a tensor representation which is then evolved into a statistical estimation problem. This enables us to treat the single-channel recording as multiple channels and subsequently allow us to develop a NTF approach for estimating the sources. The derivations of the artificial-stereo NTF source separation method are presented in Section 3.

3 PROPOSED METHOD

A. Nonnegative Tensor Factorization Separation Model

The proposed method aims to estimate the original signals $\left[x_1(t)x_2(t)\cdots x_{N_s}(t)\right]^T$ by formulating an imitated stereo mixture and using the proposed method given only one observed mixture, $y_1(t)$. The process of the proposed method is illustrated in Fig.1. Based on the linear mixing assumption in (5), the multichannel audio mixtures $y_i(t)$ of unknown sources $x_{ij}(t)$ can be generalized as:

$$y_i(t) = \sum_{j=1}^{N_s} a_{ij}x_{ij}(t) + n_i(t) \quad , \quad \forall i \tag{12}$$

where $i \in \{1,2\}$ denotes the channel number, a_{ij} corresponds to the mixing coefficient, and $n_i(t)$ is the noise. In this work, $n_i(t)$ can also represents the residue from the AR sources, i.e., $n_1(t) = 0$ and $n_2(t) = \sum_{j=1}^{N_s} r_j(t)$ as expressed in (5).

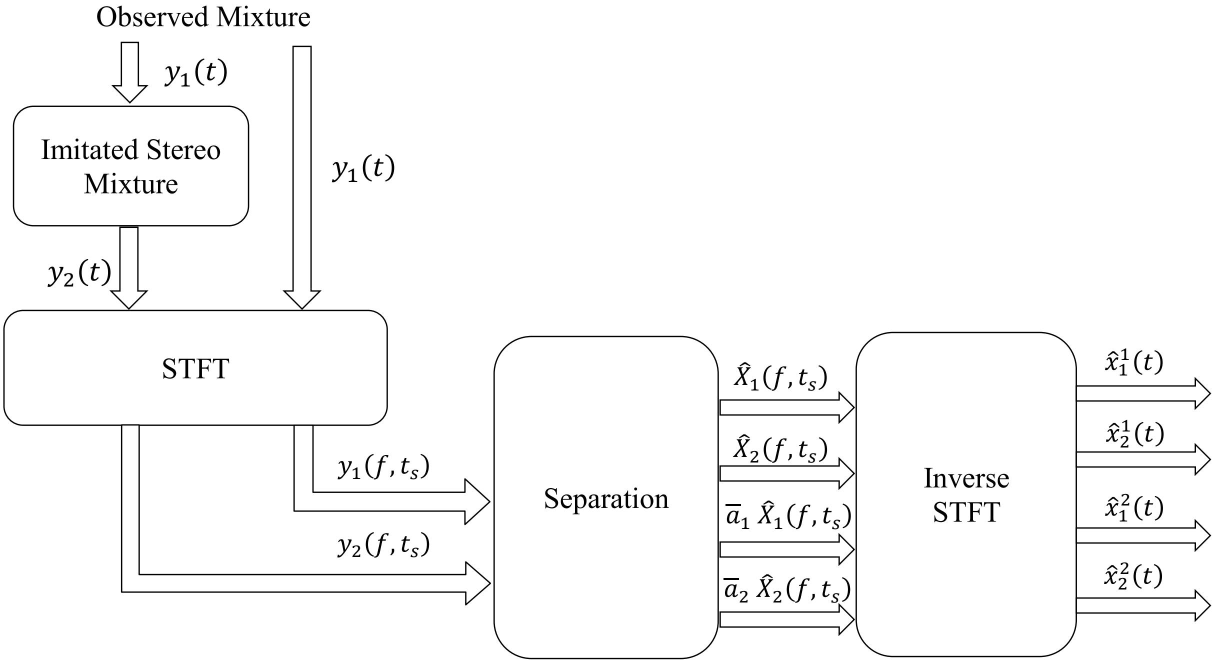


Fig. 1: Overview of the proposed method for $N_s = 2$.

The source signals can be further modeled as a sum of elementary components themselves i.e.

$$x_{ij}(t) = \sum_{k\in K_j} c_{ik}(t) \tag{13}$$

where K_j, $[K_1, \dots, K_{N_s}]$, denotes a nontrivial partition of $[1, \dots, K]$. The components $c_{ik}(t)$ will be characterized by a spectral shape d_k and a vector of activation coefficient h_k through a statistical model. Thus, the observation $y_i(t)$ can be expressed as

$$y_i(t) = \sum_{k=1}^{K} m_{ik} c_{ik}(t) + n_i(t) \tag{14}$$

where m_{ik} is defined as $m_{ik} = a_{ij}$ if and only if $k \in K_j$. The TF representation of the mixture in (14) is given by

$$Y_i(f, t_s) = \sum_{k=1}^{K} m_{ik} C_{ik}(f, t_s) + N_i(f, t_s) \tag{15}$$

where $Y_i(f, t_s)$, $C_{ik}(f, t_s)$ and $N_i(f, t_s)$ denote the TF components of $y_i(t)$, $c_{ik}(t)$, and $n_i(t)$, respectively. The time slots are given by $t_s = 1,2, \dots, T_s$ while frequencies by $f = 1,2, \dots, F$. This power spectrogram is obtained by assuming that the signals and the noise are uncorrelated. The power spectrogram of the residue-free mixture is given by

$$|\bar{Y}_i(f, t_s)|^2 = \begin{cases} |Y_i(f, t_s)|^2 \ , \ i = 1 \\ |Y_i(f, t_s)|^2 - |N_i(f, t_s)|^2 \ , \ i = 2 \end{cases} \tag{16}$$

where $|N_i(f, t_s)|^2$ is estimated using spectral subtraction method [18, 41]. Since each component is a function of t_s and f, we represent this as the 3-valence tensor of mixture STFT $\bar{\mathbf{Y}}_i = \left[\bar{Y}_{i_i}(f, t_s)\right]_{t_s=1,2,\dots,T_s}^{f=1,2,\dots,F}$, of size $I \times F \times T_s$, is modeled as a sum of K_j complex-valued latent tensor components $\mathbf{C}_{ik} = [C_{ik}(f, t_s)]_{t_s=1,2,\dots,T_s}^{f=1,2,\dots,F}$. The time-frequency spectrums of the mixtures are required to be positive values. Assuming $C_{ik}(f, t_s) \sim \mathcal{N}_c(0 | d_{fk} h_{kt_s})$ where $\mathcal{N}_c(\cdot)$ denotes the proper complex Gaussian distribution and $d_{fk} h_{kt_s}$ is the variance [1, 21, 40]. In this case, the power spectrograms $|\bar{\mathbf{Y}}_i|^2$ are approximated by a linear combination of nonnegative spectrograms $|C_{ik}(f, t_s)|^2 \simeq d_{fk} h_{kt_s}$ for each $k \in K_j$ such that

$$\begin{aligned} |\bar{Y}_i(f, t_s)|^2 &= \sum_{k=1}^{K} q_{ik} |C_{ik}(f, t_s)|^2 \\ &= \sum_{j=1}^{N_s} \sum_{k \in K_j} q_{ik} |C_{ik}(f, t_s)|^2 \\ &\simeq \sum_{j=1}^{N_s} \sum_{k \in K_j} q_{ik} d_{fk} h_{kt_s} \end{aligned} \tag{17}$$

where $q_{ik} = |m_{ik}|^2$. Denoting the non-negative matrices are $\mathbf{D} = \{d_{fk}\} = [\mathbf{d}_1 \ \cdots \ \mathbf{d}_K]$, $\mathbf{H} = \{h_{kt_s}\} = [\mathbf{h}_1^{\mathbf{T}} \ \cdots \ \mathbf{h}_K^{\mathbf{T}}]^{\mathbf{T}}$ and $\mathbf{Q} = \{q_{ik}\}$. The problem is to separate the sources $x_{ij}(t)$ given by $|\bar{Y}_i(f, t_s)|^2$ in (17). The proposed method focuses on the estimation of unknown parameters $\mathbf{Q}$, $\mathbf{D}$, and $\mathbf{H}$ of each source. The estimates of $\mathbf{Q}$, $\mathbf{D}$, and $\mathbf{H}$ are used to reconstruct the original sources which are presented in Section B.

B. Formulation of the Proposed Algorithm

The proposed algorithm is firstly formulated i.e. $\boldsymbol{V}$ is the $I \times F \times T_s$ tensor with coefficients $V_i(f,t_s) = |Y_i(f,t_s)|^2$, $\widehat{\boldsymbol{V}}$ is the estimated $I \times F \times T_s$ tensor with coefficients $\widehat{V}_i(f,t_s) = \sum_{k=1}^{K} q_{ik} d_{fk} h_{kt_s}$. The term $\boldsymbol{P} = \left\{\left|a_{ij}\right|^2\right\}$ is the $I \times J$ mixing matrix, $\boldsymbol{L} = \left\{l_{jk}\right\}$ is the $J \times K$ "labelling matrix" with only one nonzero value per column, i.e., such that

$$l_{jk} = \begin{cases} 1, if\ k \in K_j \\ 0, otherwise \end{cases} \tag{18}$$

and nonnegative vector $\boldsymbol{\lambda} = \left\{\lambda_{kt_s}\right\}$. We can express $\boldsymbol{Q}$ as follows:

$$\boldsymbol{Q} = \boldsymbol{PL} = \left\{\left|a_{ij}\right|^2 l_{jk}\right\} \tag{19}$$

Thus, we choose a prior distribution $p(D,H)$ over the factors $\{D,H\}$. It can be shown that the following optimization problem needs to be solved

$$\boldsymbol{min_{Q,D,H}\, C_{MAP}\,(D,H) \doteq -log\, p(D,H|Y,\lambda,Q)} \tag{20}$$

The posterior can be found by using Bayes' theorem as

$$\boldsymbol{p(D,H|Y,\lambda,Q) = \frac{p(Y|D,H,Q)p(D,H|\lambda)}{P(Y)}} \tag{21}$$

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

$$\boldsymbol{log\, p(D,H|Y,\lambda,Q) = log\, p(Y|D,H,Q) + log\, p(D,H|\lambda) + const} \tag{22}$$

Then, log-likelihood of the factor D, H and Q can be written as

$$\begin{aligned} \mathbf{-log\, p(Y|D,H,Q)} &\doteq \mathbf{\textstyle\sum_{ift_s} d_{IS}(V_i(f,t_s)|\widehat{V}_i(f,t_s))} \\ &= \mathbf{\textstyle\sum_{ift_s} \frac{V_i(f,t_s)}{\widehat{V}_i(f,t_s)} - log \frac{V_i(f,t_s)}{\widehat{V}_i(f,t_s)} - 1} \end{aligned} \tag{23}$$

The term c is a constant, the symbol "≐" denotes equality up to constant, and the term $d_{IS}(x|y) = \frac{x}{y} - log\frac{x}{y} - 1$ is the Itakura-Saito divergence. In our proposed model, the prior over $\boldsymbol{D}$ is assumed to be distributed as $N_m(\boldsymbol{D}|0,\boldsymbol{\Sigma}_D)$ i.e. zero-mean modified multivariate Gaussian with covariance matrix $\boldsymbol{\Sigma_D}$ which we will now develop. Since $\boldsymbol{D}$ is nonnegative, using exponential distribution can render poorer quality of sparsity than the modified Gaussian distribution. For a likelihood method based on Gaussian distribution, this is a simple Bayesian criterion for NMF. The Gaussian distribution causes the NMF will yield many locally optimal solutions. Furthermore, it does not suit with the multiplicative update algorithm. In this work, we propose the rectified Gaussian which has previously been shown to provide more flexible shapes of prior distribution [24, 26]. This benefits the distribution model to better suit the signals. The multivariate rectified Gaussian defined as

$$p(\mathbf{D}) = \Phi(-diag^{-1}(\mathbf{\Sigma}_D)\mathbf{u}_D)\delta(\mathbf{d}) + \left(\sqrt{2\pi}|\mathbf{\Sigma}_D|^2\right)^{(-1/2FK)} \exp\left(-\tfrac{1}{2}(\mathbf{d}-\mathbf{u}_D)^{\mathbf{T}}\mathbf{\Sigma}_D^{-1}(\mathbf{d}-\mathbf{u}_D)\right)\mathrm{U}(\mathbf{d}) \tag{24}$$

where $\boldsymbol{d} = vec(\boldsymbol{D}) = [\boldsymbol{d}_1^{\boldsymbol{T}} \vdots \boldsymbol{d}_2^{\boldsymbol{T}} \vdots \cdots \vdots \boldsymbol{d}_K^{\boldsymbol{T}}]^{\boldsymbol{T}}$, $vec(\cdot)$ represents the column vectorization, $\delta(\boldsymbol{d})$ is the delta function, $U(\boldsymbol{d}) = 1, \boldsymbol{d} > 0$ and zero otherwise, and $\Phi(\bullet)$ denotes the multivariate Gaussian cumulative distribution function. Considering the zero mean of the rectified Gaussian distribution (i.e. set $\boldsymbol{u}_D = \mathbf{0}$) on the latent variable would better suit most of the real-world data and can enable the induction of sparse positive factors, as results in

$$p(\mathbf{D}) = \tfrac{1}{2}\delta(\mathbf{d}) + \left(\sqrt{2\pi}|\mathbf{\Sigma}_D|^2\right)^{\frac{-1}{2FK}} \exp\left(-\tfrac{1}{2}(\mathbf{d}-\mathbf{u}_D)^{\mathbf{T}}\mathbf{\Sigma}_D^{-1}(\mathbf{d}-\mathbf{u}_D)\right) \tag{25}$$

where $\boldsymbol{\Sigma}_D = \begin{bmatrix} \boldsymbol{\Sigma}_{1,1} & \cdots & \boldsymbol{\Sigma}_{1,K} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{I,1} & \cdots & \boldsymbol{\Sigma}_{K,K} \end{bmatrix}$ is the covariance matrix of $\boldsymbol{d} = vec(\boldsymbol{D})$ and $\boldsymbol{\Sigma}_{k,n} = E[\boldsymbol{d}_k\boldsymbol{d}_n^{\boldsymbol{T}}]$ is the cross-correlation matrix between the basis vectors $\mathbf{d}_{\mathrm{k}}$ and $\boldsymbol{d}_n$, "$E[\cdot]$" denotes the statistical expectation operator. The covariance matrix $\boldsymbol{\Sigma}_D$ can be partitioned as $\boldsymbol{\Sigma}_D = \boldsymbol{\Sigma}_{diag}^{(D)} + \boldsymbol{\Sigma}_{off}^{(D)}$ where $\boldsymbol{\Sigma}_{diag}^{(D)}$ is the matrix that contains only the diagonal sub-matrices of $\boldsymbol{\Sigma}_D$ whereas $\boldsymbol{\Sigma}_{off}^{(D)}$ contains the off-diagonal sub-matrices. The inverse covariance matrix can be approximated as

$$\begin{aligned} \boldsymbol{\Sigma}_D^{-1} &= \left[\boldsymbol{\Sigma}_{diag}^{(D)} + \boldsymbol{\Sigma}_{off}^{(D)}\right]^{-1} \\ &\cong \left[\boldsymbol{\Sigma}_{diag}^{(D)}\right]^{-1} - \left[\boldsymbol{\Sigma}_{diag}^{(D)}\right]^{-1}\boldsymbol{\Sigma}_{off}^{(D)}\left[\boldsymbol{\Sigma}_{diag}^{(D)}\right]^{-1} \\ &= \boldsymbol{\Omega}_{diag}^{D} - \boldsymbol{\Omega}_{off}^{D} \end{aligned} \tag{26}$$

where $\boldsymbol{\Omega}_{diag}^{D} = \left[\boldsymbol{\Sigma}_{diag}^{(D)}\right]^{-1}$, $\boldsymbol{\Omega}_{off}^{D} = \left[\boldsymbol{\Sigma}_{diag}^{(D)}\right]^{-1}\boldsymbol{\Sigma}_{off}^{(D)}\left[\boldsymbol{\Sigma}_{diag}^{(D)}\right]^{-1}$. The $(k,n)^{\mathrm{th}}$ sub-matrix of $\boldsymbol{\Omega}_{off}^{D}$ is given by

$$\boldsymbol{\Omega}_{off,k,n}^{D} = \boldsymbol{\Sigma}_{k,k}^{-1(D)}\boldsymbol{\Sigma}_{k,n}^{D}\boldsymbol{\Sigma}_{n,n}^{-1(D)} \tag{27}$$

Using above, we may cast (27) into two terms:

$$-\log p(\boldsymbol{D}) \doteq -\log \delta(\boldsymbol{d}) + \tfrac{1}{2}\boldsymbol{d}^{\boldsymbol{T}}\boldsymbol{\Omega}_{diag}^{D}\boldsymbol{d} - \tfrac{1}{2}\boldsymbol{d}^{\boldsymbol{T}}\boldsymbol{\Omega}_{off}^{D}\boldsymbol{d} \tag{28}$$

Analyzing the above, the second term $\boldsymbol{d}^{\boldsymbol{T}}\boldsymbol{\Omega}_{diag}^{D}\boldsymbol{d} = \sum_k \boldsymbol{d}_k^{\boldsymbol{T}}\boldsymbol{P}_k^{-1}\boldsymbol{d}_k$ where $\boldsymbol{P}_k^{-1}$ is a Toeplitz matrix corresponding to the kth diagonal sub-matrix of $\boldsymbol{\Omega}_{diag}^{D}$. Since the source signals are modelled as AR processes, it is natural that $\boldsymbol{P}_k$ assumes the AR autocorrelation matrix of the following form:

$$\boldsymbol{P}_k = \sigma_k^2 \begin{bmatrix} 1 & \rho_k & \cdots & \rho_k^{F-1} \\ \rho_k & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_k \\ \rho_k^{F-1} & \cdots & \rho_k & 1 \end{bmatrix} \tag{29}$$

where ρ_k is the first-order correlation of $\boldsymbol{d}_k$. For the third term, we note that $\boldsymbol{\Omega}_{off,k,n}^{D} \overset{\text{def}}{=} \boldsymbol{\Sigma}_{k,k}^{-1(D)} \boldsymbol{\Sigma}_{k,n}^{D} \boldsymbol{\Sigma}_{n,n}^{-1(D)} = \boldsymbol{P}_k^{-1} \boldsymbol{\Sigma}_{k,n}^{D} \boldsymbol{P}_n^{-1}$. Since the elements in $\boldsymbol{P}_k$ are exponentially decaying, we can make a crude approximation that $\boldsymbol{P}_k^{-1} \boldsymbol{\Sigma}_{k,n}^{D} \boldsymbol{P}_n^{-1} \cong \mu_{kn} \boldsymbol{I}$ where $\mu_{kn} = \sigma_k^{-2} \sigma_n^{-2} c_{k,n}$ and $c_{k,n}$ is the correlation between the k^{th} and n^{th} basis vectors. Thus the term $\boldsymbol{d}^{\boldsymbol{T}} \boldsymbol{\Omega}_{off}^{D} \boldsymbol{d} = \sum_{k,n,(k \neq n)} \mu_{kn} \boldsymbol{d}_k^{\boldsymbol{T}} \boldsymbol{d}_n$ measures the sum of weighted correlation between $\boldsymbol{d}_k$ and $\boldsymbol{d}_n$ for all $k, n, (k \neq n)$. Hence, by including both of these terms, the underlying statistical correlation within and between the basis vectors can be incorporated into the matrix factorization to yield results that reflect on prior information of the AR sources. Therefore, with the factorial model in (28) the desired constraint assumes the following form:

$$f(\boldsymbol{D}) = -\log p(\boldsymbol{D}) \doteq -\sum_k \log \delta(\boldsymbol{d}_k) + \tfrac{1}{2} \sum_k \boldsymbol{d}_k^{\boldsymbol{T}} \boldsymbol{P}_k^{-1} \boldsymbol{d}_k - \tfrac{1}{2} \sum_{k,n,(k \neq n)} \mu_{kn} \boldsymbol{d}_k^{\boldsymbol{T}} \boldsymbol{d}_n \tag{30}$$

The use of multivariate rectified Gaussian prior $p(\boldsymbol{D})$ enables the matrix factorization to leverage on the statistical first order AR correlation between the basis vectors. Once the basis $\boldsymbol{d}_k$ has successfully extracted a particular spectral basis associated with a source signal, subsequent basis vectors $\{\boldsymbol{d}_{j \neq k}\}$ will leverage on $\boldsymbol{d}_k$ to extract other spectral components of the same source. However, care must be exercised in order that the basis vectors do not extract the same spectral component. Thus this necessitates us to monitor the correlation between the basis vectors i.e. μ_{kn}, and as this value gets larger, the more imperative it is to introduce pruning to prevent the basis vectors from extracting the same spectral component. This will be elaborated in Section 3.B.2). In order to turn off excess components thereby optimizing K, we choose a component-wise exponential distribution prior is imposed on $\boldsymbol{H}$, namely,

$$p(\boldsymbol{H}|\boldsymbol{\lambda}) = \prod_k \prod_{t_s} \lambda_{kt_s} exp(-\lambda_{kt_s} h_{kt_s}) \tag{31}$$

The negative log prior on $\boldsymbol{H}$ is defined as

$$\begin{aligned} f(H) &= -\log p(H|\lambda) = -\sum_k \sum_{t_s} \{\log \lambda_{kt_s} - \lambda_{kt_s} h_{kt_s}\} \\ &= -\sum_k \sum_{t_s} \log \lambda_{kt_s} + \sum_k \sum_{t_s} \lambda_{kt_s} h_{kt_s} \end{aligned} \tag{32}$$

By substituting (21), (26) and (31) into (20), the negative log posterior of $\boldsymbol{D}$ and $\boldsymbol{H}$ is given by the following:

$$-\log p(\boldsymbol{D}, \boldsymbol{H}|\boldsymbol{Y}, \boldsymbol{\lambda}, \boldsymbol{Q}) \doteq -\log p(\boldsymbol{Y}|\boldsymbol{D}, \boldsymbol{H}, \boldsymbol{Q}) - \log p(\boldsymbol{D}) - \log p(\boldsymbol{H}|\boldsymbol{\lambda}) \tag{33}$$

From (21), (30) and (32), the above can be written as

$$
\begin{aligned}
L &\doteq \sum_{ift_s} d_{IS}\left(V_i(f,t_s)\middle|\hat{V}_i(f,t_s)\right) + f(\boldsymbol{D}) + f(\boldsymbol{H}) \\
&= \sum\nolimits_{ift_s} \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - log\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - 1 - \sum\nolimits_k log\, \delta(\boldsymbol{d}_k) \\
&+ \frac{1}{2}\sum\nolimits_k \boldsymbol{d}_k^{\boldsymbol{T}} \boldsymbol{P}_k^{-1} \boldsymbol{d}_k - \frac{1}{2}\sum\nolimits_{k,j,(k\neq n)} \mu_{kn} \boldsymbol{d}_k^{\boldsymbol{T}} \boldsymbol{d}_n \\
&- \sum\nolimits_k \sum\nolimits_{t_s} log\, \lambda_{kt_s} + \sum\nolimits_k \sum\nolimits_{t_s} \lambda_{kt_s} h_{kt_s} \qquad (34)
\end{aligned}
$$

The sparsity term $\sum_k \sum_{t_s} \lambda_{kt_s} h_{kt_s}$ forms the L_1-norm regularization to resolve the permutation ambiguity by forcing all structure in $\boldsymbol{H}$ onto $\boldsymbol{D}$. Therefore, the sparseness of the solution in (34) is highly dependent on the regularization parameter λ_{kt_s}.

1). Estimation of the mixing coefficient, basis and code

In this section, we will derive the estimation of $\boldsymbol{D}$, $\boldsymbol{H}$ and $\boldsymbol{P} = \left\{\left|a_{ij}\right|^2\right\}$. The derivative of (34) with respect to $\boldsymbol{D}$ of the proposed model is given by:

$$\frac{\partial L}{\partial d_{fk}} = \sum\nolimits_{it_s} q_{ik} h_{kt_s} d'_{IS}\big(V_i(f,t_s)|\hat{V}_i(f,t_s)\big) + \sum\nolimits_n p_{k,fn} d_{fn} - \sum\nolimits_{n\neq k} \mu_{kn} d_{fn} \qquad (35)$$

where $p_{k,fn}$ is the $(f,n)^{th}$ component of the $\boldsymbol{P}_k^{-1}$ matrix. Similarly, the derivative of (34) with respect to $\boldsymbol{H}$ is given by

$$\frac{\partial L}{\partial h_{kt_s}} = \sum\nolimits_{it_s} q_{ik} d_{fk} d'_{IS}\big(V_i(f,t_s)|\hat{V}_i(f,t_s)\big) + \lambda_{kt_s} \qquad (36)$$

The derivative of (34) with respect to $\boldsymbol{P} = \left\{\left|a_{ij}\right|^2\right\}$ is given by

$$\frac{\partial L}{\partial p_{ij}} = \sum\nolimits_k l_{jk} \sum\nolimits_{f,t_s} d_{fk} h_{kt_s} d'_{IS}\big(V_i(f,t_s)|\hat{V}_i(f,t_s)\big) \qquad (37)$$

We define the term $\boldsymbol{G}$ is $I \times F \times T_s$ tensor with entries $g_{ift_s} = d'_{IS}(V_i(f,t_s)|\hat{V}_i(f,t_s))$, namely

$$d'_{IS}\left(V_i(f,t_s)\middle|\hat{V}_i(f,t_s)\right) = \frac{1}{\hat{V}_i(f,t_s)} - \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2} \qquad (38)$$

We note $\langle\bar{\boldsymbol{A}},\bar{\boldsymbol{B}}\rangle_{k_{\bar{A}},k_{\bar{B}}}$ the contracted product between tensors $\bar{\boldsymbol{A}}$ with size $I_1 \times \ldots \times I_M \times J_1 \times \ldots \times J_N$ and $\bar{\mathbf{B}}$ with size $I_1 \times \ldots \times I_M \times K_1 \times \ldots \times K_P$ and $k_{\bar{\boldsymbol{A}}}$ and $k_{\bar{\boldsymbol{B}}}$ are the sets of mode indices over which the summation take place. The contracted product $\langle\bar{\boldsymbol{A}},\bar{\boldsymbol{B}}\rangle_{\{1,\ldots,M\},\{1,\ldots,M\}}$ is a tensor of size $J_1 \times \ldots \times J_N \times K_1 \times \ldots \times K_P$ given by

$$\langle\bar{\boldsymbol{A}},\bar{\boldsymbol{B}}\rangle_{\{1,\ldots,M\},\{1,\ldots,M\}} = \sum\nolimits_{i_1=1}^{I_1} \cdots \sum\nolimits_{i_M}^{I_M} \bar{a}_{i_1,\ldots,i_M,j_1,\ldots,j_N} \bar{b}_{i_1,\ldots,i_M,k_1,\ldots,k_P} \qquad (39)$$

The contracted tensor product is a form a generalized dot product of two tensors along common modes of same dimensions. Using (39), the multiplicative update (MU) learning rules in matrix notation for $\boldsymbol{D}$, $\boldsymbol{H}$, and $\boldsymbol{P}$ become

$$\boldsymbol{D} \leftarrow \boldsymbol{D} \bullet \frac{\langle \boldsymbol{G}_-, \boldsymbol{Q} \circ \boldsymbol{H} \rangle_{\{1,3\},\{1,2\}}}{\langle \boldsymbol{G}_+, \boldsymbol{Q} \circ \boldsymbol{H} \rangle_{\{1,3\},\{1,2\}} + [\delta'(\boldsymbol{D})./\delta(\boldsymbol{D})] + \boldsymbol{D}\boldsymbol{\Xi}^{\boldsymbol{T}}}$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \bullet \frac{\langle \boldsymbol{G}_-, \boldsymbol{Q} \circ \boldsymbol{D} \rangle_{\{1,2\},\{1,2\}}}{\langle \boldsymbol{G}_+, \boldsymbol{Q} \circ \boldsymbol{D} \rangle_{\{1,2\},\{1,2\}} + \boldsymbol{\lambda}\boldsymbol{1}^{\boldsymbol{T}}}$$

$$\boldsymbol{P} \leftarrow \boldsymbol{P} \bullet \frac{\langle \boldsymbol{G}_-, \boldsymbol{D} \circ \boldsymbol{H} \rangle_{\{2,3\},\{1,2\}}\, \boldsymbol{L}^{\boldsymbol{T}}}{\langle \boldsymbol{G}_+, \boldsymbol{D} \circ \boldsymbol{H} \rangle_{\{2,3\},\{1,2\}}\, \boldsymbol{L}^{\boldsymbol{T}}} \tag{40}$$

which has a strikingly similar form with the conventional NMF update rules [16, 38]. In (40), '•' is element-wise product and $\boldsymbol{\Xi}^{\boldsymbol{T}}$ is a $K \times K$ matrix whose $(k,n)^{th}$ element is given by $P_{k,fn}$ and μ_{kn}. Here $\boldsymbol{G}_-$ follows the MU rule that denotes the negative part of the derivative of the criterion e.g. $\boldsymbol{G}_- = \left[d'_{IS}\left(V_i(f,t_s)\middle|\hat{V}_i(f,t_s)\right)\right]_- = \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}$ and $\boldsymbol{G}_+$ denotes its positive part. The term $\boldsymbol{Q} \circ \boldsymbol{H}$ denotes $I \times K \times T_s$ tensor with elements $q_{ik}h_{kt_s}$. Similarly, $\boldsymbol{Q} \circ \boldsymbol{D}$ denotes $F \times I \times K$ tensor with elements $q_{ik}d_{fk}$ and $\boldsymbol{D} \circ \boldsymbol{H}$ denotes $F \times K \times T_s$ tensor with elements $d_{fk}h_{kt_s}$.

2). Estimation of the Adaptive Sparsity Parameter

The update of λ_k follows from solving $\frac{\partial L}{\partial \lambda_{kt_s}} = 0$ which leads to $\lambda_{kt_s} = h_{kt_s}^{-1}$. However, this may cause abrupt changes in the level of sparsity. An adaptive first-order implementation that smooth over time can be obtained as follows:

$$\lambda_{kt_s}(t) = \alpha\lambda_{kt_s}(t-1) + (1-\alpha)\frac{1}{h_{kt_s}+\epsilon} \tag{41}$$

where α is the smoothing parameter and is normally set to 0.95, and $\epsilon = 10^{-9}$ is a small number to prevent division by zero. As mentioned in Section III B, pruning is exercised to prevent the basis vectors from extracting the same spectral component. First, note that the sparsity term $\sum_k \sum_{t_s} \lambda_{kt_s} h_{kt_s}$ forms the sparse NTF which aims to learn the degree of regularization from data, i.e. tune the pruning parameter, λ_{kt_s}. Second, let us assume that the factorization in (16) has an approximation error of $\sum_{i=1}^{I}\sum_{f=1}^{F}\sum_{t_s=1}^{T_s}|Y_i(f,t_s)|^2/IFT_s$. As a result of inference in (34), a subset of the λ_{kt_s} will be driven to a large upper bound, with the corresponding columns of $\boldsymbol{D}$ and rows of $\boldsymbol{H}$ driven to small values. The effective dimensionality can be deduced from the distribution of the λ_{kt_s}. We have found in practice, two clusters clearly emerge: A group of values in same order of magnitude corresponding to relevant components on columns of $\boldsymbol{D}$ and rows of $\boldsymbol{H}$, and a group of similar values of much higher magnitude corresponding to irrelevant components. Furthermore, for components which had become to zero or close to zero we set $\lambda_{kt_s} = \frac{1}{\epsilon}$. Thus, based on the above empirical observation, we propose the following pruning threshold: Let $\bar{\lambda}_k \triangleq \frac{1}{T_s}\sum_{t_s=1}^{T_s}\lambda_{kt_s}$ be the average sparseness value associated with the kth row of $\boldsymbol{H}$. If

$$\bar{\lambda}_k \ > \ \epsilon^{-1}\cdot\sqrt{\frac{\sum_{i=1}^{I}\sum_{f=1}^{F}\sum_{t_s=1}^{T_s}|Y_i(f,t_s)|^2}{IFT_s}} \tag{42}$$

then the k^{th} row of $\boldsymbol{H}$ (equivalently kth column of $\boldsymbol{D}$) is to be removed. This method allows us to estimate the effective number of components. If the prior assumptions are slightly violated or even if the likelihood function differs from the model assumption, the correct factorization rank can be determined by evaluating the above bound by the pruning threshold.

3). Estimation of source signals

For the proposed method, we obtain the estimates of $\boldsymbol{D}$, $\boldsymbol{H}$ and $\boldsymbol{P}$ that yield the smallest cost value. To reconstruct the source signals, the term $\hat{C}_{ik}(f,t_s)$ of the component k in channel i is reformulated by using the Wiener filtering as

$$\begin{aligned}\hat{C}_{ik}(f,t_s) &\stackrel{\text{def}}{=} E\{C_{ik}(f,t_s)|\boldsymbol{P},\boldsymbol{D},\boldsymbol{H},\boldsymbol{Y}\}\\ &= \frac{q_{ik}d_{fk}h_{kt_s}}{\hat{V}_i(f,t_s)}Y_i(f,t_s)\end{aligned} \tag{43}$$

where $\hat{V}_i(f,t_s) = \sum_{k=1}^{K} q_{ik}d_{fk}h_{kt_s}$. The decomposition is conservative in the sense that it satisfies $y_i(t_s) = \sum_{k=1}^{K}\hat{c}_{ik}(t_s)$. The estimated sources are converted back into time-domain by using inverse-STFT of $\hat{C}_{ik}(f,t_s)$ for all i and k. Finally, the estimated sources can be obtained as

$$\hat{x}_{ij}(t) = \sum_{k\in K_j}\hat{c}_{ik}(t) \tag{44}$$

The proposed algorithm is summarized in Algorithm 1.

Algorithm 1: Overview proposed algorithm

1. Generate the mixture $y_2(t)$ from (2) and compute the STFT of $Y_1(f,t_s)$ and $Y_2(f,t_s)$.
2. Apply spectral subtraction on $|Y_i(f,t_s)|^2$, $i = 2,\ldots,N_s$.
3. Initialize $\boldsymbol{D}$, $\boldsymbol{H}$ and $\boldsymbol{P}$ with nonnegative random values and define $\boldsymbol{L} = \{l_{jk}\}$, $l_{jk} = \begin{cases}1, if\ k\in K_j\\ 0, otherwise\end{cases}$ and $\boldsymbol{Q} = \boldsymbol{PL}$.
4. Compute the estimation statistics:
    - Observed tensor: $V_i(f,t_s) = |Y_i(f,t_s)|^2$,
    - Estimated tensor: $\hat{V}_i(f,t_s) = \sum_{k=1}^{K} q_{ik}d_{fk}h_{kt_s}$

- Positive and negative criterion of the multiplicative update (MU) learning rules: $\boldsymbol{G}_{+}$ and $\boldsymbol{G}_{-}$ according to (40).
- First-order correlation of the basis and the correlation between the basis vectors: ρ_k and μ_{kn}.

5. Update model parameters:

$$\boldsymbol{D} \leftarrow \boldsymbol{D} \bullet \frac{\langle \boldsymbol{G}_{-}, \boldsymbol{Q} \circ \boldsymbol{H} \rangle_{\{1,3\},\{1,2\}}}{\langle \boldsymbol{G}_{+}, \boldsymbol{Q} \circ \boldsymbol{H} \rangle_{\{1,3\},\{1,2\}} + [\delta'(\boldsymbol{D})./\delta(\boldsymbol{D})] + \boldsymbol{D}\boldsymbol{\Xi}^{\boldsymbol{T}}}$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \bullet \frac{\langle \boldsymbol{G}_{-}, \boldsymbol{Q} \circ \boldsymbol{D} \rangle_{\{1,2\},\{1,2\}}}{\langle \boldsymbol{G}_{+}, \boldsymbol{Q} \circ \boldsymbol{D} \rangle_{\{1,2\},\{1,2\}} + \boldsymbol{\lambda}\boldsymbol{1}^{\boldsymbol{T}}}$$

$$\boldsymbol{P} \leftarrow \boldsymbol{P} \bullet \frac{\langle \boldsymbol{G}_{-}, \boldsymbol{D} \circ \boldsymbol{H} \rangle_{\{2,3\},\{1,2\}} \boldsymbol{L}^{\boldsymbol{T}}}{\langle \boldsymbol{G}_{+}, \boldsymbol{D} \circ \boldsymbol{H} \rangle_{\{2,3\},\{1,2\}} \boldsymbol{L}^{\boldsymbol{T}}}$$

$$\boldsymbol{\lambda} = \alpha\boldsymbol{\lambda} + (\boldsymbol{1} - \alpha)\frac{\boldsymbol{1}}{\boldsymbol{H}+\epsilon} \ , \ \ \bar{\lambda}_k = \frac{1}{T_s}\boldsymbol{\delta}_k^{\boldsymbol{T}}\,\boldsymbol{\lambda}\,\boldsymbol{1}$$

6. Prune the irrelevant components of $\boldsymbol{D}$ and $\boldsymbol{H}$ using the criteria (42). Normalize $\boldsymbol{D}$ and $\boldsymbol{P}$.

$$\bar{\lambda}_k \ > \ \epsilon^{-1} \cdot \sqrt{\frac{\sum_{i=1}^{I}\sum_{f=1}^{F}\sum_{t_s=1}^{T_s}|Y_i(f,t_s)|^2}{IFT_s}}$$

Repeat Step 5 and 6 until termination (convergence, the max number of iteration)

7. Compute

$$\hat{C}_{ik}(f,t_s) = \frac{q_{ik}d_{fk}h_{kt_s}}{\hat{V}_i(f,t_s)}Y_i(f,t_s)$$

8. Transform $\hat{C}_{ik}(f,t_s)$ to the time domain $\hat{c}_{ik}(t)$ and reconstruct the sources using $\hat{x}_{ij}(t) = \sum_{k\in K_j}\hat{c}_{ik}(t)$.

4 Separability of Imitated-Stereo Mixture Model

In this section, the imitated mixture is examined the separability of the proposed method by considering $a_j(t)$ and $r_j(t)$. To achieve this, we assumed that the sources satisfy the W-disjoint orthogonality (WDO) [9] condition:

$$X_i(f,t_s)X_j(f,t_s) \approx 0, \qquad \forall i \neq j\ ,\ \forall f,t_s \tag{45}$$

The imitated-stereo mixtures of different cases based on $a_j(t)$ and $r_j(t)$ are evaluated by the selected minimum function L_i. Motivated by the separation step of the proposed algorithm, the minimum-selecting function is derived from the estimated signals in TF domain. This can be expressed by assuming that the jth source dominates at a particular TF unit as

$$\hat{X}_{ij}(f,t_s) = \left(\sum_{k\in K_l} q_{ik}d_{fk}h_{kt_s} \,/\, \sum_l q_{ik}\sum_{k\in K_l} d_{fk}h_{kt_s}\right)Y_i(f,t_s)$$

$$= \frac{E\left[Y_i(f,t_s)X_{il}^*(f,t_s)\right]}{E[|Y_i(f,t_s)|^2]}\sum_{j=1}^{N_s} m_{ij}X_{ij}(f,t_s)$$

$$= \frac{\sum_l m_{il}\sum_j m_{ij}E[|X_{il}(f,t_s)|^2]X_{ij}(f,t_s)}{\sum_l |m_{il}|^2 E[|X_{il}(f,t_s)|^2]} \tag{46}$$

If $j = l$, we then obtain

$$\hat{X}_{ij}(f,t_s) = X_{ij}(f,t_s)$$

In this light, we formulate the proposed minimum-selecting function which can be expressed as:

$$L_i = \min_l \left| X_{ij}(f,t_s) - \frac{\sum_{k\in K_l} q_{1k}d_{fk}h_{kt_s}}{\sum_{l=1}^{N_s}\sum_{k\in K_l} q_{1k}d_{fk}h_{kt_s}} Y_i(f,t_s)\right|^2$$

$$= \min_l \left|X_{ij}(f,t_s) - \hat{X}_{il}(f,t_s)\right|^2 \tag{47}$$

By evaluating the minimum-selecting function, each TF unit is marked to the l^{th} argument that yields the minimum value. Hence, the TF units of the mixture are classified into l groups of (f,t_s) units. The minimum-selected function is further analyzed in the cases of the i^{th} mixture. In the first case where $i = 1$ i.e. $Y_1(f,t_s) \ = \sum_{j=1}^{N_s} X_j(f,t_s)$, the function L_1 can be expressed as

$$L_1 = \min_l \left| X_{1j}(f,t_s) - \frac{\sum_{k\in K_l} q_{1k}d_{fk}h_{kt_s}}{\sum_{l=1}^{N_s}\sum_{k\in K_l} q_{1k}d_{fk}h_{kt_s}} Y_1(f,t_s)\right|^2$$

$$= \min_l \left| X_{1j}(f,t_s) - \frac{\sum_{k\in K_l} q_{1k}d_{fk}h_{kt_s}}{\sum_{l=1}^{N_s}\sum_{k\in K_l} q_{1k}d_{fk}h_{kt_s}} \sum_{l=1}^{2} X_l\,(f,t_s)\right|^2 \tag{48}$$

Secondly, when $i = 2$ i.e. $Y_2(f,t_s) = \sum_{j=1}^{N_s} \overline{a}_j X_j(f,t_s)$ the function L_2 can be expressed as

$$L_2 \ = \min_l \left| \overline{a}_j X_{2j}(f,t_s) - \frac{\sum_{k\in K_l} q_{2k}d_{fk}h_{kt_s}}{\sum_{l=1}^{N_s}\sum_{k\in K_l} q_{2k}d_{fk}h_{kt_s}} Y_2(f,t_s)\right|^2$$

$$= \min_l \left| \overline{a}_j X_{2j}(f,t_s) - \frac{\sum_{k\in K_l} q_{2k}d_{fk}h_{kt_s}}{\sum_{l=1}^{N_s}\sum_{k\in K_l} q_{2k}d_{fk}h_{kt_s}} \sum_{l=1}^{2} \overline{a}_l X_l(f,t_s)\right|^2 \tag{49}$$

The functions L_1 and L_2 will then be used for evaluating the separability of the proposed imitated-stereo mixture by considering $a_j(t)$ and $r_j(t)$ in the following three scenarios.

I: If $\forall j\ a_j(t) = a(t)$ and $r_j(t) = r(t)$, then $x_2(t) = \left(\frac{a(t)+\beta}{1+|\beta|}\right)x_1(t-\delta) + 2r(t)$.

The first scenario presents a situation where two identical sources are mixed in the single channel. By a weighted and time-shifting of the observed mixture, the imitated mixture is only obtained the time-delayed and scalar of the first mixture. This results no advantage of the imitated mixture at all. The separability of this case is presented by substituting the imitated-stereo mixture of Scenario I into the functions L_1 and L_2. Since both sources are identical, the minimum-selecting function of each mixture can be evaluated as follow: For $i = 1, X_j(f, t_s) = X(f, t_s)\ \forall j$, the L_1 function then becomes

$$L_1 = \min_l \left| X(f, t_s) - \frac{\sum_{\forall k} q_{1k} d_{fk} h_{kt_s}}{2 \sum_{\forall k} q_{1k} d_{fk} h_{kt_s}} 2X(f, t_s) \right|^2$$
$$= \min_l |X(f, t_s) - X(f, t_s)|^2$$
$$= 0 \text{ for } \forall l \tag{50}$$

For $i = 2$, $a_j(t)$ and $r_j(t)$ are related to the source via $\overline{a}_j$, thus $\overline{a}_j X_j(f, t_s) = \overline{a}\, X(f, t_s)\ \forall j$. Thus the L_2 function becomes:

$$L_2 = \min_l \left| \overline{a}\, X(f, t_s) - \frac{\sum_{\forall k} q_{1k} d_{fk} h_{kt_s}}{2 \sum_{\forall k} q_{1k} d_{fk} h_{kt_s}} 2\overline{a}\, X(f, t_s) \right|^2$$
$$= \min_l |\overline{a}\, X(f, t_s) - \ \overline{a}\, X(f, t_s)|^2$$
$$= 0 \text{ for } \forall l \tag{51}$$

As a result, the both minimum-selecting function are zero for all l^{th} arguments i.e. $L_1 = L_2 = 0$. In this case, the function cannot discriminate the l^{th} arguments, the mixture is not separable.

<u>**II**</u>: If $\forall j$: $a_j(t) = a(t)$ and $r_j(t) \neq r_k(t)$ for $j \neq k$ then $x_2(t) = \left(\frac{a(t)+\beta}{1+|\beta|}\right) x_1(t - \delta) + r_1(t) + r_2(t)$.

Scenario II represents different sources but setting β and δ for the imitated-stereo mixture such that $a_1(t) = \cdots = a_{N_s}(t)$. By following the steps in Case 1, the separability of this mixture can be analyzed using the functions L_1 and L_2 as

$$L_1 = \min_l \left| X_{1j}(f, t_s) - \frac{\sum_{k \in K_l} q_{1k} d_{fk} h_{kt_s}}{\sum_{l=1}^{N_s} \sum_{k \in K_l} q_{1k} d_{fk} h_{kt_s}} \sum_{l=1}^{N_s} X_{\ l}\ (f, t_s) \right|^2$$
$$= \min_l \left| X_{1j}(f, t_s) - X_l(f, t_s) \right|^2 \tag{52}$$

Since $r_j(t) \neq r_k(t)$ thus $\overline{a}_j X_j(f, t_s) \neq \overline{a}_k X_k(f, t_s)$ for $j \neq k$, we then obtain

$$L_2 = min_l \left| \overline{a}_j X_{2j}(f, t_s) - \frac{\sum_{k \in K_l} q_{2k} d_{fk} h_{kt_s}}{\sum_{l=1}^{N_s} \sum_{k \in K_l} q_{2k} d_{fk} h_{kt_s}} \sum_{l=1}^{N_s} \overline{a}_l X_l(f, t_s) \right|^2$$
$$= min_l \left| \overline{a}_j X_{2j}(f, t_s) - \ \overline{a}_l X_l(f, t_s) \right|^2 \tag{53}$$

As a result of $j = l$, the both L_1 and L_2 functions yields a zero value. The minimum-selecting functions are capable to separate the lth arguments although the sources have the same mixing attenuation; $a_1(t) = \cdots = a_{N_s}(t) = a(t)$. Therefore, the mixture of Scenario II is separable.

III: If $a_j(t) \neq a_k(t)$ and $r_j(t) \neq r_k(t)$ for $j \neq k$ then

$$x_2(t) = \sum_{j=1}^{N_s} \left(\frac{a_j(\delta)+\beta}{1+|\beta|}\right) x_j(t-\delta) + r_j(t)$$

This scenario corresponds to the most general case where the sources are distinct, and β and δ are determined arbitrarily such that the mixing attenuations and residues are also different. The L_1 function is firstly treated where the original signals differ i.e. $X_j(f,t_s) \neq X_k(f,t_s)$. Hence, the L_1 function of Scenario III provides $L_1 = min_l \left|X_{1j}(f,t_s) - X_l(f,t_s)\right|^2$ which is the same as Scenario II. Since the mixing attenuations $a_j(t_s)$ and $a_k(t_s)$ correspond respectively to $x_j(t)$ and $x_k(t)$, thus $\overline{a}_j X_j(f,t_s) \neq \overline{a}_k X_k(f,t_s)$ and $r_j(t) \neq r_k(t)$. By following similar line of the L_2 function in Scenario II, we then have

$$L_2 = \min_l \left|\overline{a}_j X_{2j}(f,t_s) - \overline{a}_l X_l(f,t_s)\right|^2 \tag{54}$$

For $j \neq l$, the L_1 and L_2 functions in Scenario III render a non-zero value. Hence, this mixture can be separated by the minimum-selecting function.

5 Results and Analysis

A. Experiment Setup

The proposed method is demonstrated by separating real-audio sources. The real-audio sources which are inherently non-stationary include vocal and music signals. All experiments are conducted using a PC with Intel® Core™ i7-6700 CPU@3.4GHz, 8GB RAM. and 4 GB RAM. MATLAB is used as the programming platform. The TF representation is computed by using the STFT of 1024-point Hanning window with 50% overlap. The experiments consist of 7 type of mixtures are generated i.e. male speech + female speech, male speech + jazz, male speech + drum, male speech + piano, jazz + drum, jazz + drum, and drum + piano. The male speech, female speech and music sources are selected from the RWC database and 3 linear instantaneous stereo mixtures of 3 sources taken from the Signal Separation Evaluation Campaign (SiSEC 2016) “Underdetermined speech and music mixtures” task development dataset [39]. Three audio datasets have been considered and are described as: 1) wdrums, a linear instantaneous stereo mixture (with positive mixing coefficients) of 2 drum sources and 1 bass line. 2) nodrums, a linear instantaneous stereo mixture (with positive mixing coefficients) of 1 rhythmic acoustic guitar, 1 electric lead guitar and 1 bass line. Both mixtures are 10 seconds-long and sampled at 16 kHz. The instantaneous mixing is characterized by static positive gains. We applied a STFT with sine bell of length 64 ms (1024 samples) leading to $F = 513$. 3) Shannonsongs Sunrise, a linear instantaneous stereo mixture of $d_{max} = 3$ musical sources (drums, lead vocals and piano) created using 3.12 seconds-excerpts of original separated tracks from the song "Sunrise" by S. Hurley and down sampled to 16 kHz. MATLAB routines for computing these criteria are obtained from the SiSEC'08 webpage [39]. The proposed method will be compared with 1) the other SCBSS method as the sparse nonnegative matrix 2-dimensional factorization (SNMF2D) [17] and the single-channel independent component analysis (SCICA) [32]. The SNMF2D parameters are set as follows [10]: number of factors is 2, sparsity weight of 1.1, number of phase shift and time shift is 31 and 7, respectively for music. As for speech, both shifts are set to 4. The TF domain used in SNMF2D is based on the log-frequency spectrogram. Cost function of SNMF2D is based on the Kullback-Leibler divergence. As for the SCICA, the number of block is 10 with time delay set to unity. We have evaluated our separation performance by measuring the distortion between original source and the estimated one according to the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and source-to-artifacts ratio (SAR) i.e. $SDR = 10\ log_{10}\left(\|s_{target}\|^2/\|e_{interf}+e_{noise}+e_{artif}\|^2\right)$, $SIR = 10\ log_{10}\left(\|s_{target}\|^2/\|e_{interf}\|^2\right)$, and $SAR = 10\ log_{10}\left(\|s_{target}+e_{interf}+e_{noise}\|^2/\|e_{artif}\|^2\right)$ where e_{interf}, e_{noise}, and e_{artif} represent the interference from other sources, noise and artifact signals.

B. Impact of weight (β) and time-delay (δ) parameters on matrix factorization and source separation

The imitated stereo mixture is formulated via determining the weight β and the time-delay δ parameters. The weight β parameter acts as a controlling factor to maintain the difference of the sources' AR coefficients and to control the amount of the residues $r_j(t;\delta,\beta)$. The impact of determination of values for β and δ parameters will be investigated on the type of sources in this section. A set of experiments has been conducted to determine the β and δ pairs by using wdrums, nodrums and Shannonsongs Sunrise mixtures. A finite range of 16 pairs of β and δ is selected to be [-4, 4] (excluding $\beta = 0$) and [1, 2] as: $\tau = \begin{Bmatrix}(-1,1),(-2,1),(-3,1),(-4,1),(1,1),(2,1),(3,1),(4,1),\\(-1,2),(-2,2),(-3,2),(-4,2),(1,2),(2,2),(3,2),(4,2)\end{Bmatrix}$. The reason is in the extreme case of $\beta = 0$, which leads to $y_1(t) = y_2(t)$ where the imitated stereo mixture cannot be formulated. In practice, the AR coefficients of sources are generally unknown. However, if one knows the source category then β and δ can be chosen from τ. Hence, this enables the algorithm to estimate β and δ for the specific type of sources.

Fig. 2 shows the separation results in terms of the SDR for the mixtures of wdrums, nodrums and Shannonsongs Sunrise. As the results, it can be seen that when $(\beta,\delta) = (1,1)$ will yield the best possible SDR overlaps with all the three categories at 13.63 dB, 7.85 dB, and 6.46 dB, respectively. This is not surprising since speech and music are mainly characterized by the initial few AR coefficients and these coefficients tend to vary for different sources. For each type of mixtures with 5% from highest SDR, the recommended pairs of β and δ ranges are {(1,-2), (2,-2)} for wdrums mixture, (2,-1) for nodrums mixture, and {(2,-2), (2,-3)} for Shannonsongs Sunrise mixture. The results indicate that only the low order AR coefficients i.e. $\delta = 1$, are beneficial for separation.

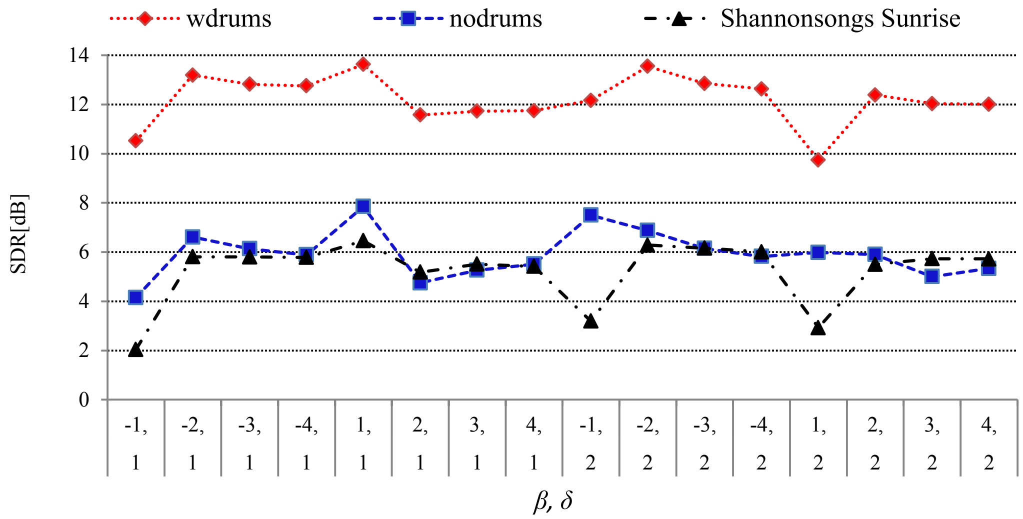
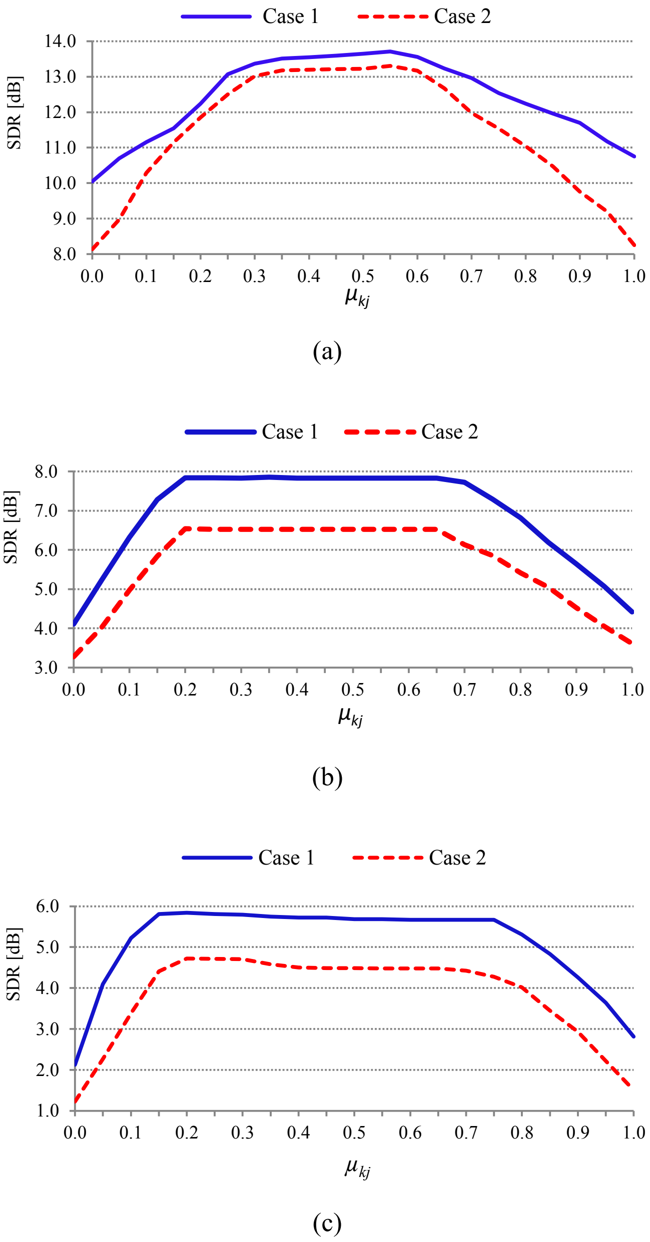


Fig. 2: Separation results of the proposed method by using different weight (β) and time-delay (δ) parameters.

C. Impact of μ_{kj} on separation performance

In this section, the impact of μ_{kj} will be investigated. In practice, the actual statistics for computing the prior on $\boldsymbol{D}$ (μ_{kj}) given in (33) is unknown. In this case, the selection of μ_{kj}will depend on the type of sources and require estimation. Hence, we investigate the effects of μ_{kj} in conjunction with the *pruning* method on the separation performance. Firstly, we estimate $\hat{\mu}_{kn} = \hat{\sigma}_k^{-2}\hat{\sigma}_n^{-2}\hat{c}_{k,n}$ using $\hat{c}_{k,n} = \boldsymbol{d}_k^{\boldsymbol{T}}\boldsymbol{d}_n$ and $\hat{\sigma}_j^{-2} = 1/\|\boldsymbol{d}_j\|^{\boldsymbol{2}}$ for $j = k, n$. We then compare the estimated $\hat{\mu}_{kn}$ with manual setting. The following two cases are considered: Case 1) with pruning and μ_{kj} is varied from 0, 0.05, 0.1,…, 1.0 Case 2) without pruning and μ_{kj} is varied from 0, 0.05, 0.1,…, 1.0. The wdrums, nodrums and Shannonsongs Sunrise datasets have been used for the above cases.



(a)

(b)

(c)

Fig.3: SDR results as a function of μ_{kj}.(a) wdrums. (b) nodrums. (c) Shannonsongs Sunrise.

Fig. 3 shows that the separation result with the pruning method yielded a total average improvement of 1.17dB over the separation method without pruning. The average SDR improvement can be summarized as follows: 1) 0.96 dB per source for wdrums mixture; 2) 1.26 dB per source for nodrums mixture; and 3) 1.29 dB per source for Shannonsongs Sunrise mixture. The results have also clearly indicated that the best performance of wdrums mixture is obtained when μ_{kj} ranges from 0.33 to 0.64 (within 2% from highest SDR) with the highest average SDR is 13.71 dB. The $\hat{\mu}_{kn}$ rendered from data estimation is 0.47 which very closely approaches the optimum SDR, which is at $\mu_{kj} = 0.52$. As for nodrums mixture, the best performance is obtained when μ_{kj} ranges from 0.17 to 0.72 with the highest average SDR is 7.85 dB where the $\hat{\mu}_{kn}$ is 0.39 and the optimum SDR is at $\mu_{kj} = 0.34$. In the case of Shannonsongs Sunrise, the best range of μ_{kj} is from 0.13 to 0.46 which yields the best performance with the highest average SDR of 5.84 dB where the $\hat{\mu}_{kn}$ is 0.23 and the optimum SDR is at $\mu_{kj} = 0.21$. From the above findings, we can conclude that for music mixtures, the best performance is obtained when μ_{kj} ranges from 0.17 to 0.72 and in the case of music and vocal mixture, the best performance is obtained when μ_{kj} ranges from 0.13 to 0.46. On the contrary, it is noted that when μ_{kj} is set either too low or high, the separation performance tends to degrade. It is also worth pointing out that the separation results are rather coarse when the factorization is non-regularized (i.e., without prior on $\boldsymbol{D}$*)* and without pruning. Here, we can see that the average SDR of without prior on $\boldsymbol{D}$ and without pruning is the lowest among the three methods across $\mu_{kj} > 0$.

By incorporating regularization (i.e., using $\mu_{kj} > 0$ and *pruning*), the performance increases significantly for all types of mixture. This is clearly evident in Figs.3 (a) - (c) where the average SDR result for separation three mixtures scales up to 9.1 dB while for the case of without regularization the average SDR result is only 7.6 dB. This amounts to a significant 1.5 dB performance improvement using the proposed regularization than that without regularization. Thanks to the modified Gaussian prior, this correlation is explicitly modeled by μ_{kj} in the proposed method. This enables the estimated basis vectors $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ to take advantage of the correlation in learning the real basis directly from the mixed pattern. This explains the reason as to why that the proposed method with pruning and with prior on $\boldsymbol{D}$ shows better performance than the proposed method with pruning and without prior on $\boldsymbol{D}$. Therefore, the analysis have unanimously indicated the importance of selecting the correct number of components and of incorporating the correlation μ_{kj} between the different basis vectors in order to arrive at the optimal performance of feature extraction.

D. Comparison Proposed Method with Other SCBSS Methods

In this section, Audio sources can be characterized as non-stationary AR processes since their AR coefficients vary with time. We have generated the mixtures of two sources which select from male speech, female speech, jazz, piano and drum. Both sources are mixed with equal power to generate the mixture. Examples of original signals, the mixture and the separated signals are respectively shown in Fig. 4. Visually, the estimated sources resemble closely to the original sources.

1) Single Channel Sources

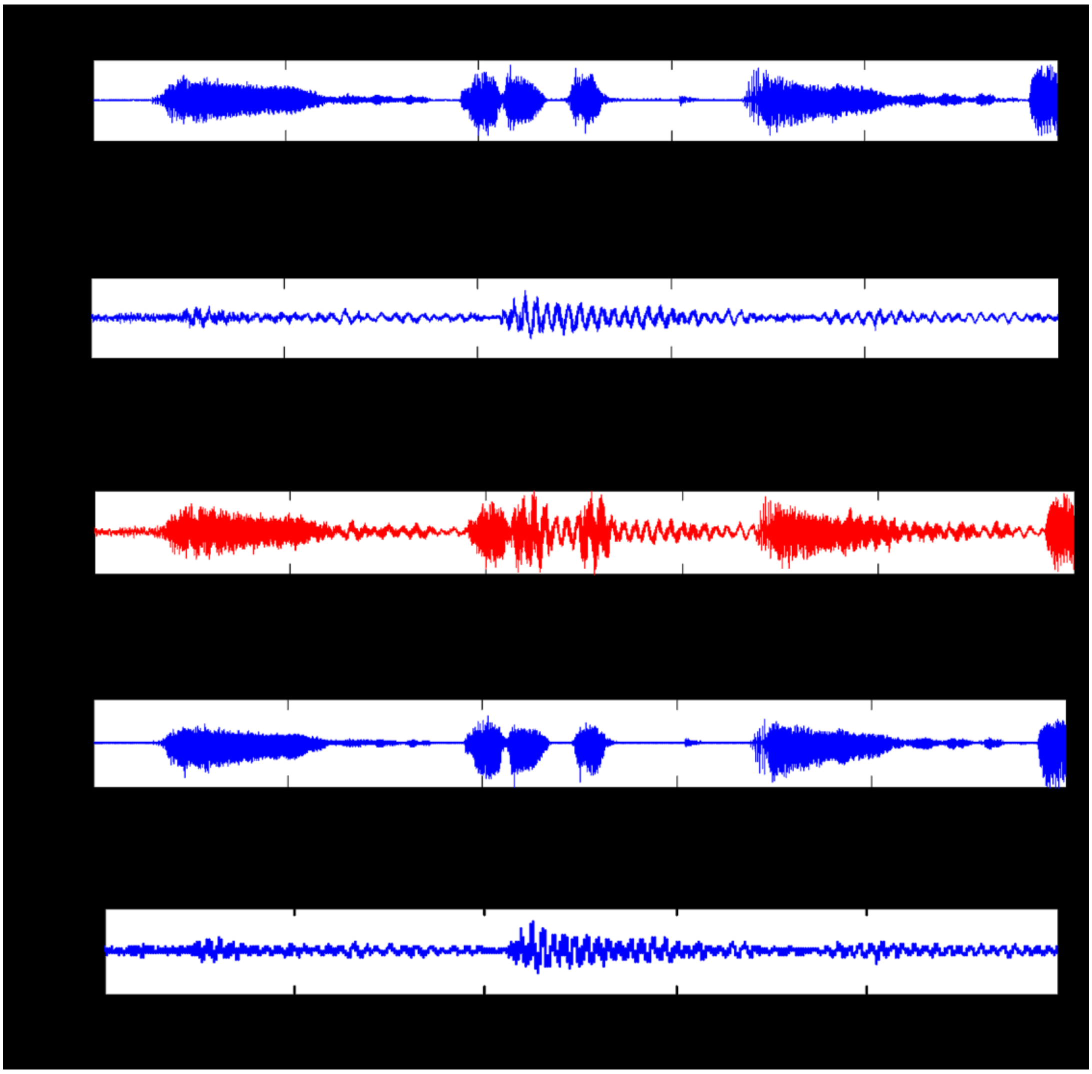


Fig.4: Original sources, single channel mixture, and estimated sources of music mixture between jazz and drum using the proposed method with $\beta = 1$ and $\delta = 1$.

The separation performance based on 3 mixture types of the proposed method was compared with the state-of-the-art of the SCBSS methods: i.e. Hilbert-SD, SCICA, EMD-ICA, SNMF2D, the imitated-stereo mixture using the degenerate unmixing estimation technique (DUET) method [45] that presented in Fig. 5. The proposed method yields the outstanding performance over the comparison methods with the total average SDR improvement at 6.62 dB per source, 6.12 dB per source, and 2.86 dB per source for music mixtures, speech and music mixtures, and speech mixtures, respectively. The reason is based on the optimal part-based factorization of the proposed method. The factorization is unique under certain conditions (e.g., adaptive sparse and nonnegative component), making it unnecessary to impose constrains in the form of statistical independence between original sources. Furthermore, the proposed method can automatically detect the optimal number of components of the individual source, thus leading to more robust separation results among the comparison methods.

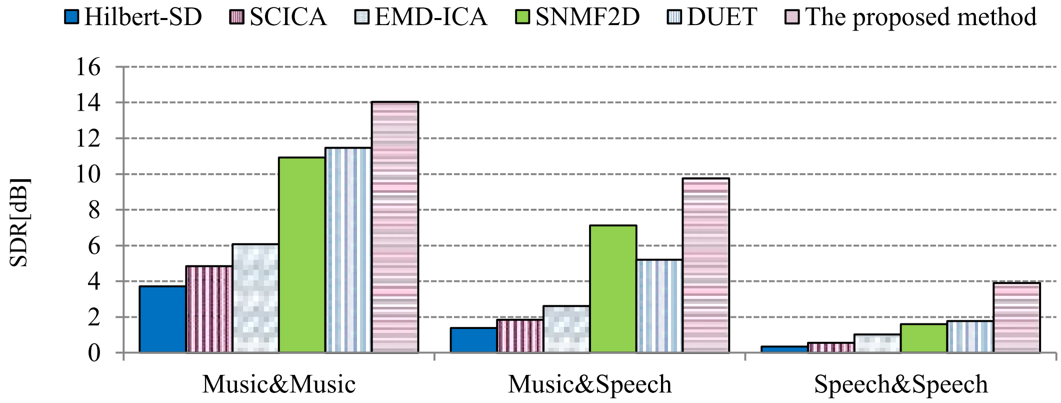


Fig.5: Comparison of average SDR performance on mixture of two audio sources with Hilbert-SD, SCICA, EMD-ICA, SNMF2D, imitated-stereo mixture using DUET, and the proposed method with $\beta = 1$ and $\delta = 1$.

2) Real Stereo signal (left channel only)

In this evaluation, three stereo signals wdrums, nodrums and Shannonsongs Sunrise are used to demonstrate the effectiveness of the proposed method in dealing with having one signal from left channel of stereo signals.$s_1(t)$ is a “left channel mixture” of stereo signal, and $s_2(t)$ is a imitated stereo mixture which was generated from (3). Fig. 6 shows the three original sources, the single channel mixture and the separated sources using the proposed method with $\beta = 1$ and $\delta = 1$. From the plots, it is visually evident that the mixture has been clearly separated in comparison with the original sources.

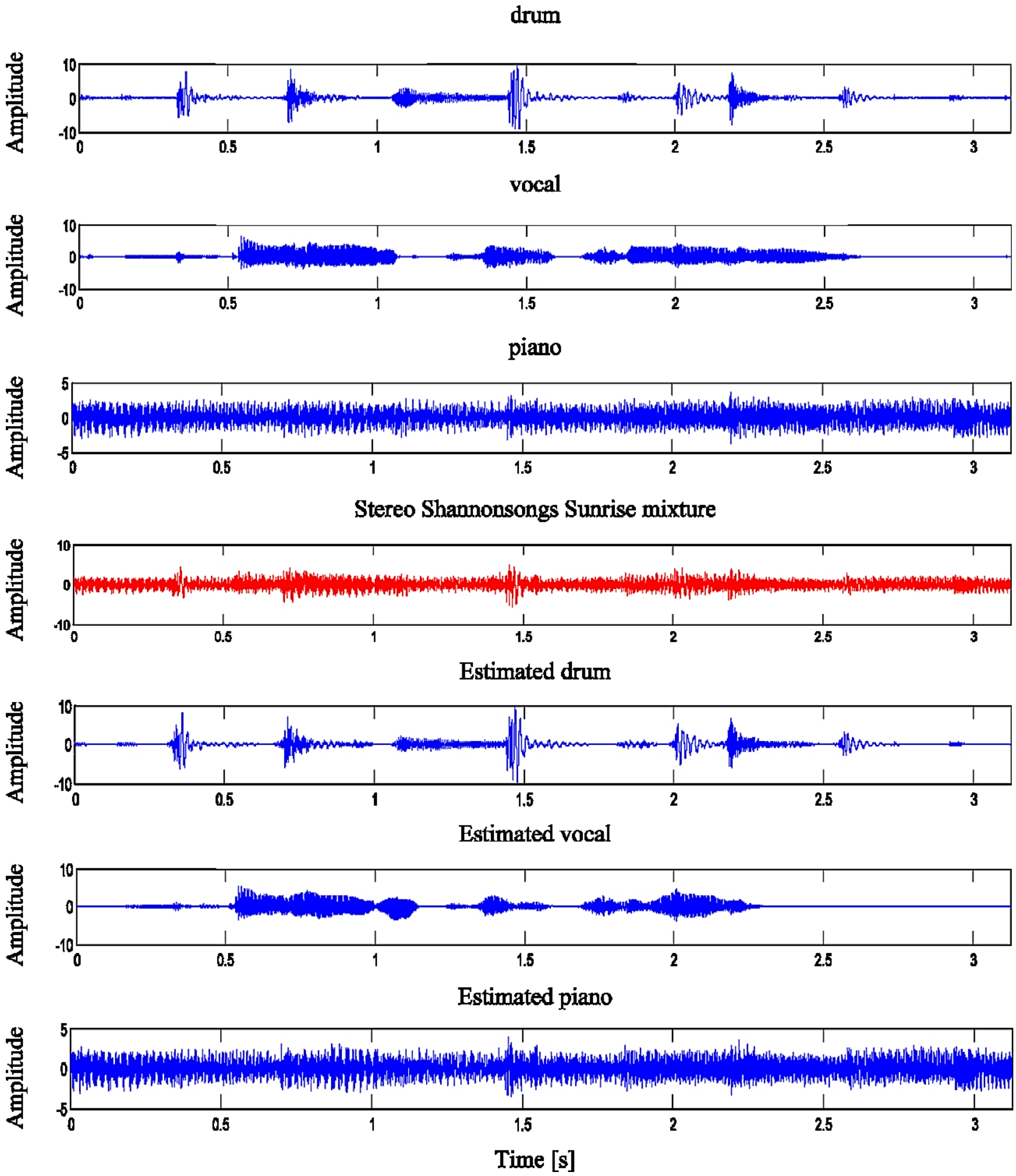


Fig. 6: Original sources, single channel mixture, and estimated sources of Shannonsongs Sunrise mixture using the proposed method with $\beta = 1$ and $\delta = 1$.

The performance evaluation of the proposed method was illustrated in Fig.7 by comparing with the imitated-stereo mixture using DUET, SNMF2D, EMD-ICA, SCICA, and Hilbert-SD methods.

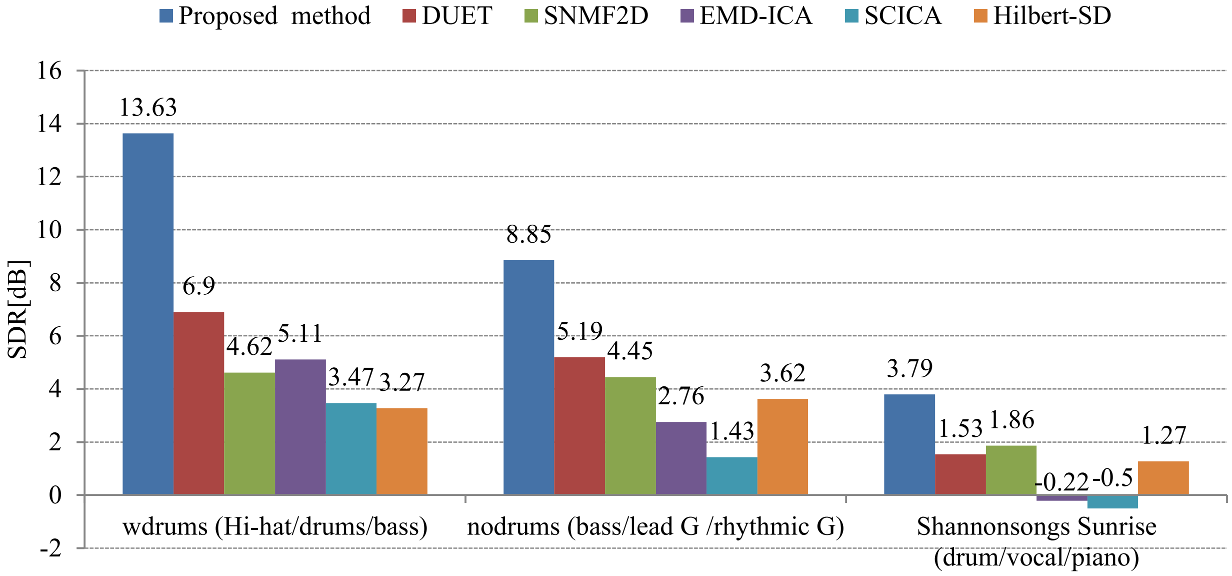


Fig. 7: Comparison of average SDR performance on mixture of two audio sources with the proposed method with $\beta = 1$ and $\delta = 1$, imitated-stereo mixture using DUET, SNMF2D, EMD-ICA, SCICA, and Hilbert-SD.

Table I presents the comparison of the proposed method and the existing well-known SCBSS methods. The proposed imitated-stereo method yields an outstanding performance over the DUET, SNMF2D, EMD-ICA, SCICA, and Hilbert-SD with a total average improvement 5.82 dB per source. In terms of percentage, the average performance improvement of the proposed method against the comparison methods are 92.9%, 140.3%, 242.1%, 497.0% and 311.1%, respectively.

Table I: Comparison of average SDR, SIR and SAR performance on three mixtures of three audio sources between Hilbert-SD, SCICA, EMD-ICA, SNMF2D, imitated-stereo mixture using DUET and the proposed method with $\beta = 1$ and $\delta = -1$.

Mixtures	**Methods**	**SDR**	**SIR**	**SAR**
wdrums (Hi-hat/drums/bass)	Proposed method	13.63	37.95	13.65
	DUET	6.90	18.45	8.72
	SNMF2D	4.62	11.90	6.45
	EMD-ICA	5.11	13.55	5.12
	SCICA	3.47	12.28	4.04
	Hilbert-SD	3.27	10.98	3.53

nodrums (bass/lead G /rhythmic G)	Proposed method	8.85	31.85	8.84
	DUET	5.19	14.71	5.43
	SNMF2D	4.45	12.15	6.13
	EMD-ICA	2.79	14.12	1.97
	SCICA	1.43	13.50	2.57
	Hilbert-SD	3.62	13.04	5.22
Shannonsongs Sunrise (drum/vocal/piano)	Proposed method	3.79	12.83	3.85
	DUET	1.53	7.36	1.47
	SNMF2D	1.86	6.24	2.14
	EMD-ICA	-0.22	4.48	-0.96
	SCICA	-0.50	3.31	-0.62
	Hilbert-SD	1.27	7.13	1.45

The proposed method yields the best separation performance for all recovered sources. The performance of SCICA method depends on the statistical independence between the sources. As this condition is relaxed, the separation performance progressively deteriorates. EMD-ICA method works similarly to SCICA but the mixed signal is firstly pre-processed by the empirical mode decomposition (EMD), which acts as a filterbank whose cut-off frequencies are determined by the signal itself. The EMD enables the mixed signal to be coarsely separated and thus extenuate the amount of mixing before to feeding the outputs to the ICA stage for finer separation. However, the EMD cannot effectively separate the mixture if both original sources share the same frequency bands. The NMF2D method extracts the time and frequency features of each the audio source and works well when the frequency bases are invariant. However, the obtained frequency bases are not enough to dynamically capture the underlying time-varying spectral patterns of the sources especially with speech. The performance of the Hilbert-SD method relies on the derived frequency independent basis vectors which are stationary over time. Therefore, good separation results can be obtained only if the basis vectors corresponding to individual source are statistical independent within the processing window. Thus, if the frequency features of the sources are too similar, it becomes difficult to obtain the independent basis vectors. This explains the reason Table I shows a relatively poorer performance when separating mixture that contains speech sources. On the other hand, since our proposed method generates two channels mixture from a single recording by using the imitated-stereo technique, it benefits from the sources' temporal correlation diversity. As long as the selected AR coefficient pertaining to each source is distinct, the mixture will be separable and the sources can be estimated using the time-frequency Wiener filter in Step 6 of the proposed method.

The benchmark methods can be classified according to their executing processes: First is time-frequency (TF) execution and second is time-series execution. On one hand, the TF execution consists of the proposed method, DUET, SNMF2D, Hilbert-SD. On another hand, the time-series execution is SCICA and EMD-ICA. General speaking, the complexity of TF execution is commonly higher than the time-series execution. Hence, the proposed method has higher complexity than SCICA and EMD-ICA. The computational complexity of the TF execution class has been elucidated.

The proposed method is based on NTF approach which is a form of factorization in 3 dimensions as: $F \times T_s \times I$ where F, T_s, and I denote number of frequency bands, number of time-index and number of mixtures, respectively. The computational complexity of the proposed method is dominated by iterating parameter update (R). The complexity of the proposed method can be express as: $C_{Proposed\ Method} = F \times T_s \times I \times R_{NTF}$. Secondly, the DUET method transforms the TF matrix of the mixtures into a power weighted histogram and the remaining steps are then performed in one-go. Thus, the complexity of the DUET method can be expressed as: $C_{DUET} = F \times T_s \times I$. The SNMF2D is based on 2 dimensional matrix factorization of a sole mixture corresponding to the number of iterative. The complexity of the SNMF2D is formed as: $C_{SNMF2D} = F \times T_s \times R_{NMF}$. Finally, the Hilbert-SD is used which combines the Hilbert transformation with the iterative EMD decomposition. The complexity of the Hilbert-SD can be written as: $C_{Hilbert-SD} = F \times T_s \times R_{EMD}$. Hence, the comparison of complexity is shown in Table II.

Table II: Comparison of complexity ration of DUET, SNMF2D, and Hilbert-SD to the proposed method.

Complexity Ratio	DUET	SNMF2D	Hilbert-SD
Proposed Method	$\frac{1}{R_{NTF}}$	$\frac{R_{NMF}}{I \times R_{NTF}}$	$\frac{R_{EMD}}{I \times R_{NTF}}$

The proposed method is augmented with high computational complexity among the benchmark methods. Future work will investigate the feasibility of the proposed method in alternative TF domains with adaptive sparseness.

6 Conclusion

A novel single channel blind source separation based on NTF is proposed. The NTF separability of the imitated stereo mixture was derived and proved that the proposed method is able to discover the original sound from a sole mixture. The conventional NTF was extended by applying the modified Gaussian prior to extract the correlation between different basis vectors. The modified Gaussian prior is modelled by μ_{kj} that allows the proposed matrix factorization to capture the features of these patterns more efficiently. Additionally, the proposed algorithm can automatically detect the optimal number of latent components of the individual source, thus enabling the spectral dictionary and temporal codes of the individual source to be estimated more efficiently. Experiments have been conducted successfully to separate real-audio mixtures. Results show that the separating performance of the proposed method yields the outstanding performance over the state-of-the-art BSS methods. However, in the case of computational complexity, the proposed method is highest among the other comparison methods. Due to, the proposed method performs iterative parameters updating and computes the nonnegative matrix decomposition given by two imitated channels. While the DUET method discovers the original signals in one-go by using the Wiener masking and the conventional NMF performs by a single channel. Therefore, in the future work, the performance improvement of the proposed method is aimed to reduce the computational time.

References

[1] Abramovich, Y. I. Besson, and Johnson, O. B. A.: Conditional expected likelihood technique for compound-Gaussian and Gaussian distributed noise mixtures. Trans. on Signal Process. (2016).

[2] Aissa-El-Bey, A. Linh-Trung, N. Abed-Meraim, K. Belouchrani, A. Grenier, Y.: Under- determined blind separation of nondisjoint sources in the time-frequency domain. IEEE Transactions on Signal Processing. 55(3), 897–907 (2007).

[3] Al-Tmeme, A. Woo, W.L. Dlay, S.S. and Gao, B.: Underdetermined Convolutive Source Separation using GEM-MU with Variational Approximated Optimum Model Order NMF2D. IEEE Trans. on Audio, Speech and Language Processing. 75(1), 35-49 (2016).

[4] Cherry C. E.: Some experiments on the recognition of speech, with one and with two ears. Journal of Acoustical Society of America. 25(5), 975-979 (1953).

[5] Cichocki, A. Zdunek, R. and Amari, S.I.: 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. 3889, 32–39 (Mar. 2006).

[6] Févotte, C. and Ozerov, A.: Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues. 7th Int. Sym. on Computer Music Modeling and Retrieval, (CMMR 2010). (2010).

[7] Févotte, C. Bertin, N. and Durrieu, J.-L.: Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Computation. 21, 793–830 (Mar. 2009).

[8] FitzGerald, D. Cranitch, M. and Coyle, E.: Non-negative Tensor Factorization for Sound Source Separation. Irish Signals and Systems Conf. Dublin, Ireland, (2005).

[9] Frein, R. de and Rickard, S.: The synchronized short-time-Fourier-transform: properties and definitions for multichannel source separation. IEEE Trans. Signal Process. 59(1), 91-103 (Jan. 2011).

[10] Gao, B. Woo, W. L. and Dlay, S. S.: Unsupervised single-channel separation of nonstationary signals using Gammatone filterbank and Itakura–Saito nonnegative matrix two-dimensional factorizations. IEEE Trans. Circuits and Sys. 60(3), 662-675 (2013).

[11] Gao, B. Woo, W. L. and Dlay, S. S.: Variational regularized 2-D nonnegative matrix factorization. IEEE Trans. Neural Netw. 23(5), 703–716 (May 2012).

[12] Ge, S. Han, J. Han, M.: Nonnegative mixture for underdetermined blind source separation based on a tensor algorithm. Circuits, Systems, and Signal Processing. 34(9), 2935–2950 (2015).

[13] Goto, M. Hashiguchi, H. Nishimura, T. and Oka, R.: RWC music database: Music genre database and musical instrument sound database. in Proc. Int. Sym. Music Inf. Retrieval (ISMIR), Baltimore. 229–230 (Oct.2003).

[14] Guo Y. Naik G. R. and Nguyen H.: Single channel blind source separation based local mean decomposition for Biomedical applications. in Proc. IEEE 35th Annual Int. Conf. Engineering in Medicine and Biology Society (EMBC). 6812-6815 (Jul. 2013).

[15] Guo, H. Li, X. Zhou, L. and Wu, Z.: Single-channel Speech Separation Using Dictionary-updated Orthogonal Matching Pursuit and Temporal Structure Information. Circuits, Systems, and Signal Processing. 34(12), 3861–3882 (Dec. 2015).

[16] Harva, M. and Kabán, A.: Variational learning for rectified factor analysis. Signal Processing. 87(3), 509–527 (2007).

[17] Hild II, K. E. Attias, H. T. and Nagarajan, S. S.: An expectation–maximization method for spatio–temporal blind source separation using an AR-MOG source model. IEEE Trans. Neural Netw. 19(3), 508-519 (Mar. 2008).

[18] Hu, K. and Wang, D. L.: Unvoiced speech separation from nonspeech interference via CASA and spectral subtraction. IEEE Trans. Audio, Speech, and Lang. Process. 19(6), 1600-1609 (Aug. 2011).

[19] Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. 10(3), 626–634 (1999).

[20] Kim, S. and Yoo, C. D.: Underdetermined blind source separation based on subspace representation. IEEE Transactions on Signal processing. 57(7), 2604–2614(2009).

[21] Kitamura, D. Ono, N. Sawada, H. Kameoka, H. and Saruwatari, H.: Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization. IEEE Trans. On Audio, Speech, and Language Procss. 24(9), 1626-1641 (Sep. 2016).

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

[23] Kumar, V. A. Ch. Rao,V. R. and Dutta, A.: Performance Analysis of Blind Source Separation Using Canonical Correlation. Circuits, Systems, and Signal Processing. 1–16, (2017).

[24] Lee, D.D. and Seung, H.S.: Algorithms for non-negative matrix factorization. In Proc. NIPS. 556–562 (2000).

[25] Lee, D.D. and Seung, H.S.: Learning the parts of objects with nonnegative matrix factorization. Nature. 401, 788–791 (1999).

[26] Lin, C.J.: On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Transactions on Neural Networks 18(6), 1589– 1596 (2007).

[27] Lu W. and Zhang B.N.: Single channel time-varying amplitude LFM interference blind separation using MHMPSO particle filtering. in Proc. IEEE Int. Conf. Signal and Image Processing Applications (ICSIPA). 425-430 (Oct. 2013).

[28] Lu, G. Xiao, M. Wei, P. and Zhang, H.: A new method of blind source separation using single-channel ICA based on higher-order statistics. Mathematical Problems in Engineering. Article ID 439264, (2015).

[29] Luengo D. Santamar´ıa I. Vielva L.: A general solution to blind inverse problems for sparse input signals, Neurocomputing. 69(1), 198–215 (2005).

[30] Luengo, D. Santamaría, I. Vielva, L. Pantaleón, C.: Underdetermined blind separation of sparse sources with instantaneous and convolutive mixtures. in IEEE 13th Workshop on: Neural Networks for Signal Processing, NNSP'03. 2003, 279 – 288 (2003).

[31] Mansour A. Benchekroun N. and Gervaise C.: Blind separation of Underwater Acoustic Signals," in Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06). 3889, 181–188 (2006).

[32] Mijovic, B. Vos, M. Gligorijevic, D. Taelman, I. J. and Haffel, S. V.: Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis. IEEE Trans. Biomedical Eng. 57(9), 2188- 2196 (Sep. 2010).

[33] Niknazar, M. Becker, H. Rivet, B. Jutten, C. Comon, P.: Blind source separation of underdetermined mixtures of event-related sources. Signal Processing. 101, 52–64 (2014).

[34] Paatero, P. and Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics. 5(2), 111–126 (1994).

[35] Parathai, P. Woo, W.L. Dlay, S.S. and Gao, B.: Single-channel blind separation using L_1-sparse complex non-negative matrix factorization for acoustic signals. The Journal of the Acoustical Society of America. 137(1), 42 – 49 (2017).

[36] Peng, T. Chen, Y. and Liu, Z. W.: Time–Frequency Domain Blind Source Separation Method for Underdetermined Instantaneous Mixtures. Circuits, Systems, and Signal Processing. 34(12), 3883–3895 (Dec. 2015).

[37] Prasad R.K. Saruwatari H. and Shikano K.: Single Channel Speech Enhancement: MAP Estimation Using GGD Prior Under Blind Setup. in Proc. 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'04). 3195, 873–880 (2004).

[38] Schachtner, R. Pöppel, G. Tomé, A.M. and Lang, E.W.: A Bayesian approach to the lee–seung update rules for nmf. Pattern Recognition Letters. 45, 251–256 (2014).

[39] Signal Separation Evaluation Campaign (SiSEC 2016). (2016). http://sisec.wiki.irisa.fr. Accessed 3 May (2017).

[40] Su, M. K. Tan, T. D. Tobias, J. O. and Gunnar, P.: On the Entropy Computation of Large Complex Gaussian Mixture Distributions. IEEE Trans. on Signal Process. 63(17), 4710 - 4723 (Sep. 2015).

[41] Vincent, E. Gribonval, R. and Févotte, C.: Performance measurement in blind audio source separation. IEEE Trans. Speech, Audio Lang. Process. 14(4), 1462–1469 (Jul. 2005).

[42] Weninger, F., Lehmann, A., Schuller, B.: OpenBliSSART: Design and evaluation of a research toolkit for Blind Source Separation in Audio Recognition Tasks. in Proc. IEEE Int. Conf Acoustics, Speech and Signal Processing (ICASSP). 1625-1628 (May 2011).

[43] Xiang, Y. Ng, S. K. and Nguyen, V. K.: Blind Separation of Mutually Correlated Sources Using Precoders. IEEE Trans. Neural Netw. 21(1), 82–90 (2010).

[44] Xie, S. Yang, L. Yang, J.-M. Zhou, G. Xiang, Y.: Time-frequency approach to underdetermined blind source separation. IEEE Transactions on neural networks and learning systems. 23(2), 306–316 (2012)

[45] Yilmaz, Ö. and Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. 52(7), 1830–1847 (Jul. 2004).

[46] Zhu, H. Zhang, S. and Zhao, H.: Single-Channel Source Separation of Multi-Component Radar Signal with the Same Generalized Period Using ICA. Circuits, Systems, and Signal Processing. 35(1), 353–363 (Jan. 2016).

[47] Zibulevsky M. Pearlmutter B. A.: Blind source separation by sparse decomposition in a signal dictionary, Neural computation. 13(4), 863–882 (2001).