Underdetermined Convolutive Source Separation using GEM-MU with Variational Approximated Optimum Model Order NMF2D

Ahmed Al-Tmeme, W.L. Woo[1], *Senior Member, IEEE*, S.S. Dlay and B. Gao

***Abstract* — An unsupervised machine learning algorithm based on nonnegative matrix factor 2D deconvolution (NMF2D) with approximated optimum model order is proposed. The proposed algorithm adapted under the hybrid framework that combines the generalized EM algorithm with multiplicative update (GEM-MU). As the number of parameters in the NMF2D grows exponentially as the number of frequency basis increases linearly, the issues of model order fitness, initialization and parameters estimation become ever more critical. This paper proposes a variational Bayesian method to optimize the number of components in the NMF2D by using the Gamma-Exponential process as the observation-latent model. In addition, it is shown that the proposed Gamma-Exponential process can be used to initialize the NMF2D parameters. Finally, the paper investigates the issue and advantages of using different window length. Experimental results for the synthetic convolutive mixtures and live recordings verify the competence of the proposed algorithm.**

***Index Terms* — Audio source separation, variational Bayesian, nonnegative matrix factorization, optimum model order selection, generalized expectation-maximization algorithm**

I. Introduction

NONNEGATIVE matrix factorization (NMF) [1, 2] is an important machine learning method in many scientific fields [3-10]. One such field that uses NMF extensively is audio source separation. Audio source separation means estimating the sources from their mixtures and if there is no information about the sources, then the separation will be achieved blindly and the technique will be called blind audio source separation (BSS) [11-13], or it can be achieved in supervised way using the deep neural network (DNN) [14-16] that models the nonlinear relationship between the trained parameters of the targeted speech signal and the mixture signal. Until now audio source separation is an open problem as it does not have the same ability of humans to listen and distinguish between different sources.

Paper is received on 27th April 2016, revised on 4th Oct 2016 and accepted on 14th Oct 2016.

[1] A. Al-Tmeme, W.L. Woo and S.S. Dlay are with School of Electrical and Electronic Engineering, Newcastle University, England, UK. A. Al-Tmeme is on study leave from Al khawarizmi College of Engineering, University of Baghdad, Iraq.

B. Gao is with School of Automation, University of Electronic Science and Technology of China, Chengdu, China.

The work is supported by the Ministry of Higher Education and Scientific Research, Iraq.

(Corresponding author: w.l.woo@ncl.ac.uk)

Audio source separation can be classified according to (i) Input representation: It is related to the time-frequency (TF) representation of the signal, whether it is linear; such as the short time Fourier transform (STFT) [17-20], or quadratic; such as the equivalent rectangular bandwidth (ERB) scale [21]. (ii) Problem dimensionality: It is related to the number of channels and number of sources together; whether the sources are less, equal or greater than the number of channels, and whether the channel is single, or not. Let J be the number of sources and I be the number of channels, then, the following cases can be considered: If $I=1$; then it is a single channel case [22-25]. If $1<I<J$; then it is the underdetermined [17-21]. If $I \geq J$; then it is the Over-determined case [26]. (iii) Mixture: It is related to the type of mixtures; whether it is linear [17-21] or non-linear [27], and the mixing operation; whether it is convolutive [17, 19-21] or instantaneous [28]. The convolutive case is more realistic than the instantaneous one because it considers the reverberation of the channel.

In addition to the above classification, nonnegative matrix factor 2D deconvolution (NMF2D) [29-33], can be appended to the above classification. These methods consider a single channel with linear instantaneous mixture. The problem of NMF2D is that it uses single frequency basis (single component) for each source that is convolved in both time and frequency by a time–pitched weighted matrix [29-33]. It is more suitable for simple musical instruments than complex sound e.g. speech. To overcome this problem, multiple frequency basis (multiple components) are needed; in other words, NMF2D with multiple components where each component is similarly convolved in both time and frequency by a different time-pitched weighted matrix. Consequently, in this paper, a NMF2D with multiple components will be used.

Most methods on BSS that uses the NMF2D are largely confined to instantaneous mixture. Hence there seems to be a gap to the applicability of NMF2D for convolutive mixture. This is not surprising due to the inherent inseparability between the convolutive channel and the convolutive factor used in NMF2D. This paper is an attempt to rigorously overcome this limitation. This paper will also tackle a more challenging case of underdetermined convolutive mixture. The proposed NMF2D with adaptive sparsity will be developed within the framework of the GEM-MU algorithm [34]. The sparsity is the penalty on the activation matrix that ensures only a few units (out of a large population) will be active at the same time [35]. Furthermore, we control the factors that effect on the NMF2D including the cost function, initialization, windows length, and convolutive parameters. Itakura-Saito divergence is considered due its advantage of scale invariance properties [36]. This is important because

source separation requires us to deal with the low and high energy components equally. Compared with the Least Square (LS) distance and Kullback-Leibler (KL) divergence, both methods favor the high energy components but suppress the low energy ones. Furthermore, as each musical instrument has its own characteristics in terms of the spectral and temporal features e.g., drum instrument has a high pitch with low temporal note while the opposite is true for the piano; then different windows length will be considered in the separation. To understand the effects of the convolutive parameters on the separation performance, we briefly describe how the NMF2D works. Let $C(n,m)$ be a data matrix of size $N \times M$ with nonnegative entries, then $C(n,m)$ is approximated with two nonnegative tensors $A(n,k,\tau)$ and $B(k,m,\phi)$ as $C(n,m) \approx \sum_{k=0}^{K}\sum_{\tau=0}^{\tau_{max}-1}\sum_{\phi=0}^{\phi_{max}-1} A(n-\phi,k,\tau)B(k,m-\tau,\phi)$. From the auditory point of view $A(n,k,\tau)$ represents the spectral basis and $B(k,m,\phi)$ represents the temporal code for each spectral basis, the terms K, τ_{max} and ϕ_{max} are the number of components, and the number of the convolutive parameters τ and ϕ, respectively. If τ_{max} and ϕ_{max} are chosen more than the actual requirement, then they will break the structure of the audio signal, i.e., $A(n,k,\tau)$ and $B(k,m,\phi)$ will be shifted more than the actual requirement. This will generate undesirable spurious artefacts to the audio signal and subsequently leads to interference. Therefore, in this paper a novel method will be proposed to estimate the convolutive parameter. Another dimension for consideration is initialization which forms an integral part of the NMF and NMF2D. Good initialization of the model parameters is required for faster convergence to the desired solution.

The novelty of this paper can be summarized as follows: Firstly, a variational Bayesian estimation method using the Gamma-Exponential observation process is proposed to estimate the model order of NMF2D i.e. the optimal number of components, K and the number of convolutive parameters (τ_{max}, ϕ_{max}). Secondly, we propose an initialization scheme for the spectral and temporal parameters in NMF2D. To the best of our knowledge, this is the first research paper that investigates model order estimation and initialization of parameters in NMF2D. Thirdly, the NMF2D with adaptive sparsity will be developed using the GEM-MU algorithm for faster convergence and ensuring the non-negativity of the parameters is preserved. Finally, most current research on NMF2D has been limited to instantaneous mixture [29-33], the present work fills the missing gap by developing the NMF2D with approximated optimum model order for underdetermined convolutive mixture.

The paper is organized as follows: The details of the source model and the development of GEM-MU algorithm to work with NMF2D and adaptive sparsity will be presented in Section II. Section III presents the proposed Gamma-Exponential process for estimating the numbers of components and convolutive parameters, and the initialization of the NMF2D. Experimental results will be presented in Section IV. The effects of the sparsity, initialization, and model order selection on the proposed separation algorithm will be shown in Section V. Finally, conclusions are drawn in Section VI.

II. PROPOSED GEM-MU BASED ADAPTIVE SPARSE NMF2D ALGORITHM

Consider the underdetermined channel with convolutive mixture, namely:

$$\tilde{x}_i(t) = \sum_{j=1}^{J}\sum_{\tau=0}^{L-1} \tilde{a}_{ij}(\tau)\tilde{s}_j(t-\tau) + \tilde{b}_i(t). \quad (1)$$

where $\tilde{x}_i(t)(i = 1,\dots,I, t = 1,\dots,T)$ is the sampled mixture signal and I is number of channels, $\tilde{s}_j$ $(j = 1,\dots,J)$ is the source signal and J is the number of sources, $\tilde{a}_{ij}(\tau)$ is the finite-impulse response of some (causal) filter, and $\tilde{b}_i(t)$ is some additive noise. By assuming that the mixing channel is time-invariant then the short-time Fourier transform (STFT) of (1) can be expressed as

$$x_{i,fn} \approx \sum_{j=1}^{J} a_{ij,f} s_{j,fn} + b_{i,fn} \quad (2a)$$

and in matrix form

$$X_f \approx A_f S_f + B_f \quad (2b)$$

where $X_f = [x_{i,fn}]_f \in \mathbb{C}^{I\times N}$, $A_f = [a_{ij,f}]_f \in \mathbb{C}^{I\times J}$, $S_f = [s_{j,fn}]_f \in \mathbb{C}^{J\times N}$, and $B_f = [b_{i,fn}]_f \in \mathbb{C}^{I\times N}$ and $f = 1,\dots,F$ is the index of a frequency bin. As the NMF2D with multiple components will be considered as the spectral variance model in this paper instead of the NMF spectral model [36], then each source in the STFT can be expressed by K_j complex-valued latent components, i.e., $s_{j,fn} = \sum_{k=1}^{K_j} c_{k,j,fn}$, and can be modeled as realization of proper complex zero-mean variables:

$$c_{k,j,fn} \sim \mathcal{N}_c\left(0, \sigma^2_{k,j,fn}\right) = \mathcal{N}_c\left(0, \sum_{\tau=0}^{\tau_{max,k}-1}\sum_{\phi=0}^{\phi_{max,k}-1} w^{\tau,j}_{f-\phi,k} h^{\phi,j}_{k,n-\tau}\right) \quad (3)$$

where $\mathcal{N}_c(\mu,\Sigma)$ is the proper complex Gaussian distribution [37], $w^{\tau,j}_{f,k}$ represents the spectral basis of the j^{th} source, and $h^{\phi,j}_{k,n}$ represents the temporal code for each spectral basis element of the j^{th} source, for $f = 1,\dots,F, n = 1,\dots,N, j = 1,\dots,J$, and $k = 1,\dots K_j$. The terms $\tau_{max,k}$ and $\phi_{max,k}$ refer to the number of temporal and frequency shifts of the k-th component in the NMF2D model. The noise $b_{i,fn}$ is assumed to be stationary and spatially uncorrelated, i.e.

$$b_{i,fn} \sim \mathcal{N}_c\left(0, \sigma^2_{i,f}\right), \text{ and } \Sigma_{b,f} = diag[\sigma^2_{i,f}]. \quad (4)$$

In this work, the parameters to be estimated are $A_f, \Sigma_{b,f}, \boldsymbol{\Lambda}, C = \{c_{k,j,fn}\}, \boldsymbol{W} = \{w^{\tau,j}_{f,k}\}$, and $\boldsymbol{H} = \{h^{\phi,j}_{k,n}\}$ which are obtained via the posterior probability:

$$P(C, \boldsymbol{W}, \boldsymbol{H} | X_f, A_f, \Sigma_{b,f}, \boldsymbol{\Lambda}) = \frac{P(X_f|C, A_f, \Sigma_b)P(C|\boldsymbol{W},\boldsymbol{H})P(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda})}{P(X_f|A_f,\Sigma_{b,f})}. \quad (5)$$

and their negative log-posterior probabilities are given by

$$-\log P(C, \boldsymbol{W}, \boldsymbol{H} | X_f, A_f, \Sigma_{b,f}, \boldsymbol{\Lambda}) = -\log P(X_f|C, A_f, \Sigma_{b,f}) - \log P(C|\boldsymbol{W},\boldsymbol{H}) - \log P(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda}) + const. \quad (6)$$

where $\boldsymbol{\Lambda}$=$\{\lambda^{\phi,j}_{k,n}\}$ is a 4-dimensional tensor $\mathbb{R}^{K\times N\times \phi_{max}\times J}$ that contains the sparsity terms. The sparsity is the penalty on the

activation matrix that ensures only a few units (out of a large population) will be active at the same time, which can be added as a constraint to the cost function [35].

The GEM-MU combines both the expectation maximization (EM) algorithm and the multiplicative update (MU) algorithm [34]. The source power spectrogram posterior estimates ($\hat{p}_{j,fn}$) (see (8)), the mixing parameters, and the noise covariance will be estimated in the E-step of the EM algorithm, while the parameters $\boldsymbol{W}$ and $\boldsymbol{H}$ will be estimated in the M-step of the EM algorithm by using the MU algorithm coupled with adaptive sparsity.

A. E-Step: Conditional expectations of natural statistics

The negative log-likelihood in the right hand side of (6) can be expressed as

$$\begin{aligned}&-\log P\left(X_f|C,A_f,\Sigma_{b,f}\right)\\&=\sum_{fn}\left(\boldsymbol{x}_{fn}-A_f\boldsymbol{s}_{fn}\right)^H\Sigma_{b,f}^{-1}\left(\boldsymbol{x}_{fn}-A_f\boldsymbol{s}_{fn}\right)+\sum_f\log\left|\Sigma_{b,f}\right|\\&=N\sum_f tr\{\Sigma_{b,f}^{-1}R_{XX,f}\}-N\sum_f tr\{A_f^H\Sigma_{b,f}^{-1}R_{XS,f}\}\\&\quad-N\sum_f tr\left\{\Sigma_{b,f}^{-1}A_f\left(R_{XS,f}\right)^H\right\}+N\sum_f tr\{A_f^H\Sigma_{b,f}^{-1}A_fR_{SS,f}\}\\&\quad+\sum_f\log\left|\Sigma_{b,f}\right|\end{aligned}\tag{7}$$

where the superscript H is the Hermitian transpose, $R_{XX,f}=\frac{1}{N}\sum_n\boldsymbol{x}_{fn}\boldsymbol{x}_{fn}^H$, $R_{SS,f}=\frac{1}{N}\sum_n\boldsymbol{s}_{fn}\boldsymbol{s}_{fn}^H$, $R_{XS,f}=\frac{1}{N}\sum_n\ \boldsymbol{x}_{fn}\boldsymbol{s}_{fn}^H$.The source power spectrogram posterior estimates is as follows:

$$\hat{p}_{j,fn}=\hat{R}_{SS,fn}(j,j)\tag{8}$$

where

$$\hat{R}_{SS,fn}=\mathbb{E}\left[\boldsymbol{s}_{fn}\right]\mathbb{E}\left[\boldsymbol{s}_{fn}^H\right]+\hat{\Sigma}_{s,fn}=\hat{\boldsymbol{s}}_{fn}\hat{\boldsymbol{s}}_{fn}^H+\hat{\Sigma}_{s,fn}\tag{9}$$

$$\hat{\boldsymbol{s}}_{fn}=\Sigma_{s,fn}A_f^H\Sigma_{x,fn}^{-1}\boldsymbol{x}_{fn}\tag{10}$$

$$\hat{\Sigma}_{s,fn}=\left(I_J-\Sigma_{s,fn}A_f^H\Sigma_{x,fn}^{-1}A_f\right)\Sigma_{s,fn}\tag{11}$$

$$\Sigma_{X,fn}=A_f\Sigma_{s,fn}A_f^H+\Sigma_{b,f}\tag{12}$$

$$\Sigma_{s,fn}=diag\left(\left[\sum_{k=1}^{K_j}\sum_{\tau=0}^{\tau_{max,k}-1}\sum_{\phi=0}^{\phi_{max,k}-1}w_{f-\phi,k}^{\tau,j}w_{k,n-\tau}^{\phi,j}\right]_j\right)\tag{13}$$

Detailed derivations of (9) - (13) follow immediately from the linear Gaussian process model [37].

B. M Step: Update of parameters

To find A_f *and* $\Sigma_{b,f}$, we set

$$\frac{\partial}{\partial A_f}\log P\left(C,\boldsymbol{W},\boldsymbol{H}\left|X_f,A_f,\Sigma_{b,f}\right.\right)=0\tag{14}$$

which leads to

$$A_f=\hat{R}_{XS,f}\hat{R}_{SS,f}^{-1}.\tag{15}$$

Similarly,

$$\frac{\partial}{\partial\Sigma_{b,f}^{-1}}\log P\left(C,\boldsymbol{W},\boldsymbol{H}\left|X_f,A_f,\Sigma_{b,f}\right.\right)=0\tag{16}$$

which leads to

$$\Sigma_{b,f}=diag\left(\hat{R}_{XX,f}-\hat{R}_{XS,f}\hat{R}_{SS,f}^{-1}\hat{R}_{XS,f}^H\right).\tag{17}$$

where $\hat{R}_{XS,f}=\frac{1}{N}\sum_n\ \boldsymbol{x}_{fn}E\left[\boldsymbol{s}_{fn}^H\right]=\frac{1}{N}\sum_n\ \boldsymbol{x}_{fn}\hat{\boldsymbol{s}}_{fn}^H$, $\hat{R}_{SS,f}=\frac{1}{N}\sum_n\ \hat{R}_{SS,fn}$ and $\hat{R}_{XX,f}=R_{XX,f}$. As $\hat{p}_{j,fn}$ is estimated from the E-step, the second term in the right hand side of (6) can be written in term of $\hat{p}_{j,fn}$ and expressed with Itakura-Saito divergence as

$$-\log P\left(\hat{P}|\boldsymbol{W},\boldsymbol{H}\right)=\sum_{j,f,n}D_{IS}\left(\hat{p}_{j,fn}\left|\sum_{k,\tau,\phi}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}\right.\right)\tag{18}$$

where $\hat{P}=\left\{\hat{p}_{j,fn}\right\}_{j,fn}$. The third term in the right hand side of (6) is the prior information on $\boldsymbol{W}$ and $\boldsymbol{H}$. An improper prior is assumed for $\boldsymbol{W}$ and factor-wise normalized to unit length i.e. $p(\boldsymbol{W})=\prod_j\delta(\left\|\boldsymbol{W}^j\right\|_2-1)$ where $\boldsymbol{W}^j=\left\{w_{f,k}^{\tau,j}\right\}$ is the spectral basis that belongs to the j-th source. Each element of $\boldsymbol{H}$ has independent decay parameter $\lambda_{k,n}^{\phi,j}$ with exponential distribution:

$$\begin{aligned}p(\boldsymbol{H}|\boldsymbol{\Lambda})&=\prod_{j,k}p\left(H_k^j|\Lambda_k^j\right)=\prod_{j,k,n,\phi}p\left(h_{k,n}^{\phi,j}|\lambda_{k,n}^{\phi,j}\right)\\&=\prod_{j,k,n,\phi}\lambda_{k,n}^{\phi,j}exp\left(-\lambda_{k,n}^{\phi,j}h_{k,n}^{\phi,j}\right)\end{aligned}\tag{19}$$

The negative log-likelihood for prior on $\boldsymbol{H}$ is derived such as

$$\begin{aligned}-\log p(\boldsymbol{H}|\boldsymbol{\Lambda})&=-\log\left(\prod_{j,k,n,\phi}\lambda_{k,n}^{\phi,j}exp\left(-\lambda_{k,n}^{\phi,j}h_{k,n}^{\phi,j}\right)\right)\\&=\sum_{j,k,f,n}\left(\lambda_{k,n}^{\phi,j}h_{k,n}^{\phi,j}-\log\lambda_{k,n}^{\phi,j}\right)\end{aligned}\tag{20}$$

By adding (20) to IS divergence derived in (18), we obtain

$$\begin{aligned}&-\log P(C|\boldsymbol{W},\boldsymbol{H})-\log P(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda})\\&=\sum_{j,f,n}D_{IS}\left(\hat{p}_{j,fn}\left|\sum_{k,\tau,\phi}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}\right.\right)-\sum_j\log\delta\left(\left\|\boldsymbol{W}^j\right\|_2-1\right)\\&\quad+\sum_{j,k,n,\phi}\left(\lambda_{k,n}^{\phi,j}h_{k,n}^{\phi,j}-\log\lambda_{k,n}^{\phi,j}\right)\\&=\sum_{j,k,f,n}\left(\frac{\hat{p}_{j,fn}}{\sum_{\tau,\phi}\left(w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}\right)}-\log\frac{\hat{p}_{j,fn}}{\sum_{\tau,\phi}\left(w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}\right)}-1\right)\\&\quad-\sum_j\log\delta\left(\left\|\boldsymbol{W}^j\right\|_2-1\right)+\sum_{j,k,n,\phi}\lambda_{k,n}^{\phi,j}h_{k,n}^{\phi,j}-\sum_{j,k,n,\phi}\log\lambda_{k,n}^{\phi,j}.\end{aligned}\tag{21}$$

Let

$$v_{j,fn}=\sum_k\sum_\tau\sum_\emptyset w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}\tag{22}$$

then the derivatives of individual component for proposed model with respect to $w_{f,k}^{\tau,j}$ and $h_{k,n}^{\phi,j}$ can be derived as:

$$\begin{aligned}&\frac{\partial}{\partial w_{f',k'}^{\tau',j'}}\log P\left(C,\boldsymbol{W},\boldsymbol{H}\left|X_f,A_f,\Sigma_{b,f}\right.\right)\\&=-\sum_{\phi,n}\hat{p}_{j',f'+\phi,n}v_{j',f'+\phi,n}^{-2}\ h_{k',n-\tau'}^{\phi,j'}+\sum_{\phi,n}v_{j',f'+\phi,n}^{-1}h_{k',n-\tau'}^{\phi,j'}\end{aligned}\tag{23}$$

Similarly,

$$\begin{aligned}&\frac{\partial}{\partial h_{k',n'}^{\phi',j'}}\log P\left(C,\boldsymbol{W},\boldsymbol{H}\left|X_f,A_f,\Sigma_{b,f}\right.\right)\\&=-\sum_{f,\tau}\hat{p}_{j',f,n'+\tau}v_{j',f,n'+\tau}^{-2}\,w_{f-\phi',k'}^{\tau,j'}+\sum_{f,\tau}v_{j',f,n'+\tau}^{-1}w_{f-\phi',k'}^{\tau,j'}\\&\quad+\lambda_{k',n'}^{\phi',j'}\end{aligned}\tag{24}$$

For each individual component, the standard gradient descent method is applied with $w_{f',k'}^{\tau',j'}\leftarrow w_{f',k'}^{\tau',j'}-\eta_w\frac{\partial C_{IS}}{\partial w_{f',k'}^{\tau',j'}}$ and $h_{k',n'}^{\phi',j'}\leftarrow h_{k',n'}^{\phi',j'}-\eta_h\frac{\partial C_{IS}}{\partial h_{k',n'}^{\phi',j'}}$ where η_w and η_h are the positive learning rate. Based on [2], the positive learning rate can be set as $\eta_w=w_{f',k'}^{\tau',j'}/\sum_{\phi,n}v_{j',f'+\phi,n}^{-1}h_{k',n-\tau'}^{\phi,j'}$ and $\eta_h=$

$h_{k',n'}^{\phi',j'}/(\sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'})$. This gives the multiplicative update (MU) rules for $w_{f,k}^{\tau,j}$:

$$w_{f',k'}^{\tau',j'} \leftarrow w_{f',k'}^{\tau',j'} \left(\frac{\sum_{\phi,n} \hat{p}_{j',f'+\phi,n} v_{j',f'+\phi,n}^{-2} h_{k',n-\tau'}^{\phi,j'}}{\sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}} \right) \quad (25)$$

In order to satisfy the constraint $\delta\left(\|\boldsymbol{W}^j\|_2 - 1\right)$, each spectral dictionary is explicitly normalized to unity i.e. $w_{f,k}^{\tau,j} = w_{f,k}^{\tau,j} / \sqrt{\sum_{f,\tau,k} \left(w_{f,k}^{\tau,j}\right)^2}$. Similarly, for $h_{k,n}^{\phi,j'}$ we have

$$h_{k',n'}^{\phi',j'} \leftarrow h_{k',n'}^{\phi',j'} \left(\frac{\sum_{f,\tau} \hat{p}_{j',f,n'+\tau} v_{j',f,n'+\tau}^{-2} w_{f-\phi',k'}^{\tau,j'}}{\sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}} \right). \quad (26)$$

For the sparsity term, the update is obtained by driving the derivative with respect to $\lambda_{k,n}^{\phi,j}$ to zero:

$$\frac{\partial}{\partial \lambda_{k',n'}^{\phi',j'}} \log P\left(C, \boldsymbol{W}, \boldsymbol{H} \,|X_f, A_f, \Sigma_{b,f}\right)$$

$$= \frac{\partial \left(\sum_{fn} \left(\frac{\hat{p}_{j,fn}}{v_{j,f,n}} - log \frac{\hat{p}_{j,fn}}{v_{j,f,n}} - 1 \right) + \sum_{n,\phi} h_{k,n}^{\phi,j} \lambda_{k,n}^{\phi} - \sum_{n,\phi} \log \lambda_{k,n}^{\phi,j} \right)}{\partial \lambda_{j',n'}^{\phi'}}$$

$$= h_{k',n'}^{\phi',j'} - \frac{1}{\lambda_{k',n'}^{\phi',j'}} \quad (27)$$

Therefore, the solution for $\lambda_{k',n'}^{\phi',j'}$ is given by

$$\lambda_{k',n'}^{\phi',j'} = \frac{1}{h_{k',n'}^{\phi',j'} + \epsilon} \quad (28)$$

where ϵ is a small positive random value to prevent division by zero when $h_{k',n'}^{\phi',j'} = 0$. The introduction of ϵ in (28) is necessary to ensure $h_{k,n}^{\phi,j}$ remains sparse while $\lambda_{k,n}^{\phi,j}$ finite.

C. *Components Reconstruction*

The estimated sources ($\hat{\boldsymbol{s}}_{fn}$) can be reconstructed by using Wiener filtering (10) or signal presence probability [38]), and due to the linearity of the STFT, the inverse-STFT can be used to transform it to the time domain.

III. Estimating the Optimum Number of Components and Number of Convolutive Parameters in NMF2D

A. Variational Bayesian Formulation

The determination of the number of components in NMF has been previously investigated in [39] by means of nonparametric statistical fit, and in [40] by a Bayesian model based on automatic relevance determination (ARD). These methods have their own merits. However, they may not suitable for NMF2D model as the number of convolutive parameters and number of components will be lumped into a single entity and thus will estimate an overfit model. In this paper, we propose a constrained Gamma-Exponential process to estimate the convolutive parameters and the number of components of the NMF2D. The proposed Gamma-Exponential process introduces a hidden tensor of nonnegative values $\theta_k^{\tau,\phi}$ that weight each element of the factor model ($\sum_{j,k,\tau,\phi} \theta_k^{\tau,\phi} |a_{ij,f}|^2 w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}$) such that the number of components and convolutive parameters are inferred automatically based on the mixture power spectrogram $p_{i,fn}^x$ which can estimated from the observations as $|x_{i,fn}|^2$. The model order k, τ, and ϕ are assigned to a large integer values (ideally infinity) and the proposed model will retain a finite number of each subset corresponding to the active elements in θ. To the best of our knowledge, this is the first proposed method to estimate the number of convolutive parameters of the NMF2D model.

The generative process of the mixture power spectrogram is assumed to follow the Gamma-Exponential process as

$$p_{i,fn}^x \sim Exponential\left(\sum_{j,k,\tau,\phi} \theta_k^{\tau,\phi} r_{ij,f} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j} \right) \quad (29)$$

$$w_{f,k}^{\tau,j} \sim Gamma\left(a_k^{\tau,j}, a_k^{\tau,j}\right) \quad (30)$$

$$h_{k,n}^{\phi,j} \sim Gamma\left(b_k^{\phi,j}, b_k^{\phi,j}\right) \quad (31)$$

$$r_{ij,f} \sim Gamma\left(c_{ij}, c_{ij}\right) \quad (32)$$

$$\theta_k^{\tau,\phi} \sim Gamma\left(\frac{\alpha_k^{\tau,\phi}}{L + \phi_{max} + \tau_{max}}, \alpha_k^{\tau,\phi} d \right) \quad (33)$$

where $r_{ij,f} = |a_{ij,f}|^2$, L is the truncation level, k is the number of components, α, a, b, and c are the shape parameters, and d is the inverse shape parameter $d = \frac{1}{\bar{x}}$, where $\bar{x}$ is the empirical mean of $p_{i,fn}^x$ expressed as:

$$\mathbb{E}_p\left[p_{i,fn}^x\right] = \sum_{j,k,\tau,\phi} \mathbb{E}_p\left[\theta_k^{\tau,\phi}\right] \mathbb{E}_p\left[r_{ij,f}\right] \mathbb{E}_p\left[w_{f-\phi,k}^{\tau,j}\right] \mathbb{E}_p\left[h_{k,n-\tau}^{\phi,j}\right] = \frac{1}{d} \quad (34)$$

We approximate the posterior distribution of parameters $\Omega = \left\{\{\theta_k^{\tau,\phi}\}, \{r_{ij,f}\}, \{w_{f,k}^{\tau,j}\}, \{h_{k,n}^{\phi,j}\}\right\}$ by resorting to the generalized inverse Gaussian (GIG) distribution, the statistical properties of the GIG can be found in [41]. The PDF of the GIG distribution is given by

$$GIG(y; \gamma, \rho, \beta) = \frac{y^{\gamma - 1} \exp\left(-\rho y - \frac{\beta}{y}\right) \left(\frac{\rho}{\beta}\right)^{\frac{\gamma}{2}}}{2\mathcal{K}_\gamma\left(2\sqrt{\rho\beta}\right)} \quad (35)$$

where $\mathcal{K}_\gamma(\cdot)$ is the modified Bessel function of the second kind and $y \geq 0$, $\rho \geq 0$, and $\beta \geq 0$. Using the GIG, the approximate distribution assumes the form of $q(\Omega) = q\left(\{\theta_k^{\tau,\phi}\}, \{r_{ij,f}\}, \{w_{f,k}^{\tau,j}\}, \{h_{k,n}^{\phi,j}\}\right) = q\left(\theta_k^{\tau,\phi}\right) q\left(r_{ij,f}\right) q\left(w_{f,k}^{\tau,j}\right) q\left(h_{k,n}^{\phi,j}\right)$ where

$$q\left(w_{f,k}^{\tau,j}\right) = GIG\left(\gamma_{w,f,k}^{\tau,j}, \rho_{w,f,k}^{\tau,j}, \beta_{w,f,k}^{\tau,j}\right) \quad (36)$$

$$q\left(h_{k,n}^{\phi,j}\right) = GIG\left(\gamma_{h,k,n}^{\phi,j}, \rho_{h,k,n}^{\phi,j}, \beta_{h,k,n}^{\phi,j}\right) \quad (37)$$

$$q\left(r_{ij,f}\right) = GIG\left(\gamma_{r,ijf}, \rho_{r,ijf}, \beta_{r,ijf}\right) \quad (38)$$

$$q\left(\theta_k^{\tau,\phi}\right) = GIG\left(\gamma_{\theta,k}^{\tau,\phi}, \rho_{\theta,k}^{\tau,\phi}, \beta_{\theta,k}^{\tau,\phi}\right) \quad (39)$$

The variational Bayesian solution is given by

$$\log q^*(\Omega_a) = \mathbb{E}_{q(\Omega/a)}\left[\log p\left(p_{fn}, \Omega\right)\right] \quad (40)$$

where

$$\mathbb{E}_{q(\Omega/a)}\left[\log p\left(p_{i,fn}^x, \Omega\right)\right] = \int \log p\left(p_{i,fn}^x, \Omega\right) \prod_{b \neq a} q(\Omega_b)\, d\Omega_b$$

is the expectation of the logarithm of the joint probability of the mixture power spectrogram and the NMF2D model parameters. The marginal likelihood of $p_{i,fn}^x$ can be shown to be lower bounded given by

$$\log p\left(p_{i,fn}^x \middle| \alpha_k^{\tau,\phi}, a_k^{\tau,j}, b_k^{\phi,j}, c_{ij}\right) \geq$$

$$\begin{aligned}&\mathbb{E}_q\left[\log p\left(p_{i,fn}^x\middle|w_{f,k}^{\tau,j},h_{k,n}^{\phi,j},r_{ij,f},\theta_k^{\tau,\phi}\right)\right]\\&+\mathbb{E}_q\left[\log p\left(w_{f,k}^{\tau,j}\middle|a_k^{\tau,j}\right)\right]-\mathbb{E}_q\left[\log q\left(w_{f,k}^{\tau,j}\right)\right]\\&+\mathbb{E}_q\left[\log p\left(h_{k,n}^{\phi,j}\middle|b_k^{\phi,j}\right)\right]-\mathbb{E}_q\left[\log q\left(h_{k,n}^{\phi,j}\right)\right]\\&+\mathbb{E}_q\left[\log p\left(r_{ij,f}\middle|c_{ij}\right)\right]-\mathbb{E}_q\left[\log q\left(r_{ij,f}\right)\right]\\&+\mathbb{E}_q\left[\log p\left(\theta_k^{\tau,\phi}\middle|\alpha_k^{\tau,\phi},d\right)\right]-\mathbb{E}_q\left[\log q\left(\theta_k^{\tau,\phi}\right)\right].\end{aligned}\quad(41)$$

The first term of the right hand side of (41) is intractable. However, by using first-order Taylor series expansion, it can be shown that this term has a closed-form expression as:

$$\begin{aligned}&\mathbb{E}_q\left[\log p\left(p_{i,fn}^x\middle|w_{f,k}^{\tau,j},h_{k,n}^{\phi,j},r_{ij,f},\theta_k^{\tau,\phi}\right)\right]\\&\geq-\sum_{f,n}\sum_{j,k,\tau,\phi}p_{i,fn}^x\left(\varphi_{f,n,j,k}^{\tau,\phi,i}\right)^2\mathbb{E}_q\left[\frac{1}{\theta_k^{\tau,\phi}r_{ij,f}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}}\right]\\&\quad-\log\left(\omega_{f,n}^i\right)+1-\frac{1}{\omega_{f,n}^i}\mathbb{E}_q\left[\theta_k^{\tau,\phi}r_{ij,f}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}\right]\end{aligned}\quad(42)$$

where

$$\varphi_{f,n,j,k}^{\tau,\phi,i}\propto\mathbb{E}_q\left[\frac{1}{\theta_k^{\tau,\phi}r_{ij,f}w_{f,k}^{\tau,j}h_{k,n}^{\phi,j}}\right]^{-1}\quad(43)$$

and

$$\omega_{f,n}^i=\sum_{j,k,\tau,\phi}\mathbb{E}_q\left[\theta_k^{\tau,\phi}r_{ij,f}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}\right]\quad(44)$$

Using the variational Bayesian solution in (40), this leads to the following parameter updates:

$$\gamma_{w,f,k}^{\tau,j}=a_k^{\tau,j}\quad(45a)$$

$$\rho_{w,f,k}^{\tau,j}=a_k^{\tau,j}+\mathbb{E}_q\left[r_{ij,f}\right]\sum_{n,\phi}\frac{\mathbb{E}_q\left[\theta_k^{\tau,\phi}h_{k,n-\tau}^{\phi,j}\right]}{\omega_{f,n}^i}\quad(45b)$$

$$\beta_{w,f,k}^{\tau,j}=\mathbb{E}_q\left[\frac{1}{r_{ij,f}}\right]\sum_{n,\phi}p_{i,fn}^x\left(\varphi_{f,n,j,k}^{\tau,\phi,i}\right)^2\mathbb{E}_q\left[\frac{1}{\theta_k^{\tau,\phi}h_{k,n-\tau}^{\phi,j}}\right]\quad(45c)$$

$$\gamma_{h,k,n}^{\phi,j}=b_k^{\phi,j}\quad(46a)$$

$$\rho_{h,k,n}^{\phi,j}=b_k^{\phi,j}+\sum_{f,\tau}\frac{\mathbb{E}_q\left[\theta_k^{\tau,\phi}r_{ij,f}w_{f-\phi,k}^{\tau,j}\right]}{\omega_{f,n}^i}\quad(46b)$$

$$\beta_{h,k,n}^{\phi,j}=\sum_{f,\tau}p_{i,fn}^x\left(\varphi_{f,n,j,k}^{\tau,\phi,i}\right)^2\mathbb{E}_q\left[\frac{1}{\theta_k^{\tau,\phi}r_{ij,f}w_{f-\phi,k}^{\tau,j}}\right]\quad(46c)$$

$$\gamma_{r,ijf}=c_{ij}\quad(47a)$$

$$\rho_{r,ijf}=c_{ij}+\sum_{n,k,\tau,\phi}\frac{\mathbb{E}_q\left[\theta_k^{\tau,\phi}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}\right]}{\omega_{f,n}^i}\quad(47b)$$

$$\beta_{r,ijf}=\sum_{n,k,\tau,\phi}p_{i,fn}^x\left(\varphi_{f,n,j,k}^{\tau,\phi,i}\right)^2\mathbb{E}_q\left[\frac{1}{\theta_k^{\tau,\phi}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}}\right]\quad(47c)$$

$$\gamma_{\theta,k}^{\tau,\phi}=\frac{\alpha_k^{\tau,\phi}}{L+\phi_{max}+\tau_{max}}\quad(48a)$$

$$\rho_{\theta,k}^{\tau,\phi}=\alpha_k^{\tau,\phi}d+\sum_{f,n,j}\frac{\mathbb{E}_q\left[r_{ij,f}w_{f,k}^{\tau,j}h_{k,n}^{\phi,j}\right]}{\omega_{f,n}^i}\quad(48b)$$

$$\beta_{\theta,k}^{\tau,\phi}=\sum_{f,n,j}p_{i,fn}^x\left(\varphi_{f,n,j,k}^{\tau,\phi,i}\right)^2\mathbb{E}_q\left[\frac{1}{r_{ij,f}w_{f,k}^{\tau,j}h_{k,n}^{\phi,j}}\right]\quad(48c)$$

The expectations over $q(\Omega)$ can be computed by

$$\mathbb{E}_q[y]=\frac{\mathcal{K}_{\gamma+1}\left(2\sqrt{\rho\beta}\right)\sqrt{\beta}}{\mathcal{K}_\gamma\left(2\sqrt{\rho\beta}\right)\sqrt{\rho}}\quad(49)$$

$$\mathbb{E}_q\left[\frac{1}{y}\right]=\frac{\mathcal{K}_{\gamma-1}\left(2\sqrt{\rho\beta}\right)\sqrt{\rho}}{\mathcal{K}_\gamma\left(2\sqrt{\rho\beta}\right)\sqrt{\beta}}\quad(50)$$

Once the GIG statistics are computed, the model order of the NMF2D can be readily estimated and these will be detailed in Section IV (see (52)-(58)). The Gamma-Exponential process should be executed before the proposed estimation algorithm in order to tune the convolutive parameters and number of components.

B. Initialization

The initialization is an essential part for the separation since the NMF2D and its variants are very sensitive to the initialization. We propose the initialization for the spectral basis and temporal code as the variational approximated posterior mean i.e., $\mathbb{E}_q\left[w_{f,k}^{\tau,j}\right]$ and $\mathbb{E}_q\left[h_{k,n}^{\phi,j}\right]$ given by:

$$w_{f,k}^{\tau,j(initial)}=\frac{\sqrt{\beta_{w,f,k}^{\tau,j}/\rho_{w,f,k}^{\tau,j}}\,\mathcal{K}_{\gamma_{w,f,k}^{\tau,j}+1}\left(2\sqrt{\rho_{w,f,k}^{\tau,j}\beta_{w,f,k}^{\tau,j}}\right)}{\mathcal{K}_{\gamma_{w,f,k}^{\tau,j}}\left(2\sqrt{\rho_{w,f,k}^{\tau,j}\beta_{w,f,k}^{\tau,j}}\right)}\quad(51a)$$

$$h_{k,n}^{\phi,j(initial)}=\frac{\sqrt{\beta_{h,k,n}^{\phi,j}/\rho_{h,k,n}^{\phi,j}}\,\mathcal{K}_{\gamma_{h,k,n}^{\phi,j}+1}\left(2\sqrt{\rho_{h,k,n}^{\phi,j}\beta_{h,k,n}^{\phi,j}}\right)}{\mathcal{K}_{\gamma_{h,k,n}^{\phi,j}}\left(2\sqrt{\rho_{h,k,n}^{\phi,j}\beta_{h,k,n}^{\phi,j}}\right)}\quad(51b)$$

Table I summarizes the main steps of the proposed algorithm.

Table I: Proposed algorithm

1. Estimate the number of components and convolutive parameters by using the proposed Gamma-Exponential process in (45)-(48) and compute $\mathbb{E}_q\left[\theta_k^{\tau,\phi}\right]$.
2. Initialize $w_{f,k}^{\tau,j}$ and $h_{k,n}^{\phi,j}$ with the proposed Gamma-Exponential process spectral and temporal tensors in (51a) and (51b), and initialize $\lambda_{k,n}^{\phi,j}$ with positive value.
3. E-step: compute $\hat{p}_{j,fn}$ and $\hat{\boldsymbol{s}}_{fn}$ using (8) and (10).
4. M-step: compute $A_f,\Sigma_{b,f}$, $w_{f,k}^{\tau,j},h_{k,n}^{\phi,j}$ and $\lambda_{k,n}^{\phi,j}$ using (15), (17), (25), (26), and (28).
5. Normalize $w_{f,k}^{\tau,j}=w_{f,k}^{\tau,j}/\sqrt{\sum_{f,k,\tau}\left(w_{f,k}^{\tau,j}\right)^2}$.
6. Repeat E- and M-steps, and the normalization until convergence is achieved i.e. rate of cost change is below a prescribed threshold, ψ.
7. Take inverse STFT of $\hat{s}_{j,fn}$ to obtain $\tilde{s}_j(t)$.

IV. Results and Discussions

The effect of the sparsity on the separation performance will be investigated by comparing between the uniform sparsity and the adaptive sparsity. The experiment has been ran for different values of the uniform sparsity for three sources that are convolutively mixed in stereo mixture. The latter has 1m space between the microphones, 130ms reverberation time, and with 16 kHz sampling frequency. The following parameters were set for the proposed algorithm; $K_j=5$ components per source, $\tau_{max,k}=5$, and $\phi_{max,k}=2$. Furthermore, in order to focus on the sparsity effects only an oracle initialization (where the input parameters are known)

has been used. Fig. 1 shows the average signal-to-distortion ratio (SDR) [42] with respect to the different values of sparsity. The SDR shows a total separation performance that includes a degree of separation and absence of nonlinear distortion. It is clear from Fig. 1 that the adaptive sparsity

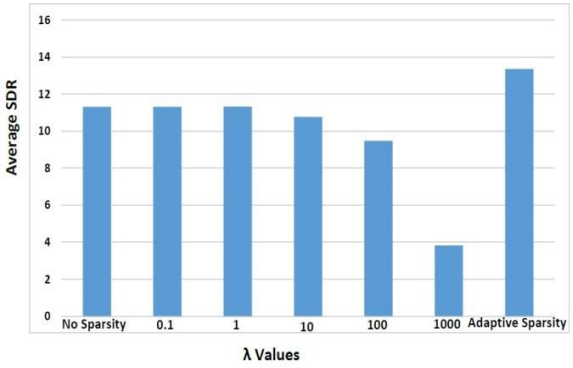

Fig. 1: Average SDR w.r.t different sparsity values

gives the highest SDR as it has a specific sparsity value for each element of the tensor $\boldsymbol{H} = \{h_{k,n}^{\phi,j}\}$ instead of constant value as in the case of constant uniform sparsity. Furthermore, the spectrogram of one of the estimated source for adaptive sparsity, over-sparsity, and the under-sparsity is shown in Fig. 2. It is clear from Fig. 2 that the over-sparsity has nullified many parts of the spectrums from the estimated source, as it assigned far too many zero values to the $\boldsymbol{H}$ tensor. On the other hand, the under-sparsity setting has given rise to redundant spectrum due to the unrestrained elements in the $\boldsymbol{H}$ tensor. This issue is addressed through the adaptive sparsity by specifying a correct sparsity to each element of the $\boldsymbol{H}$ tensor according to (28).

The proposed algorithm will be compared with the standalone

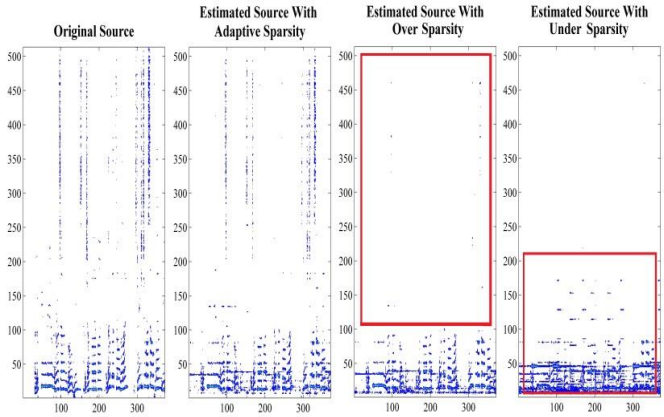

Fig. 2: Effects of sparsity on the estimated source.

EM and MU based algorithms [19], GEM–MU based NTF [34] with adaptive sparsity, the GEM-MU based NMF (this is obtained by setting the convolutive parameters of the proposed algorithm to zero i.e. $\tau_{max} = 1$ and $\phi_{max} = 1$) with adaptive sparsity and proposed initialization, Adiloglu *et al.* [43], and Sawada *et al.* [44]. As our results will be compared with the benchmark MU and EM algorithms of [19], we will consider the same datasets of the synthetic convolutive and the live recording (convolutive) stereo mixture of three sources vocal, percussive musical instruments, and non-percussive (pitched) musical instruments, which matches with the dataset dev2 of SiSEC'08 "under-determined speech and music mixtures". All the mixtures were 10s long, and sampled at 16 kHz. Also, they have 130ms of reverberation time with 1m space between their microphones. Different windows length will be used in the STFT with 50% overlaps. To evaluate the proposed algorithm the performance will be measured using the SDR which measures an overall sound quality of the source separation where it combines the signal-to-interference ratio (SIR), signal-to-noise ratio (SNR), and the signal-to-artifact ratio (SAR) into one measurement. Three dataset will be used in the experiments:

A. Synthetic convolutive dataset: This dataset consist of two groups. The wdrum group which consists of three percussive instruments, and the ndrum group which consists of three non-percussive instruments.

(1) wdrum case: As all the musical instruments are percussive that have short temporal then the STFT with window length of 512-sample is selected. Firstly we will investigate the effect of the proposed Gamma-Exponential process in estimating the number of components and the convolutive parameters. The bounds of the proposed Gamma-Exponential process set as follows: $\tau = \{0, 1, 2, \ldots, 10\}$, $\phi = \{0,1,2, \ldots, 10\}$, and $K = 24$. The results of the proposed Gamma-Exponential process are shown in Figs. 3 and 4. We propose that the number of effective components in the NMF2D is estimated according to the hidden latent variable in (29) as

$$\begin{aligned}\mathbb{E}_q[\theta_k] &= \int \theta_k \, q(\theta_k) d\theta_k \\ &= \int \sum_{\tau=0}^{\tau_{max}-1} \sum_{\phi=0}^{\phi_{max}-1} \theta_k \, q(\theta_k|\tau,\phi) q(\tau) q(\phi) \, d\theta_k \\ &= \frac{1}{\tau_{max}\phi_{max}} \sum_{\tau=0}^{\tau_{max}-1} \sum_{\phi=0}^{\phi_{max}-1} \mathbb{E}_q\left[\theta_k^{\tau,\phi}\right] \end{aligned} \quad (52)$$

where

$$\begin{aligned}\mathbb{E}_q\left[\theta_k^{\tau,\phi}\right] &= \int \theta_k \, q(\theta_k|\tau,\phi) d\theta_k \\ &= \frac{\sqrt{\beta_{\theta,k}^{\tau,\phi}/\rho_{\theta,k}^{\tau,\phi}}\,\mathcal{K}_{\gamma_{\theta,k}^{\tau,\phi}+1}\left(2\sqrt{\rho_{\theta,k}^{\tau,\phi}\beta_{\theta,k}^{\tau,\phi}}\right)}{\mathcal{K}_{\gamma_{\theta,k}^{\tau,\phi}}\left(2\sqrt{\rho_{\theta,k}^{\tau,\phi}\beta_{\theta,k}^{\tau,\phi}}\right)} \end{aligned} \quad (53)$$

The above statistics are obtained from the GIG distribution. It is assumed that both $q(\tau)$ and $q(\phi)$ are uniformly distributed. We define the *effective component* as

$$k_* = \arg_k \left\{ \mathbb{E}_q[\theta_k] \Big/ \sum_{k=1}^{K} \mathbb{E}_q[\theta_k] \; \geq \varepsilon \right\} \quad (54)$$

where ε is a small constant. Through the experiments we found that selecting $\varepsilon = 0.1$ will best fit the proposed algorithm. Therefore, we treat $\mathbb{E}_q[\theta_k]$ as a histogram and select the effective component as those that exceeds 10% of the overall sum. Fig. 3 shows the values of $\mathbb{E}_q[\theta_k]$ for $k = 1, \ldots, 24$ which are predominantly zero except for $k = 3, 8, 11$ and 20 whose $\mathbb{E}_q[\theta_k]$ values are $1.46, 0.07, 2.1$ and 3.23, respectively. The term $\sum_{k=1}^{K} \mathbb{E}_q[\theta_k]$ has been calculated to be 6.86 and thus, the effective components are only $k_* = 3, 11$ and 20. Let $K_* = \#\, k_*$, that is, the number of effective components e.g. in Fig. 3 this corresponds to $K_* = 3$. Since there are $J = 3$ sources, then $K_j = K_*/J = 1$ for $j = 1, 2, 3$. In addition, for each k_* effective component, we have determined distribution for (τ, ϕ) which is given by

$\mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi}\right]$. These are shown in Fig. 4. We select the optimum model for (τ,ϕ) by treating each $\mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi}\right]$ for various values of (τ,ϕ) as a histogram. Thus the optimum model for (τ,ϕ) is given by the average of non-zero components:

$$\hat{\tau}_{max,k_*} = \frac{\sum_{l=0}^{\phi max-1} F_l^{(\tau)}}{\#\left(F_l^{(\tau)} \neq 0, \forall l\right)} \quad (55)$$

$$\hat{\phi}_{max,k_*} = \frac{\sum_{l=0}^{\tau max-1} F_l^{(\phi)}}{\#\left(F_l^{(\phi)} \neq 0, \forall l\right)} \quad (56)$$

where

$$F_l^{(\tau)} = \#component\ in\ \frac{\mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi=l}\right]}{\sum_\tau \mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi=l}\right]} \geq \varepsilon \quad (57)$$

$$F_l^{(\phi)} = \#component\ in\ \frac{\mathbb{E}_q\left[\theta_{k=k_*}^{\tau=l,\phi}\right]}{\sum_\phi \mathbb{E}_q\left[\theta_{k=k_*}^{\tau=l,\phi}\right]} \geq \varepsilon \quad (58)$$

The term $F_l^{(\tau)}$ counts the number of τ components in the normalized $\mathbb{E}_q\left[\theta_k^{\tau,\phi=l}\right]$ that exceeds ε, and $\#\left(F_l^{(\tau)} \neq 0, \forall l\right)$ counts the number of entries of $F_l^{(\tau)}$ that is non-zero. The same interpretation is applied to $F_l^{(\phi)}$ and $\#\left(F_l^{(\phi)} \neq 0, \forall l\right)$ for determining the model order ϕ_{max}. Thus from Fig. 4, we calculate that $\hat{\tau}_{max,k_*} = 5$ and $\hat{\phi}_{max,k_*} = 11$ for all k_*. Hence, the optimum model order for the NMF2D model in (3) is given by $K_j = 1, \tau_{max,k} = 5$ and $\phi_{max,k} = 11$.

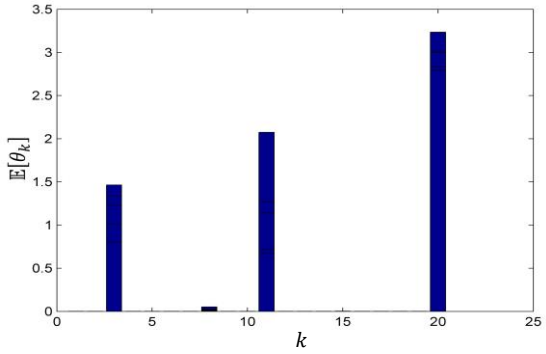

Fig. 3: Number of estimated components.

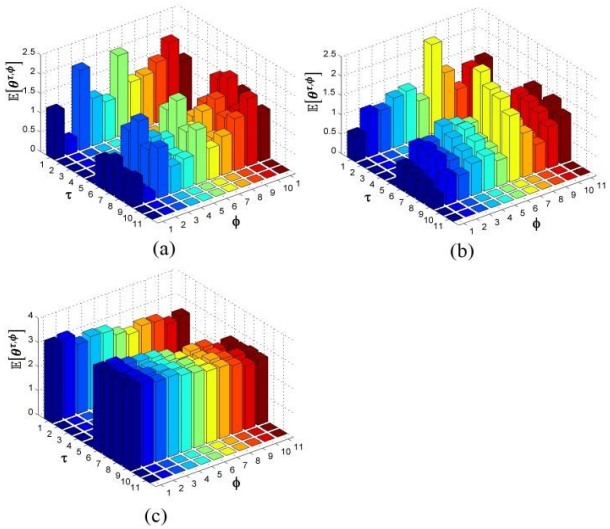

Fig. 4: Convolutive parameters distribution of $\theta_k^{\tau,\phi}$ corresponding to (a) $k = 3$, (b) $k = 11$, and (c) $k = 20$ in Fig. 3.

The tensors of the Gamma-Exponential process (51a) and (51b) are used to initialize the GEM-MU based NMF2D algorithm, and the separation performance is tabulated in Table II. It can be seen that the SDRs of the proposed

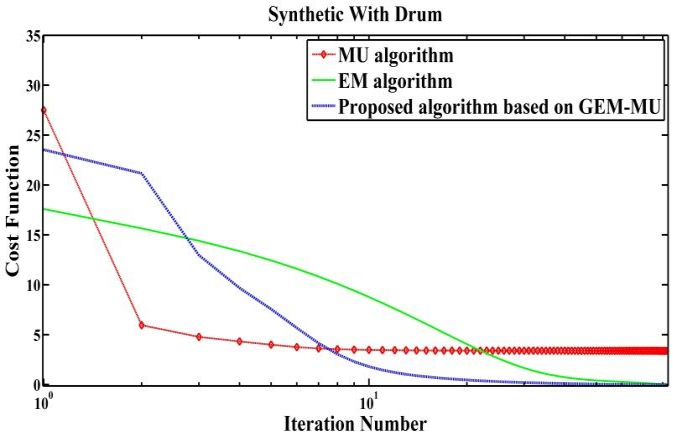

Fig. 5: Convergence of the cost functions.

GEM-MU based NMF2D is better than the other algorithms. This shows that by using the proposed Gamma-Exponential process, we are able to estimate the number of components, convolutive parameters, and initialize the separation algorithm. Furthermore, we plot the cost function (i.e. (6)) versus number of iteration in Fig. 5 (a constant value has been added to the curve to ensure positivity). Fig. 5 shows the convergence trajectory of the tested algorithms. The plot is obtained by evaluating the cost function in (6) and averaging over 200 independent runs. The plot shows that the proposed algorithm has better convergence than both the MU and EM algorithms. It converges in less than 40 iterations. Finally the waveforms of the estimated sources are shown in Fig. 6.

Table II: Convolutive mixture with drum (wdrurm)

Algorithm	**Parameters**	**SDRs**			**Avrg SDR**
		s_1	**s_2**	**s_3**	
EM NMF [19]	Window=512	6.89	-4.83	1.75	1.27
MU NMF [19]	Window=512	5.10	-9.87	2.46	-0.77
GEM-MU NTF [34]	Window=512	6.18	-1.32	3.00	2.62
GEM-MU NMF	Window=512	5.54	-0.28	1.21	2.16
Proposed algorithm	Window=512 $K_j = 1\ \forall j$ $\tau_{max,k} = 5$ $\phi_{max,k} = 11$	**7.99**	**0.22**	**3.86**	**4.02**

(*2) ndrum case*: Since most musical instruments are pitched (non-percussive) and have long temporal characteristics then the STFT with window length of 2048-sample will be selected. By following the same procedure of the wdrum case, the number of components and convolutive parameters are selected from Fig. 7 and Fig. 8, respectively. From Fig. 7, it is calculated that $K_* = 5$ and the effective components are $k_* = \{3, 7, 11, 13, 20\}$. Since there are 3 sources, there is a

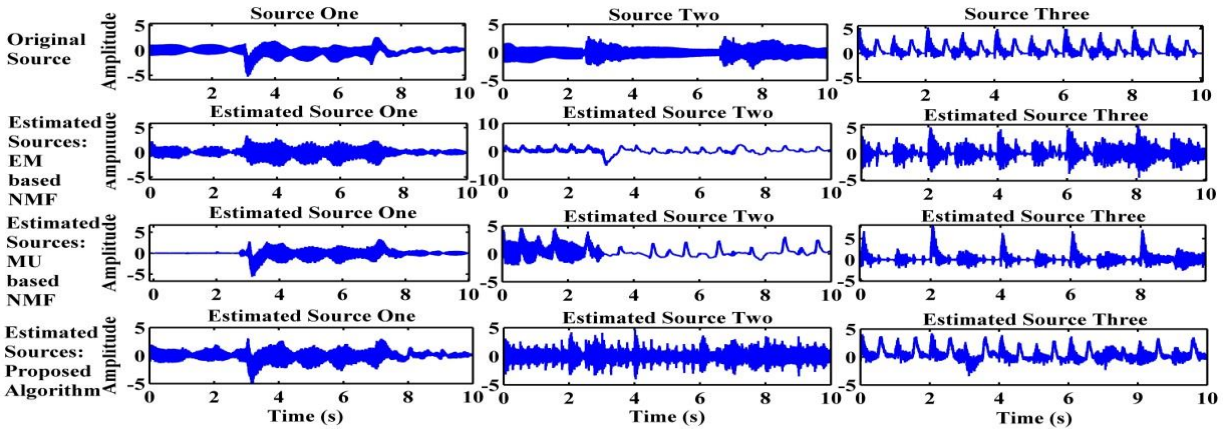

Fig. 6: Waveforms of the estimated sources for drum case.

need to determine which component belongs to which source. To this end, we perform the *k*-means clustering using Kullback-Leibler divergence on the estimated set of effective

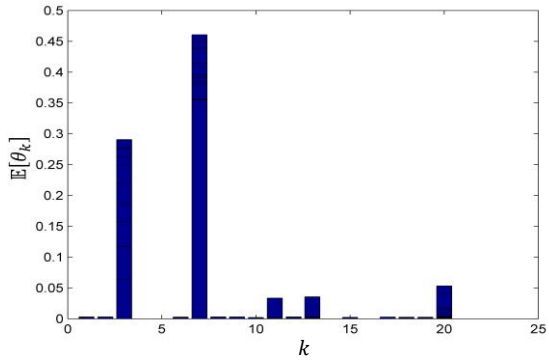

Fig. 7: Number of estimated components.

spectral basis i.e., $\{\mathbb{E}_q[w_{f,k=k_*}^{\tau,j}]\}$ in (51a). Subsequently, this leads to the partition of K_* into $K_1 = K_2 = 2$ and $K_3 = 1$. Also from Fig. 8, the convolutive model order are determined as follows: $(\hat{\tau}_{max,k=3} = 11, \hat{\phi}_{max,k=3} = 11)$, $(\hat{\tau}_{max,k=7} = 1, \hat{\phi}_{max,k=7} = 11)$, $(\hat{\tau}_{max,k=11} = 11, \hat{\phi}_{max,k=11} = 11)$, $(\hat{\tau}_{max,k=13} = 1, \hat{\phi}_{max,k=13} = 11)$, and $(\hat{\tau}_{max,k=20} = 4, \hat{\phi}_{max,k=20} = 11)$. The waveforms of the estimated sources is shown in Fig. 9. Furthermore, all the results are tabulated in Table III. It can be seen that the average SDRs of the proposed algorithm with window 2048-sample are better than other algorithms.

B. Live recording (convolutive) dataset: This dataset is more complicated than the Synthetic convolutive case as it contains different musical instruments with vocal signal. It consists of two groups: (1) wdrum group which consists of vocal and musical instrument with drum, and (2) ndrum group which consists of vocal and musical instruments without drum.

(1) wdrum case: By following similar procedure in the previous section, window length of 2048-sample is selected for the STFT, the number of components and convolutive parameters are selected from Fig. 10 and Fig. 11, respectively. From Fig. 10, it is calculated that the effective number of components is $K_* = 8$. By using the k-means clustering, this leads the partition of K_* into $K_1 = K_3 = 3$ and $K_2 = 2$.The convolutive model orders are determined from Fig. 11 as

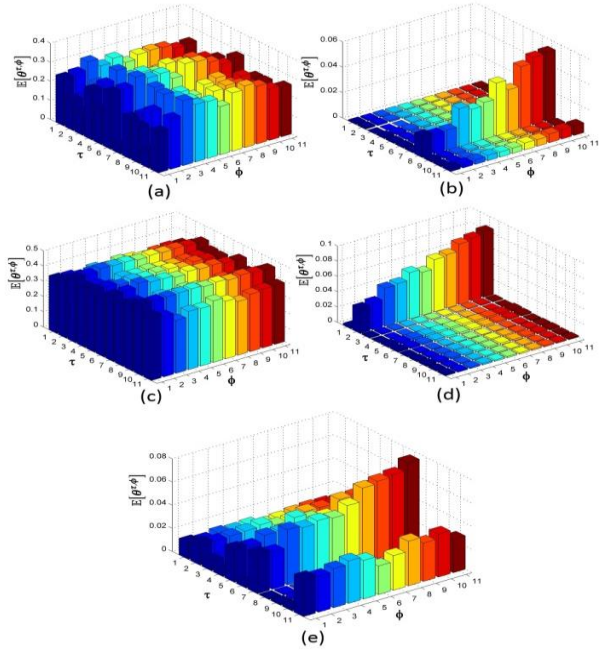

Fig. 8: Convolutive Parameters distribution corresponding to each effective component in Fig. 7.

follows: $(\hat{\tau}_{max,k=3} = 1, \hat{\phi}_{max,k=3} = 6)$, $(\hat{\tau}_{max,k=5} = 2, \hat{\phi}_{max,k=5} = 2)$, $(\hat{\tau}_{max,k=6} = 3, \hat{\phi}_{max,k=6} = 2)$, $(\hat{\tau}_{max,k=10} = 2, \hat{\phi}_{max,k=10} = 4)$, $(\hat{\tau}_{max,k=11} = 2, \hat{\phi}_{max,k=11} = 4)$, $(\hat{\tau}_{max,k=15} = 2, \hat{\phi}_{max,k=15} = 3)$, $(\hat{\tau}_{max,k=16} = 2, \hat{\phi}_{max,k=16} = 3)$ and $(\hat{\tau}_{max,k=17} = 1, \hat{\phi}_{max,k=17} = 6)$. Fig. 12 shows the convergence of the proposed algorithms by averaging 200 independent runs. Additionally, all the results are tabulated in Table IV which shows that the SDRs of the proposed algorithm have been superior. Finally the waveforms of the estimated sources in are shown in Fig. 13.

(2) ndrum case: Since this dataset contains pitched instruments and vocal, and as the vocal sound acts like percussive instrument in long window, then a long window of 4096-sample is selected for the STFT. The number of components and convolutive parameters are selected from Fig. 14 and Fig. 15, respectively. From Fig. 14, it is calculated that $K_* = 15$ and using the k-means clustering this leads the partition of K_* into $K_j = 5$ for $j = 1,2,3$. The convolutive model orders are determined from Fig. 15. It is interesting to

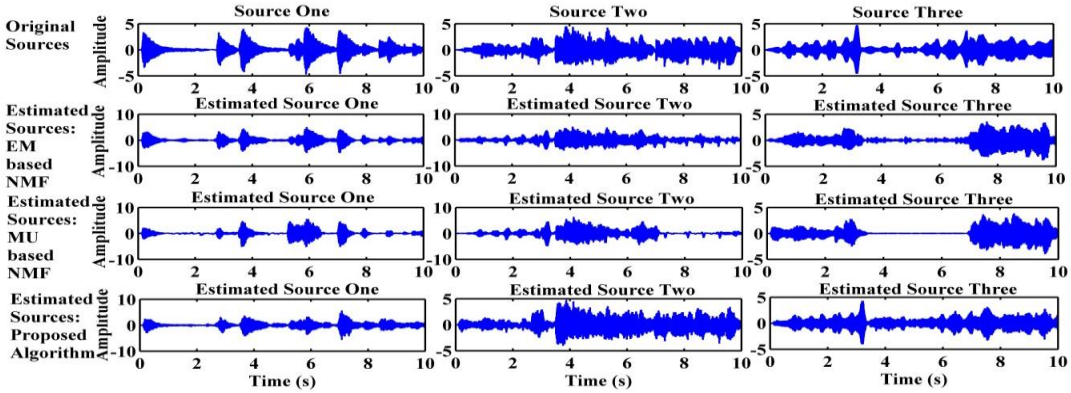

Fig. 9: Waveforms of the estimated sources for no drum case.

Table III: Synthetic convolutive without drum (ndrum).

Algorithm	Parameters	SDRs			Avrg SDR
		s_1	s_2	s_3	
EM NMF [19]	Window=2048	4.18	1.02	-1.8	1.10
MU NMF [19]	Window=2048	2.89	1.04	-2.09	0.61
GEM-MU NTF [34]	Window=2048	2.93	3.09	1.57	2.53
GEM-MU NMF	Window=2048	2.98	2.57	1.15	2.23
Proposed algorithm	Window=2048 $K_1 = K_2 = 2$, and $K_3 = 1$	**4.75**	**3.93**	**4.75**	**4.48**

Table IV: Live recording with drum (wdrum).

Algorithm	Parameters	SDRs			Avr SDR
		s_1	s_2	s_3	
EM NMF [19]	Window=2048	4.96	5.55	8.03	6.18
MU NMF [19]	Window=2048	4.19	4.50	7.58	5.42
GEM-MU NTF [34]	Window=2048	5.89	7.90	7.68	7.16
GEM-MU NMF	Window=2048	5.99	7.74	7.58	7.10
Proposed algorithm	Window=2048 $K_1 = K_3 = 3$, $K_2 = 2$	**6.98**	**8.85**	**8.92**	**8.25**

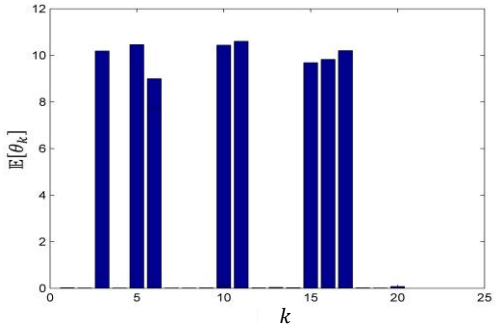

Fig. 10: Number of estimated components.

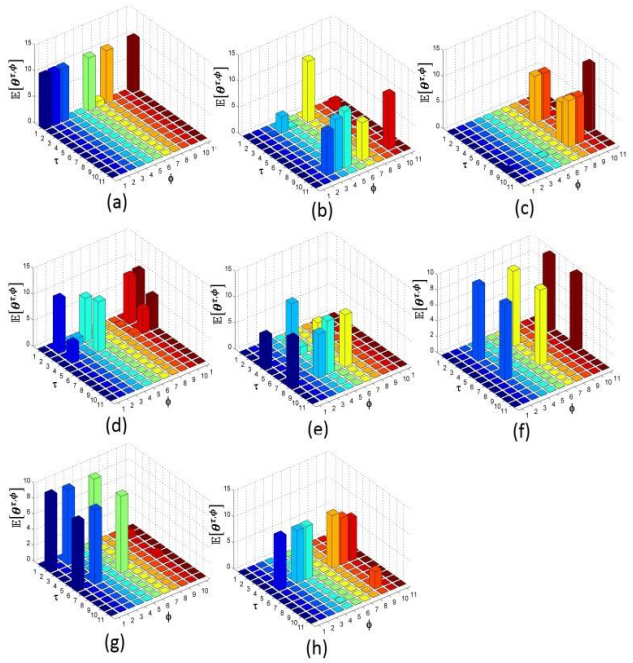

Fig. 11: Convolutive parameters distribution corresponding to each effective component in Fig. 10.

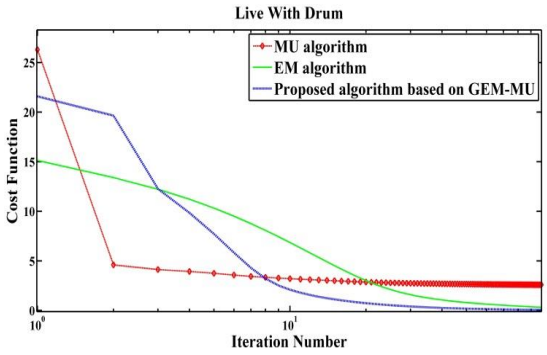

Fig. 12: Convergence of cost functions.

note that, on average, we have $\tau_{\max,k} = 1$ or 2 while $\phi_{max,k} = 6$ or 7. All the result has been tabulated in Table V. Finally, the waveforms of the estimated sources are shown in Fig. 16.

C. Adiloglu et al. algorithm [43]: In this subsection the proposed algorithm will be compared with Adiloglu *et al.*

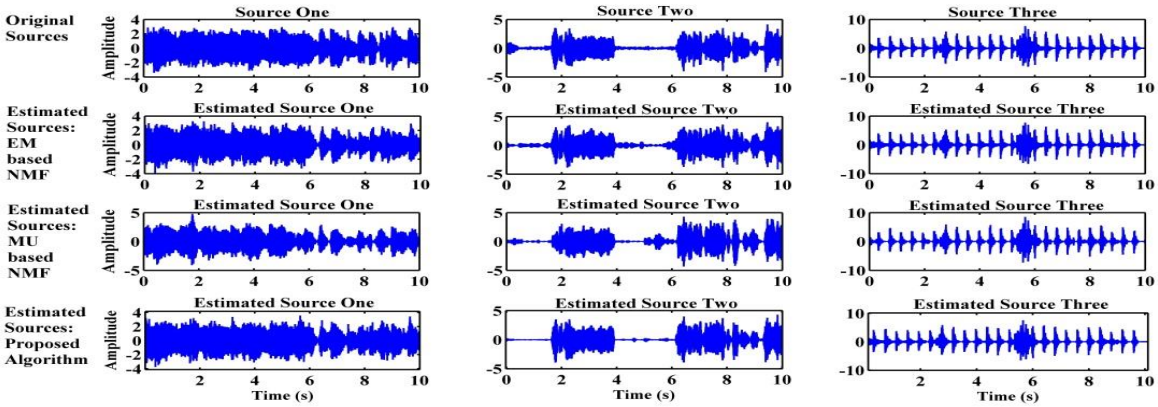

Fig. 13: Waveforms of the estimated sources for the live recording with drum case.

algorithm [43] that considers the fully Bayesian source separation algorithm based on variational inference method [45], the time difference-of-arrival (TDOA) as an initialization algorithm [46], and the multi-level-NMF [7] as a source variance. It consists of two groups: (i) live recording wdrum group which consists of musical instruments with drum and (ii) live recording ndrum group which consists of musical instruments without drum. Both groups have 250 ms reverberation time, and 5 cm and 1 m microphone spacing which matches with the dataset dev1 of SiSEC'13 "under-determined speech and music mixtures". First, the Gamma-Exponential process is used to estimate the number of components and convolutive parameters. The proposed separation algorithm is initialized using the estimated values from the Gamma-Exponential process. The final separation results are tabulated in Table VI. It is seen that the proposed algorithm achieved higher SDRs than Adiloglu *et al.* algorithm which emphasizes that with the correct initialization and correct number of components, the NMF2D-based algorithm can yield high separation performance.

Table V: Live recording without drum (ndrums)

Algorithm	**SDR**			**Avg SDR**
	S_1	**S_2**	**S_3**	
EM NMF [19]	6.02	1.68	-0.91	2.26
MU NMF [19]	4.27	0.05	-3.14	0.39
GEM-MU NTF [31]	7.71	3.60	-0.40	3.64
GEM-MU NMF	6.80	2.10	-0.24	2.89
Proposed algorithm	**8.93**	**4.83**	**3.18**	**5.65**

D. Sawada et al. algorithm [44]: In this subsection the proposed algorithm will be compared with Sawada *et al.* algorithm [44], which is an underdetermined convolutive blind source separation algorithm that carried out in two stages scenario. In the first stage the frequency bin-wise clustering is applied by using EM algorithm. While in the second stage the permutation ambiguities that occurred from the first stage is solved. It uses the same dataset of the previous section and by following the same procedure applied in previous section the results are tabulated in Table VI. Although the performance is still quite good, it is noted that the proposed algorithm achieved higher SDRs than Sawada *et al.* algorithm. This is attributed to the optimal model order used in NMF2D and the source estimation rendered by the GEM-MU framework.

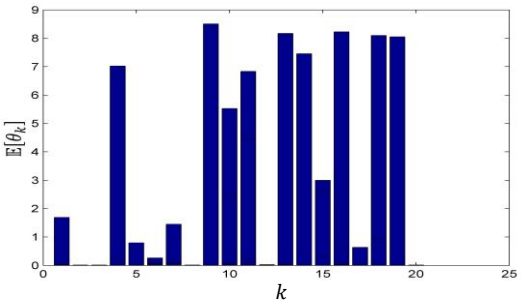

Fig. 14: Number of estimated components.

V. Effects of the sparsity, initialization, and model order selection on the separation Performance

In this section the effects of the sparsity, initialization, and model order selection on the performance of the proposed separation algorithm will be shown. This will be carried out on the same datasets used in experiments A and B above.

A. Effects of the sparsity

Three cases will be considered here, the no sparsity case by setting $\lambda = 0$, fixed uniform sparsity case by setting $\lambda = c$, where c is constant value, and adaptive sparsity case by setting λ accoring to (28). The results are tabulated in Table VII. It can be seen that the best result is obtained from the adaptive sparsity as it assigns a specific sparsity value for each element in the $\boldsymbol{H}$ tensor, while the fixed uniform sparsity assigns fixed value for the entire elements of the $\boldsymbol{H}$ tensor. Assigning fixed large value causes over-sparseness (which removes many elements from the $\boldsymbol{H}$) or under-sparseness (which retain many unwanted elements in $\boldsymbol{H}$) if the value is low, as visually shown in Fig. 2.

B. Effects of the initialization

Depending on how to initialize $\boldsymbol{W}$ and $\boldsymbol{H}$ tensors three cases will be considered here, the random initialization (which is the average of 100 runs), the singular value decomposition (SVD) [47] after adapting it to work with the NMF2D, and the proposed Gamma-Exponential process that initializes $\boldsymbol{W}$ and

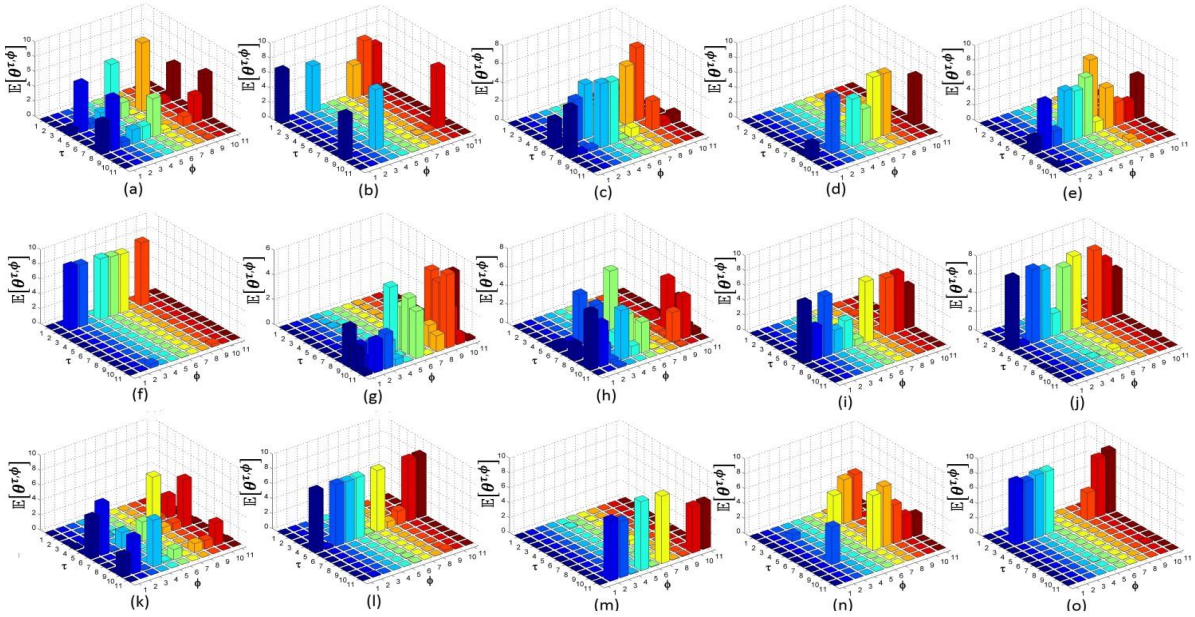

Fig. 15: Convolutive parameters distribution corresponding to each effective component in Fig. 14.

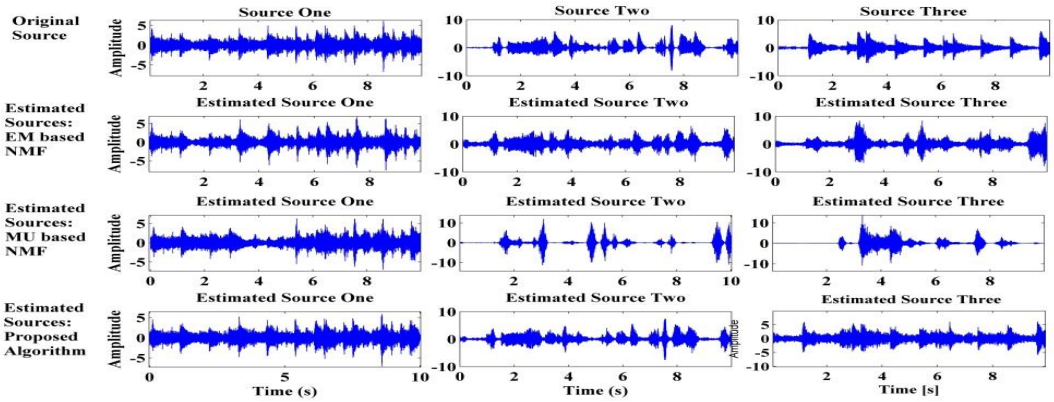

Fig. 16: Waveforms of the estimated sources for the live recording no drum case.

$\boldsymbol{H}$ according to (51a) and (51b), respectively. The results are tabulated in Table VIII. The table shows that initialization through the Gamma-exponential process achieved the highest results, as the $\boldsymbol{W}$ and $\boldsymbol{H}$ are initialized using the Gamma-Exponential process which ensures that they start closed to the desired solution and avoid divergence. However it is time consuming as it is offline initialization process that should be run and converged before initializing the tensors of the proposed separation algorithm.

C. *Effects of the model order selection*

It is not straightforward to compare the proposed Gamma-exponential process with other methods in terms of estimating the model order of the NMF2D. However we proposed to compare with the mesh method that compute the SDR for each single selection of the convolutive parameter (for $\tau = \{0, 1, \dots, 10\}$ and $\phi = \{0,1, \dots, 20\}$) and check the convolutive parameters that give the highest SDR. For fair comparison the SVD [47] has been used to initialize the tensors of the NMF2D. This method is time consuming and unrealistic as it required the original sources to compute their SDRs. We apply it on the case of synthetic convolutive with drum, as shown in Fig. 17. The figure shows the results of the mesh method of running the NMF2D algorithm for every

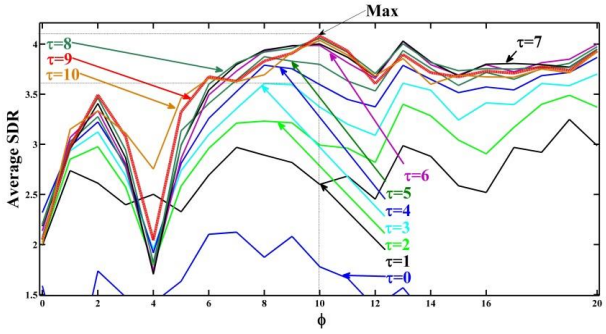

Fig. 17: Results of mesh method

possible case of τ and ϕ. In total, there are $11 \times 21 = 231$ possible model order. The highest SDR is obtained at SDR = 4.08dB with $\tau = \{0, 1, \dots, 9\}$ and $\phi = \{0,1, \dots,10\}$ i.e., this corresponds to $\tau_{max} = 10$ and $\phi_{max} = 11$ in the proposed model in (3). In addition, the figure reveals that a range of relatively high SDR is attained around the model order $\tau_{max} = 5$ to 10, and $\phi_{max} = 10$ to 12. On the other hand, the result attained using the Gamma-Exponential process indicates a model order of $\tau_{max} = 5$ and $\phi_{max} = 11$ which gives a SDR of 3.71dB. Note that the attained model order using Gamma-Exponential process lies within the range of high SDR performance obtained by the mesh method. Thus

the result shows that the Gamma-Exponential process does not only lead to a model order that maintain a high SDR performance but also a sparse model.

Table VI: Adiloglu *et al*. and Sawada *et al*. algorithms

			ndrums		wdrums	
Reverberation Time (ms)			**250**		**250**	
Microphone Distance (cm)			**5**	**100**	**5**	**100**
Adiloglu *et al.* [43]	SDR	s_1	-5.5	-0.6	7.0	2.4
		s_2	-1.2	-0.0	-0.1	3.0
		s_3	3.7	0.6	-0.5	-11.1
		Avg	-1.0	0.0	2.1	-1.9
Sawada *et al.* [44]	SDR	s_1	1.19	0.13	5.00	4.83
		s_2	3.48	1.67	0.60	-0.24
		s_3	2.89	1.81	-5.48	-6.68
		Avg	2.52	1.20	0.04	-0.69
Proposed algorithm		K	1		4	
		τ_{max}	11		1	
		ϕ_{max}	11		6	
	SDR	s_1	1.57	1.42	7.17	7.81
		s_2	5.30	1.97	0.92	0.2
		s_3	2.76	1.65	1.44	-3.78
		Avg	3.21	1.68	3.17	1.41

Table VII: Effects of the sparsity on the proposed algorithm

			No Sparsity	**Uniform sparsity**	**Adaptive sparsity**
Synthetic Convolutive with drum	SDR	s_1	1.89	7.40	7.95
		s_2	0.08	-0.02	0.22
		s_3	-2.47	3.80	3.90
		Avg	-0.17	3.72	4.02
Synthetic convolutive without drum	SDR	s_1	2.81	3.20	4.75
		s_2	2.68	3.36	3.93
		s_3	1.36	2.51	4.75
		Avg	2.28	2.69	4.48
Live recording with drum	SDR	s_1	5.70	6.23	6.98
		s_2	7.47	8.04	8.85
		s_3	6.78	7.57	8.92
		Avg	6.65	7.28	8.25
Live recording without drum	SDR	s_1	7.35	8.86	8.93
		s_2	3.28	4.76	4.83
		s_3	-0.34	-0.01	3.18
		Avg	3.43	4.54	5.65
Avg of all datasets	**SDR**		**3.05**	**4.56**	**5.60**

VI. CONCLUSIONS

In this paper, an approximated optimal NMF2D with adaptive sparsity has been proposed within the linear Gaussian framework in the time-frequency domain for separating the underdetermined convolutive mixture. The parameters are estimated using the GEM-MU algorithm which has superior performance to efficiently initialize the NMF2D model. Furthermore, a variational Bayesian approach using the generalized inverse Gaussian model has been developed to estimate the number of components and the number of convolutive parameters. In addition, the window length used in the STFT has been taken advantage to match the characteristics of the audio signals. It is shown that for the mixture containing sources that exhibit percussive-like characteristics, a short-time processing window will extract these sources more efficiently. Conversely, a larger processing window is more suitable for pitch-like sources. The efficacy of the proposed algorithm has been demonstrated on synthetic and live recording of underdetermined convolute mixture. Results have shown that the proposed algorithm is very promising, considerable more flexible and offers a considerable better approach to the EM- and MU-based NMF, or NTF.

Table VIII: Effects of the initialization on the proposed separation algorithm.

Initialization			**Random**	**SVD [47]**	**Gamma-Exponential Process**
Synthetic Convolutive with drum	SDR	s_1	6.59	7.12	7.95
		s_2	-5.41	-0.98	0.22
		s_3	3.32	3.88	3.90
		Avg	1.50	3.34	4.02
Synthetic convolutive without drum	SDR	s_1	2.22	1.85	4.75
		s_2	2.67	3.33	3.93
		s_3	1.00	4.75	4.75
		Avg	1.96	3.31	4.48
Live recording with drum	SDR	s_1	3.21	6.47	6.98
		s_2	3.03	7.42	8.85
		s_3	7.87	8.11	8.92
		Avg	4.70	7.33	8.25
Live recording without drum	SDR	s_1	3.68	8.33	8.93
		s_2	-3.37	4.59	4.83
		s_3	-5.61	-0.23	3.18
		Avg	-1.77	4.23	5.65
Avg of all datasets	**SDR**		**1.60**	**4.55**	**5.60**

REFERENCES

[1] D. D. Lee, and H. S. Seung, "Algorithms for non-negative matrix factorization," in Advances in Neural Information Processing Systems 13, (2001), pp. 556-562.

[2] D. D. Lee, and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, Oct 21, 1999.

[3] D. Wang, X. Gao, and X. Wang, "Semi-Supervised Nonnegative Matrix Factorization via Constraint Propagation," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1-1, 2015.

[4] S. Nikitidis, A. Tefas, and I. Pitas, "Projected Gradients for Subclass Discriminant Nonnegative Subspace Learning," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2806-2819, Dec, 2014.

[5] Y. H. Xiao, Z. F. Zhu, Y. Zhao, Y. C. Wei, S. K. Wei, and X. L. Li, "Topographic NMF for Data Representation," *IEEE Transactions on Cybernetics*, vol. 44, no. 10, pp. 1762-1771, Oct, 2014.

[6] R. C. Zhi, M. Flierl, Q. 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, no. 1, pp. 38-52, Feb, 2011.

[7] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE*

Trans. Audio, Speech, Lang. Process, vol. 20 no. 4, pp. 1118–1133, May, 2012.

[8] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and Dynamic Source Separation Using Nonnegative Factorizations [A unifed view]," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66-75, May, 2014.

[9] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928-1936, Aug, 2012.

[10] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950-1960, Aug, 2012.

[11] S. Makino, T.-W. Lee, and H. Sawada, *Blind Source Separation*: New York: Springer, 2007.

[12] P. Comon, and C. Jutten, "Handbook of Blind Source Separation Independent Component Analysis and Applications," Academic Press, (2010), p. 856.

[13] Y. Xianchuan, H. Dan, and X. Jindong, "Blind Source Separation: Theory and Applications," JohnnWiley & Sons Singapore Pte. Ltd., (2014), p. 416.

[14] X. L. Zhang, and D. L. Wang, "A Deep Ensemble Learning Method for Monaural Speech Separation," *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 5, pp. 967-977, May, 2016.

[15] X. L. Zhang, and J. Wu, "Deep Belief Networks Based Voice Activity Detection," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 4, pp. 697-710, Apr, 2013.

[16] Y. X. Wang, and D. L. Wang, "Towards Scaling Up Classification-Based Speech Separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 7, pp. 1381-1390, Jul, 2013.

[17] R. Zdunek, "Improved Convolutive and Under-Determined Blind Audio Source Separation with MRF Smoothing," *Cognitive Computation*, vol. 5, no. 4, pp. 493-503, Dec, 2013.

[18] K. Takeda, H. Kameoka, H. Sawada, S. Araki, S. Miyabe, T. Yamada, and S. Makino, "Underdetermined BSS With Multichannel Complex NMF Assuming W-Disjoint Orthogonality of Source," in IEEE Region 10 Conference Tencon, 2011, pp. 413-416.

[19] A. Ozerov, and C. Fevotte, "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 3, pp. 550-563, Mar, 2010.

[20] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 7, pp. 1830-1840, Sep, 2010.

[21] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-Determined Reverberant Audio Source Separation Using Local Observed Covariance and Auditory-Motivated Time-Frequency Representation," in 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10), 2010, pp. 73-80.

[22] B. J. King, and L. Atlas, "Single-Channel Source Separation Using Complex Matrix Factorization," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 8, pp. 2591-2597, Nov, 2011.

[23] Q. Wang, W. L. Woo, and S. S. Dlay, "Informed Single-Channel Speech Separation Using HMM-GMM User-Generated Exemplar Source," *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 12, pp. 2087-2100, Dec, 2014.

[24] P. Parathai, W. L. Woo, S. S. Dlay, and B. Gao, "Single-channel blind separation using L1-sparse complex non-negative matrix factorization for acoustic signals," *Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. El124-El129, Jan, 2015.

[25] N. Tengtrairat, Bin Gao, W.L. Woo and S.S. Dlay, "Single-Channel Blind Separation using Pseudo-Stereo Mixture and Complex 2-D Histogram," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 11, pp. 1722-1735, 2013.

[26] J. L. Yao, X. N. Yang, J. D. Li, and Z. Li, "An MRC Based Over-determined Blind Source Separation Algorithm," in 2010 IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), 2010, pp. 309-313.

[27] A. M. Darsono, G. Bin, W. L. Woo, and S. S. Dlay, "Nonlinear single channel source separation," in Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010 7th International Symposium on, 2010, pp. 507-511.

[28] A. Nesbit, E. Vincent, and M. D. Plumbley, "Benchmarking Flexible Adaptive Time-Frequency Transforms for Underdetermined Audio Source Separation," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), (2009), pp. 37-40.

[29] M. N. Schmidt, and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in 6th Intl. Conf. on Independent Component Analysis and Signal Separation (ICA '06), Charleston, USA, 2006, pp. 700–707.

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

[31] B. Gao, W. L. Woo, and S. S. Dlay, "Variational Regularized 2-D Nonnegative Matrix Factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 5, pp. 703-716, May, 2012.

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

[33] B. Gao, W. L. Woo, and B. W. K. Ling, "Machine Learning Source Separation Using Maximum A Posteriori Nonnegative Matrix Factorization," *IEEE Transactions on Cybernetics*, vol. 44, no. 7, pp. 1169-1179, Jul, 2014.

[34] A. Ozerov, C. Fevotte, R. Blouet, and J. L. Durrieu, "Multichannel Nonnegative Tensor Factorization with Structured Constraints for User-Guided Audio Source Separation," in 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011, pp. 257-260.

[35] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, Nov, 2004.

[36] 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, Mar, 2009.

[37] F. D. Neeser, and J. L. Massey, "Proper Complex Random-Processes with Applications to Information-Theory," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1293-1302, Jul, 1993.

[38] N. Tengtrairat, W.L. Woo, S.S. Dlay, and B. Gao, "Online Noisy Single-Channel Source Separation Using Adaptive Spectrum Amplitude Estimator and Masking," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1881-1895, Apr 1, 2016.

[39] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," in International Conference on Machine Learning (ICML), 2010, pp. 439-446.

[40] V. Y. F. Tan, and C. Fevotte, "Automatic Relevance Determination in Nonnegative Matrix Factorization with the beta-Divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592-1605, Jul, 2013.

[41] B. Jørgensen, *Statistical properties of the generalized inverse Gaussian distribution*, New York: Springer-Verlag, 1982.

[42] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462-1469, Jul, 2006.

[43] K. Adiloglu, H. Kayser, and L. Wang. "A variational inference based source separation approach for the separation of sources in underdetermined recording," (2013); http://www.onn.nii.ac.jp/sisec13/evaluation_result/UND/submission/ob/Algorithm.pdf (date last viewed 01/06/15).

[44] H. Sawada, S. Araki, and S. Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 3, pp. 516-527, Mar, 2011.

[45] K. Adiloglu, and E. Vincent, *"Variational Bayesian interference for source separation and robust feature extraction*, Tech. Rep. RT-0428, Inria, Augest 2012.

[46] C. Knapp, and G. C. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320-327, 1976.

[47] C. Boutsidis, and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350-1362, Apr, 2008.

Ahmed Al-Tmeme received the BSc and MSc degrees in electrical\ computer and control engineering from Baghdad University, Iraq, in 2000 and 2002, respectively, and he is currently working toward the Ph.D. degree at the School of Electrical and Electronics Engineering, Newcastle University, UK. He is on study leave from the Al-Khwarizmi College of Engineering, University of Baghdad, Iraq. His research interests include blind and informed source separation, signal processing, and machine learning.

W. L. Woo 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 Reader in Intelligent Signal Processing. His major research is in the mathematical theory and algorithms for nonlinear signal and image processing. He has published over 250 papers on these topics on various journals and international conference proceedings. Dr Woo is a senior member of the IEEE and member of the Institution Engineering Technology (IET).

S. S Dlay received his BSc (Hons.) degree in Electrical and Electronic Engineering and his PhD in VLSI Design from the Newcastle University. In 1986 he re-joined the Newcastle University as a Lecturer in the School of Electrical, Electronic and Computer Engineering and later appointed to a Personal Chair in Signal Processing Analysis. He has published over 250 research papers ranging from biometrics and security, biomedical signal processing and implementation of signal processing architectures. Professor Dlay is a College Member of the EPSRC.

Bin Gao received the B.S. degree in communications and signal processing from Southwest Jiao Tong University, Chengdu, China, in 2005, the M.Sc. (Hons.) degree in communications and signal processing, and the Ph.D. degree from Newcastle University, Newcastle, U.K., in 2011. He is currently an Professor with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu.