IET Signal Processing

Research Article

Journals
The Institution of
Engineering and Technology

Two-stage blind audio source counting and separation of stereo instantaneous mixtures using Bayesian tensor factorisation

ISSN 1751-9675
Received on 20th October 2014
Revised on 9th March 2015
Accepted on 23rd March 2015
doi: 10.1049/iet-spr.2014.0404
www.ietdl.org

Sayeh Mirzaei[1] ✉, *Yaser Norouzi*[2], *Hugo Van Hamme*[1]

[1]*Department of Electrical Engineering, Amirkabir University, Tehran, Iran*
[2]*Department of Electrical Engineering, Amirkabir University, Tehran, Iran*
✉ *E-mail: smirzaei@esat.kuleuven.be*

Abstract: In this paper, the authors address the tasks of audio source counting and separation for two-channel instantaneous mixtures. This goal is achieved in two steps. First, a novel scheme is proposed for estimating the number of sources and the corresponding channel intensity difference (CID) values. For this purpose, an angular spectrum is evaluated as a function of the ratio of the magnitude spectrogram of the two channels and the peak locations of that spectrum are obtained. In the second stage, a new approach is developed for extracting the individual source signals exploiting a Bayesian non-parametric modelling. The mean field variational Bayesian approach is applied for inferring the unknown parameters. Classification is then performed on the inferred active CID values to obtain the individual source magnitude spectrograms. This way, the number of spectral components used for modelling each source is found automatically from the data. The Bayesian approach is compared with the standard Kullback–Leibler non-negative tensor factorisation method to illustrate the effectiveness of Bayesian modelling. The performance of the source separation is measured by obtaining the existing metrics for multichannel blind source separation evaluation. The experiments are performed on instantaneous mixtures from the dev2 database.

1 Introduction

Audio source separation is still a challenging task that is relevant in several fields such as polyphonic music separation, automatic meeting transcription and speech recognition. Generally, the channel characteristics between the sources and the sensors are unknown and we are dealing with a so-called blind source separation (BSS) task. Underdetermined mixtures are more often separated using time–frequency masking techniques [1] or classical sparse approaches such as [2–4]. Much research has been driven using non-negative matrix factorisation (NMF) to decompose the mixture audio signal into its spectral components from single channel measurements [5–7]. The data matrix is usually constructed based on the magnitude or power spectrogram of the audio mixture signal. NMF is applied for approximating the data matrix $\boldsymbol{X}$ by a product of two non-negative matrices $\boldsymbol{W}$ and $\boldsymbol{H}$. The columns of $\boldsymbol{W}$ specify the spectral components and the rows of $\boldsymbol{H}$ represent the time activations of these components at different time frames. The well-known multiplicative update (MU) rules are used for solving this optimisation problem [8]. These rules have been derived for different measures to express the discrepancy between the data matrix $\boldsymbol{X}$ and the model $\boldsymbol{W} \times \boldsymbol{H}$ such as the generalised Kullback–Leibler (KL) divergence, the Itakura–Saito (IS) distance or the Euclidean distance. The probabilistic extension of NMF is accomplished by presuming a generative probability distribution for the data and finding the maximum-likelihood (ML) solution through the generalised EM algorithm. For some specific choices of data distribution, this is equivalent to the standard NMF solution given by the MU rules. For instance, the Poisson assumption for the data distribution is equivalent to KL divergence [9]. If in addition a suitable prior distribution is assumed for the unknown parameters $\boldsymbol{W}$ and $\boldsymbol{H}$, we deal with a Bayesian modelling of the data [10]. In this case, maximum a posteriori (MAP) estimation of the parameters needs the exact derivation of the parameter posteriors. Since this is not tractable in many cases, variational Bayesian techniques are the most promising methods for approximating the posterior. Subsequently, the optimal values of the parameters are inferred as the expectation under the approximated variational posterior distribution. With a suitable choice of the prior distributions, Bayesian modelling can outperform the ML solution.

In the case of multichannel recordings, the channel mixture coefficients are an additional unknown parameter set which is to be estimated. In [11], a probabilistic non-negative tensor factorisation (NTF) approach is applied for obtaining the channel mixing coefficients along with the spectral components and the time activation matrices corresponding to individual sources. Each source's STFT is described by a generative model of superimposed *Gaussian* components, which is equivalent to taking the IS divergence as the difference measure in standard NTF, whereas the EM algorithm is equivalent to the MU rules. In [12], a PARAFAC-NTF approach is considered on both the absolute value and power of the signal's STFT, considering KL and IS divergence measures, respectively. These two scenarios are then compared with their probabilistic counterparts. In [13], an improved version of shift-invariant tensor factorisation has been proposed for musical sound source separation. In [14], a source separation algorithm is proposed based on NTF which uses a known spatial cue. The above-mentioned techniques have in common that they need to cluster the obtained spectral components and associate them to different audio sources.

One shortcoming of the methods discussed above is that they assume the number of sources as well as the total number of spectral components for modelling the sources is known in advance; however, this is not the case for many applications. Non-parametric Bayesian modelling techniques have been introduced for fixing this issue in single channel source separation [15–17]. The major assumption behind non-parametric modelling is the infinite number of latent components composing the data. However, in the inference stage, only a limited subset of these components remains active. This is achieved by choosing a sparse prior for the weight parameters specifying the gain of the latent components in the mixture. Here, we construct a proper non-parametric Bayesian model for the multichannel case to avoid having to assume a predefined number of spectral components for each source. In [18], a so-called permutation-free infinite sparse factor analysis, based on a non-parametric Bayesian framework is introduced that enables inference without a pre-determined number of sources. However, these and similar methods based on infinite independent component analysis are just applicable when the number of sensors is larger or at least equal to the number of sources. Here we propose a more general approach which can also be applied to the underdetermined BSS.

Despite the fact that the instantaneous audio mixtures are rarely encountered in real-world due to multipath effect and delayed reception, many papers have considered the instantaneous scenario [19–22]. The instantaneous mixture assumption is reasonable when *coincident microphone arrays* are applied [23–25]. In this case, we can disregard the phase difference because the microphones lie at the same position; however, the received signal amplitudes are not the same due to the different directivity patterns of the microphones. Therefore, using coincident boundary microphones we can configure an acoustic mixture system accomplishing the instantaneous mixture model conditions. Instantaneous mixtures are also common in synthetically mixed music.

In this paper, we partition the task of source separation into two steps. In the first step, the number of sources and the channel intensity difference (CID) values are estimated by evaluating a metric as a function of the ratio (represented as the tangent of an angle) of the magnitude spectrograms of both channels, and this for every time–frequency (TF) cell of the mixture signal STFT. An angular spectrum is calculated applying a pooling operation over time and frequency. The number of sources can then be estimated by finding the peak locations of the angular spectrum subject to some constraints. Enforcing these constraints can eliminate spurious peaks produced due to mismatch between model and data such as sensor noise. Besides inferring the number of sources, the proposed approach enables us to avoid clustering schemes for obtaining the channel coefficients, which is advantageous in practice since clustering is a delicate procedure [13]. We relax the sparsity assumption in the sense that we require a sufficiently large number of sparse TF cells in the analysis window, such that clear peaks emerge in the angular spectrum. Another contribution of this paper consists of proposing a novel approach for estimating the number of the sources and channel coefficients in advance.

For the second step, we propose a Bayesian non-parametric NTF modelling scheme to decompose the magnitude spectrograms into channel coefficients, spectral components and time activation matrices. The combination of Bayesian and NTF modelling is introduced for the first time here and makes our developed method different from what is followed in [9–12]. The elements of the spectral components and the time activation matrices are assumed to have a *Gamma* prior distribution. The concepts are similar to our proposed Bayesian non-parametric factorisation method for model order estimation [26] and can be regarded as an NTF extension of that model; in the present paper, we develop a non-parametric Bayesian model for multichannel recordings that enable us to infer the number of spectral components required for modelling individual sources automatically. This goal is achieved by presuming a large value for the total number of spectral components. After estimating the parameters through an iterative variational Bayesian procedure, we can group the components into sources by classifying the inferred channel coefficient ratios (intensity difference values), where the centroids obtained in the first step are used for classification. Derivation of the required parameter posterior update relations has been done for the first time in this paper. The obtained magnitude spectrogram of the sources is enriched with the phase of the original mixture STFT to derive their complex STFT. The individual time-domain source image signals can then be obtained by inverse STFT on each channel. In the experiments, the effectiveness of the proposed Bayesian scheme is demonstrated through comparison with the standard KL-NTF method.

The novel aspects of our proposed approach can be summarised as follows:

- Estimation of the number of the sources through introducing the angular spectrum and also estimation of the channel coefficients.
- Development of the Bayesian NTF framework and deriving the model parameters' update relations.
- Inferring the model order (total number of components) through imposing sparsity to the channel coefficients.

The rest of this paper is organised as follows. The CID estimation and the source counting scheme are described in Section 2. Section 3 is devoted to the Bayesian non-parametric model for multichannel source separation and to the derivation of the variational Bayesian update equations. Section 4 presents the experiments and a discussion on the results. Conclusions are drawn in Section 5.

2 Estimation of the CID and the number of sources

Generally, the channel mixture coefficients are estimated based on a sparsity assumption for audio signals [2, 27]. This means that just one source is assumed active in each TF cell. Therefore, the channel coefficients are estimated for each TF cell and the final solution is obtained by applying a clustering scheme on the histogram of these estimates. The main drawback of these methods is that they are very sensitive to the sparsity assumption. Moreover, clustering schemes are generally sensitive to the initial choice of the centroids. Moreover, the number of sources needs to be known. In [28, 29], less strict sparsity is assumed by detecting TF regions where one source is dominant over others. In [28], a local confidence measure is obtained in a given TF region and a clustering algorithm called DEMIX, based on the confidence measure, is proposed for counting and locating sources.

Motivated by the above-mentioned disadvantages, we do not deal with clustering-based methods in this paper. Instead, our proposed approach computes an angular spectrum from CID values obtained for each TF cell. The CID values and the number of sources are then estimated by finding the peak locations of this spectrum.

The instantaneous mixture signal in the STFT domain can be written as

$$X_{ift} = \sum_{j=1}^{J} a_{ij} S_{jft} + n_{ift}, \quad i = 1,\ 2 \tag{1}$$

where X_{ift} denotes the complex value of the mixture signal STFT in frequency bin f and time frame t for the ith channel and a_{ij} represents the channel mixing coefficient which takes a real positive value. S_{jft} is the complex contribution of each source in each TF bin and J is the total number of sources. n_{ift} represents additive noise.

We define the CID measure as follows

$$\phi(f,\ t) = 2\tan^{-1}\left(\frac{\left|X_{2ft}\right|}{\left|X_{1ft}\right|}\right) \tag{2}$$

Then, we evaluate a metric against angles θ corresponding to CID values, using equal spacing in the interval $[0,\ \pi]$

$$R(f,\ t,\ \theta) = \cos\big(\phi(f,\ t) - \theta\big) \tag{3}$$

For increasing the angular resolution, a monotonically decreasing non-linear function in the range $[0,\ 1]$ is applied to this metric, a technique that is inspired by direction-of-arrival estimation [30]

$$M(f,\ t,\ \theta) = 1 - \tanh\left(\alpha\sqrt{1 - R(f,\ t,\ \theta)}\right) \tag{4}$$

where α is the non-linearity parameter. Using this non-linear function is a computationally cheap method to sharpen the peaks corresponding to the true source CIDs. To obtain the final angular spectrum $F(\theta)$, a summation over all frequency bins and a maximisation over all time frames are performed

$$F(\theta) = \max_t \sum_f M(f,\ t,\ \theta) \tag{5}$$

To purify the angular spectrum, we can obtain it based on the TF cells which more probably correspond to one dominant active source. This way, we can enhance the true peak levels corresponding to the actual source CIDs and consequently diminish the effect of the spurious peaks. For identifying the mentioned cells, we introduce the following weighting function

$$\lambda(f,\ t) = \begin{cases} 1, & \text{if}\Big(\cos\left(\angle X_{2ft} - \angle X_{1ft}\right) > \beta_1\Big) \\ & \wedge\Big(\Big(\left|X_{1ft}\right| + \left|X_{2ft}\right|\Big) > \beta_2\Big) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where β_1 is set to 0.99 and $\beta_2 = \max_t(\mathrm{mean}_f(|X_{1ft}| + |X_{2ft}|))$.

The first condition in (6) is aimed to recognise the cells with one dominant source and the second one eliminates the contribution of the cells with a low magnitude. The metric expressed in (4) is then multiplied by this weighting to provide an improved angular spectrum. The effect of the weighting will be illustrated in the experiments in Section 4. The tasks of CID estimation and source counting are accomplished by exploiting a peak finding algorithm. First, the minimum value of the angular spectrum $F(\theta)$ is subtracted and then it is normalised, that is, the vector is divided by its maximum value. Afterwards, two constraints on the minimum distance between the peaks and minimum peak height eliminate the irrelevant peak locations found by the peak finder algorithm. To choose values for these thresholds, we studied the behaviour of the peak finding algorithm in a setting where the time-domain observations are corrupted by Gaussian noise at an signal-to-noise ratio >15 dB (note this scenario differs from our evaluation setup in Section 4). From this paper, we found that thresholds for minimum peak height of 0.55 and for minimum peak distances of 5° gave satisfactory performance. The retained peak locations will be referred to below as *estimated source CID* values.

Note that the peak heights are not proportional to the source signal strengths, but to the number of TF cells in which a source is dominant. This can be seen from 5: the derived two-dimensional spectrum is computed as the sum over all frequencies of values between 0 and 1, measuring agreement of data and hypothesis, and this at the time t where the source can produce the best match. This clarifies our choice of threshold for the minimal peak height; implying sources should at some point in time be dominant in at least 55% of the number of cells compared with the most dominant source. The minimal angular distance between peaks is set based on the practical consideration that sources are spatially distributed.

3 Source separation

Most of the BSS techniques work based on the sparsity assumption for the audio sources. Therefore, the dominant source in each bin is identified and then individual binary masks are applied to segregate the source signals [1, 27]. A common issue with binary TF masking methods is the musical noise artefacts which degrade the separation performance. This noise consists of short tones randomly distributed in the separated signal over time and frequency. Furthermore, the sparsity is, in general, not a valid assumption for mixtures in music. Therefore, a benefit of the proposed source separation approach is it does not involve an explicit step of constructing a binary TF mask based on this assumption. Cells where the sparsity assumption is violated will not contribute to the peaks in the angular spectrum $F(\theta)$. As long as there are sufficient numbers of sparse contributions, $F(\theta)$ will show peaks at the correct locations. Each source is subsequently modelled by a sum of components each described by spectral patterns, time activations and channel mixing coefficients. The attribution of a component to a source is achieved by classification of its channel mixing coefficient ratios (IDQ – see Section 3.3) to one of the estimated source CID values.

3.1 Standard KL-NTF framework

To decompose the magnitude spectrogram of the audio mixture signal $Y_{ift} = |X_{ift}|$ into K spectral components in the multichannel case, we can consider a standard NTF framework with KL difference measure as follows

$$\begin{aligned} &\min_{\boldsymbol{Q},\boldsymbol{W},\boldsymbol{H}} \sum_{i,f,t} D_{KL}(Y_{ift}|\hat{Y}_{ift}), \quad \text{subject to}\,\boldsymbol{Q},\ \boldsymbol{W},\ \boldsymbol{H} \geq \boldsymbol{0} \\ &\hat{Y}_{ift} = \sum_{k=1}^{K} Q_{ik} W_{fk} H_{kt} \end{aligned} \tag{7}$$

where the columns of the $\boldsymbol{W}$ matrix denote the spectral components, the $\boldsymbol{H}$ elements specify the time activations and the $\boldsymbol{Q}$ matrix is representative of the channel mixing coefficients corresponding to each component. For solving the above optimisation problem, MU rules have already been derived [12, 31].

3.2 Proposed Bayesian NTF framework

To avoid overfitting and hence obtaining more robust performance, a Bayesian extension of NTF is derived here. We assume the following Bayesian generative model

$$\begin{aligned} &Y_{ift} \sim \delta(Y_{ift} - \sum_{k=1}^{K} C_{ikft}), \quad i = 1,\ 2 \\ &f = 1,\ \ldots,\ F\, t = 1,\ \ldots,\ T \\ &C_{ikft} \sim \mathrm{Poisson}\Big(Q_{ik} W_{fk} H_{kt}\Big) \\ &W_{fk} \sim \mathrm{Gamma}\Big(a^w_{fk},\ b^w_{fk}\Big) \\ &H_{kt} \sim \mathrm{Gamma}\big(a^h_{kt},\ b^h_{kt}\big) \\ &Q_{ik} \sim \mathrm{Gamma}\big(a^q_{ik},\ b^q_{ik}\big) \end{aligned} \tag{8}$$

where the *Poisson* and *Gamma* distributions are defined, respectively, by $\mathrm{Poisson}(x|\lambda) = \mathrm{e}^{-\lambda}(\lambda^x)/(\Gamma(x+1))$ and $\mathrm{Gamma}(x|\alpha, \beta) = [\beta^{\alpha}\Gamma(\alpha)]^{-1} x^{\alpha\ -1} \mathrm{e}^{-x/\beta}$, $x \geq 0$, $\alpha\ > 0$, $\beta\ > 0$, respectively. We refer to α as the shape parameter and β as the scale parameter. C_{ikft} denotes the latent components in our model which are assumed to follow a *Poisson* distribution. The elements of the $\boldsymbol{W}$, $\boldsymbol{H}$ and $\boldsymbol{Q}$ matrices are assumed to have a *Gamma* prior distribution. The *Gamma* distribution is the conjugate prior for *Poisson* distribution, a choice which leads to simplification of the inference procedure. The *Gamma* distribution is also selected because it can impose sparsity on the inferred parameters. The evaluation in Section 4 will reveal that these assumptions are acceptable in the sense that they lead to strong results on data from a competitive evaluation campaign. By introducing the above generative model, we aim to provide a method for automatically inferring the number of spectral components required for modelling each source. Therefore, the total number of components, K, is set to a large number. Owing to the choice of a sparse prior for the elements of $\boldsymbol{Q}$, the remaining active components after inference will be limited. The active components can be specified by observing the l_1-norm of the inferred model components on the first channel which is defined as $\zeta_k = \sum_{f,t} Q_{1k} W_{fk} H_{kt}$ for $k = 1 \ldots K$ and discarding the components with near-zero l_1-norm. This Bayesian NTF framework can also be regarded as an extension of the Bayesian NMF model proposed in [10].

MAP estimation of the parameters is not straightforward since the exact posterior expression is intractable. Variational Bayesian inference is a common method for finding optimal parameters in this case. This approach tries to obtain an analytic approximation of the posterior distribution of the parameters given the data by maximising a lower bound of the data likelihood. Here, we use the mean field variational scheme which presumes a factorised form for the approximate posterior [32]. If $\boldsymbol{Z}$ denotes the unknown parameters in the model, these parameters are partitioned into M groups specified by $\boldsymbol{Z}_i$, $i = 1, 2, \ldots, M$. The variational posterior is then chosen with the following factorised form

$$g(\boldsymbol{Z}) = \prod_{i=1}^{M} g_i(\boldsymbol{Z}_i) \tag{9}$$

The log-likelihood of data can be written as

$$\ln p(\boldsymbol{X}) = \int g(\boldsymbol{Z}) \ln\left(\frac{p(\boldsymbol{X}, \boldsymbol{Z})}{g(\boldsymbol{Z})}\right) \mathrm{d}\boldsymbol{Z} - \int g(\boldsymbol{Z}) \ln\left(\frac{p(\boldsymbol{Z}|\boldsymbol{X})}{g(\boldsymbol{Z})}\right) \mathrm{d}\boldsymbol{Z} \tag{10}$$

where $\boldsymbol{X}$ denotes the observed data and the integration domain is the set of allowable parameter values for Z. The first term in (10) specifies a lower bound on the log-likelihood. The aim of the variational inference algorithm is to find an approximate posterior $g(\boldsymbol{Z})$ which maximises this lower bound. Then, the optimal values of the parameters are obtained as the expectation of the individual parameters with respect to the variational posterior distribution $g(\boldsymbol{Z})$. The optimal solution for the variational distribution of each group, $g_j(\boldsymbol{Z}_j)$, can be expressed as [32]

$$g_j^*\left(\boldsymbol{Z}_j\right) = \frac{\exp\left(\langle \ln p(\boldsymbol{X}, \boldsymbol{Z})\rangle_{\overline{g_j(\boldsymbol{Z}_j)}}\right)}{\int \exp\left(\langle \ln p(\boldsymbol{X}, \boldsymbol{Z})\rangle_{\overline{g_j(\boldsymbol{Z}_j)}}\right) \mathrm{d}\boldsymbol{Z}_j} \tag{11}$$

where $\langle\rangle_{\overline{g_j(\boldsymbol{Z}_j)}}$ denotes the expectation with respect to the variational distribution of all other groups, $g_i(\boldsymbol{Z}_i)$, $i \neq j$. On the basis of (11), the variational distributions of the four unknown parameter groups $\boldsymbol{C}$, $\boldsymbol{Q}$, $\boldsymbol{W}$ and $\boldsymbol{H}$ are derived. The variational distribution of the components C_{ikft} is proportional to the following expression (see (12) at the bottom of the page)

which is in the form of a multinomial distribution with cell probabilities p_{ikft}

$$p_{ikft} = \frac{\exp\left(\langle \ln Q_{ik}\rangle_{\overline{g(\boldsymbol{C})}} + \langle \ln W_{fk}\rangle_{\overline{g(\boldsymbol{C})}} + \langle \ln H_{kt}\rangle_{\overline{g(\boldsymbol{C})}}\right)}{\sum_k \exp\left(\langle \ln Q_{ik}\rangle_{\overline{g(\boldsymbol{C})}} + \langle \ln W_{fk}\rangle_{\overline{g(\boldsymbol{C})}} + \langle \ln H_{kt}\rangle_{\overline{g(\boldsymbol{C})}}\right)} \tag{13}$$

The expectation of C_{ikft} with respect to this multinomial distribution is equal to $p_{ikft} Y_{ift}$.

Again using (11), the variational distributions of the parameters Q_{ik}, W_{fk} and H_{kt} are obtained. They have the form of a *Gamma* distribution

$$\begin{aligned} g\left(W_{fk}\right) &= \mathrm{Gamma}\left(\alpha_{fk}^{w}, \beta_{fk}^{w}\right) \\ \alpha_{fk}^{w} &= a_{fk}^{w} + \sum_{i,t} \langle C_{ikft}\rangle, \\ \beta_{fk}^{w} &= \left(\frac{1}{b_{fk}^{w}} + \sum_{i,t} \langle Q_{ik}\rangle\langle H_{kt}\rangle\right)^{-1} \end{aligned} \tag{14}$$

Similarly

$$\begin{aligned} g\left(H_{kt}\right) &= \mathrm{Gamma}\left(\alpha_{kt}^{h}, \beta_{kt}^{h}\right) \\ \alpha_{kt}^{h} &= a_{kt}^{h} + \sum_{i,f} \langle C_{ikft}\rangle, \\ \beta_{kt}^{h} &= \left(\frac{1}{b_{kt}^{h}} + \sum_{i,f} \langle Q_{ik}\rangle\langle W_{fk}\rangle\right)^{-1} \end{aligned} \tag{15}$$

and

$$\begin{aligned} g\left(Q_{ik}\right) &= \mathrm{Gamma}\left(\alpha_{ik}^{q}, \beta_{ik}^{q}\right) \\ \alpha_{ik}^{q} &= a_{ik}^{q} + \sum_{f,t} \langle C_{ikft}\rangle, \\ \beta_{ik}^{q} &= \left(\frac{1}{b_{ik}^{q}} + \sum_{f,t} \langle W_{fk}\rangle\langle H_{kt}\rangle\right)^{-1} \end{aligned} \tag{16}$$

The $\langle\rangle$ operator denotes the expected value of the enclosed random variable. The logarithmic expectations with respect to a *Gamma* distribution which appear in (13) are given by the following expression:

$$v \sim \mathrm{Gamma}(\alpha, \beta) \Rightarrow \langle \ln(v)\rangle = \varphi(\alpha) + \ln(\beta) \tag{17}$$

where $\varphi(.)$ is the digamma function. The log-likelihood lower bound is calculated as

$$\begin{aligned} \mathrm{LB} &= \int g(\boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{H}) \ln\left[\frac{P(\boldsymbol{Y}, \boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{H})}{g(\boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{H})}\right] \mathrm{d}\boldsymbol{H}\, \mathrm{d}\boldsymbol{W}\, \mathrm{d}\boldsymbol{Q}\, \mathrm{d}\boldsymbol{C} \\ &= -\sum_{i,k,f,t} \langle Q_{ik}\rangle\langle W_{fk}\rangle\langle H_{kt}\rangle \\ &\quad + \sum_{i,k} \langle \ln\left(Q_{ik}\right)\rangle \left[a_{ik}^{q} - 1 + \sum_{f,t} \langle C_{ikft}\rangle\right] \\ &\quad + \sum_{f,k} \langle \ln\left(W_{fk}\right)\rangle \left[a_{fk}^{w} - 1 + \sum_{i,t} \langle C_{ikft}\rangle\right] \\ &\quad + \sum_{k,t} \langle \ln\left(H_{kt}\right)\rangle \left[a_{kt}^{h} - 1 + \sum_{i,f} \langle C_{ikft}\rangle\right] \\ &\quad + \sum_{i,k} \left[\frac{-\langle Q_{ik}\rangle}{b_{ik}^{q}} - \ln\Gamma\left(a_{ik}^{q}\right) - a_{ik}^{q} \ln\left(b_{ik}^{q}\right)\right] \\ &\quad + \sum_{f,k} \left[\frac{-\langle W_{fk}\rangle}{b_{fk}^{w}} - \ln\Gamma\left(a_{fk}^{w}\right) - a_{fk}^{w} \ln\left(b_{fk}^{w}\right)\right] \\ &\quad + \sum_{k,n} \left[\frac{-\langle H_{kt}\rangle}{b_{kt}^{h}} - \ln\Gamma\left(a_{kt}^{h}\right) - a_{kt}^{h} \ln\left(b_{kt}^{h}\right)\right] \\ &\quad + \sum_{i,f,t} \left[-\log\Gamma\left(Y_{ift} + 1\right) - \sum_{k} \langle C_{ikft}\rangle \ln p_{ikft}\right] \\ &\quad + \sum_{f,k} \left[-\left(\alpha_{fk}^{w} - 1\right)\varphi\left(\alpha_{fk}^{w}\right) + \ln\beta_{fk}^{w} + \alpha_{fk}^{w} + \ln\Gamma\left(\alpha_{fk}^{w}\right)\right] \\ &\quad + \sum_{i,k} \left[-\left(\alpha_{ik}^{q} - 1\right)\varphi\left(\alpha_{ik}^{q}\right) + \ln\beta_{ik}^{q} + \alpha_{ik}^{q} + \ln\Gamma\left(\alpha_{ik}^{q}\right)\right] \\ &\quad + \sum_{k,t} \left[-\left(\alpha_{kt}^{h} - 1\right)\varphi\left(\alpha_{kt}^{h}\right) + \ln\beta_{kt}^{h} + \alpha_{kt}^{h} + \ln\Gamma\left(\alpha_{kt}^{h}\right)\right] \end{aligned} \tag{18}$$

$$g(C_{ikft}) \propto \delta\left(Y_{ift} - \sum_k C_{ikft}\right) \exp\left(\sum_k C_{ikft}\left(\langle \ln Q_{ik}\rangle_{\overline{g(\boldsymbol{C})}} + \langle \ln W_{fk}\rangle_{\overline{g(\boldsymbol{C})}} + \langle \ln H_{kt}\rangle_{\overline{g(\boldsymbol{C})}}\right) - \ln\Gamma(C_{ikft} + 1)\right) \tag{12}$$

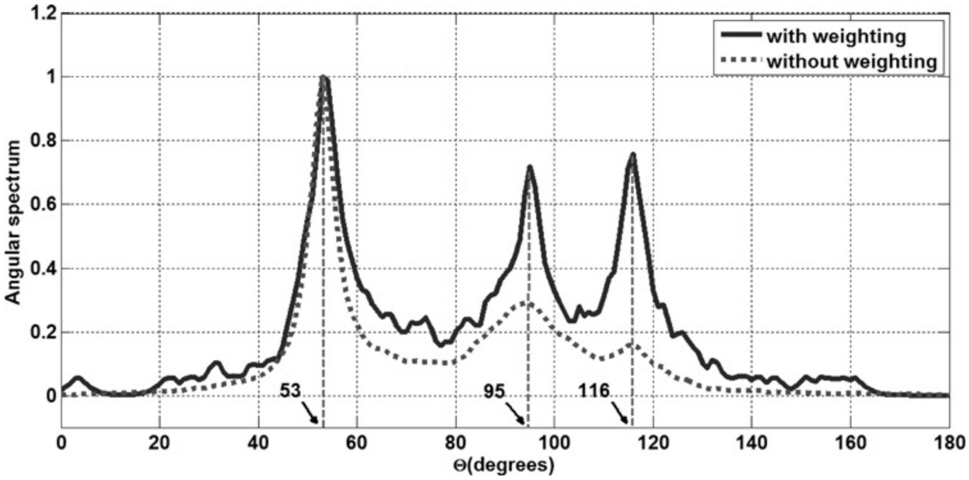

Fig. 1 *Angular spectrum*

After initialising the parameters by randomly drawing from the prior and iteratively running the above coupled update equations, the log-likelihood lower bound increases monotonically till convergence. The optimal parameters $\boldsymbol{W}$, $\boldsymbol{H}$ and $\boldsymbol{Q}$ are then taken to be the expectation with respect to the variational distributions, that is, the product of the shape and scale parameter of their *Gamma* distribution. Convergence is achieved when the relative increase of the lower bound falls below a threshold.

3.3 Extracting individual source signals

The estimated $\boldsymbol{Q}$ parameters are classified based on the estimated CID values in the previous step. This can be performed by first evaluating the inferred intensity difference measures IDQ corresponding to $\boldsymbol{Q}$ coefficients as $\text{IDQ}_k = 2\tan^{-1}(Q_{2k}/Q_{1k})$, $k = 1\dots K$. The IDQ values are classified to J groups where J is the estimated number of sources. This grouping is done by associating each IDQ element to the nearest (in Euclidean distance) CID value (centroid) obtained in the first step. For the standard NTF framework, the classification is done on all of the estimated components. Subsequently, the complex spectrogram of the jth source spatial image on channel i, S^{Im}_{ijft}, is given by the following relation

$$S^{\text{Im}}_{ijft} = \frac{\sum_{K_j} Q_{ik} W_{fk} H_{kt}}{\sum_{k=1}^{K} Q_{ik} W_{fk} H_{kt}} X_{ift} \tag{19}$$

where K_j denotes the subset of components associated to the source j. The phase of the mixture signal is exploited to retrieve the source contributions.

For Bayesian NTF, the individual complex spectrogram of the jth source is derived as

$$S^{\text{Im}}_{ijft} = \sum_{K_j} p_{ikft} X_{ift}, \quad i = 1,\ 2 \quad j = 1\dots J \tag{20}$$

Then, the individual source signals in time domain are obtained by applying inverse STFT operation to (19) or (20).

4 Experiments

The experiments are performed on instantaneous stereo mixtures from dev2 dataset of SiSEC'08 'underdetermined speech and music mixtures' task [33]. The signals are constructed by combining three static sources scaled by real positive gains. Here, the original mixture signals are used and no noise is added. The sampling frequency is 16 kHz. The time duration of all individual sources is 10 s. The STFT is computed with half-overlapping sine windows of length 1024. The non-linearity parameter α is taken equal to 15. The *nodrums* data is a mixture of three non-percussive music sources and *wdrums* is the combination of three music sources including drums.

4.1 Effectiveness of applying weighting function for obtaining the angular spectrum

The angular spectrum is calculated for 180 uniformly spaced angles in the interval [0 π]. The angular spectrum is calculated for the nodrums instantaneous mixture. The number of sources is estimated correctly to be three using the peak finder algorithm. The estimated CID values are equal to [53 95 116] in degrees that perfectly match with the true CID of the instantaneous mixtures of dev2 database. Fig. 1 shows the effectiveness of applying the mentioned weighting scheme to derive the angular spectrum ($F(\theta)$) for the nodrums data. It can be observed that the peaks corresponding to the actual CID values of the sources are considerably enhanced. The success of the first stage implies that there are indeed enough TF cells where the relaxed sparsity assumption holds for both the wdrums and nodrums data. Finding the correct number of sources is a prerequisite for a good performance in the second stage.

4.2 Source counting and CID estimation

To evaluate our proposed source counting method, the rate of success of the algorithm in the estimation of the true number of sources is calculated and reported in Table 1. For this purpose, J sources with corresponding CID values uniformly spaced in the interval [0 π] are mixed together. This is done for J from 2 to 10. For each J, we generate ten different mixtures by randomly selecting the original signals from speech/music sources of the dev2 dataset. The results are compared with the results of the DEMIX-Inst method [34] which are obtained using the software provided in [35]. The success rate is calculated as the percentage of correct estimations out of ten trials. As can be observed, the algorithms perform perfectly up to five sources. For the case of

Table 1 Percentage of correct estimation of the number of sources

Number of sources	2	3	4	5	6	7	8	9	10
angular spectrum based	100	100	100	100	80	70	50	30	30
DEMIX-Inst	100	100	100	100	90	80	60	20	0

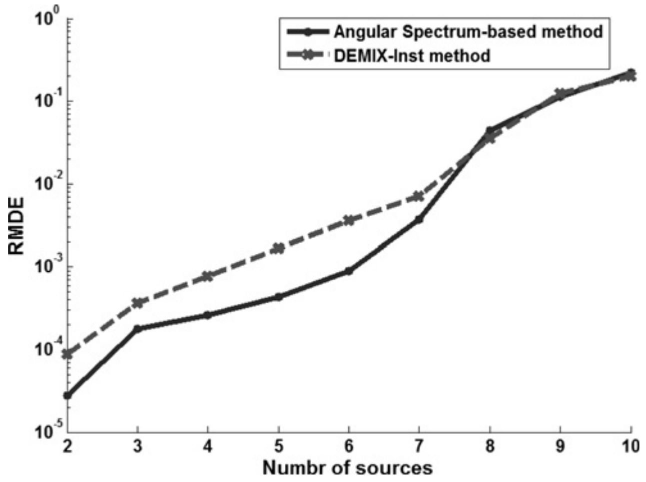

Fig. 2 *RMDE as a function of the number of sources*

erroneous estimations, our angular spectrum-based algorithm often finds more peaks than the actual number of sources.

To also evaluate the CID estimation performance, the mean direction error (MDE) is calculated similar to what is proposed in [34]. Given the true CID values $\mathbf{\Theta} = [\theta_1 \ldots \theta_J]$ and the estimated ones $\hat{\mathbf{\Theta}} = [\hat{\theta}_1 \ldots \hat{\theta}_J]$, the MDE is defined as

$$\mathrm{MDE}(\mathbf{\Theta},\ \hat{\mathbf{\Theta}}) = \min_{P \in S_J} \frac{1}{J} \sum_{j=1}^{J} \left| \theta_j - \hat{\theta}_{P(j)} \right| \tag{21}$$

where S_J is the permutation group of size J. Toward this goal, the J highest peaks of the spectrum found by the peak finder algorithm are considered as the estimated CID values of the sources. Similarly, for DEMIX-Inst algorithm, we set and fix the number of sources to J. To measure the error in terms of relative precision, the relative MDE (RMDE) is defined as

$$\mathrm{RMDE}(\mathbf{\Theta},\ \hat{\mathbf{\Theta}}) = \frac{\mathrm{MDE}(\mathbf{\Theta},\ \hat{\mathbf{\Theta}})}{\min\limits_{j \neq j'} \left| \theta_j - \theta_{j'} \right|} \tag{22}$$

where the denominator denotes the minimum distance between the true CID values. RMDE values are plotted in Fig. 2 against the number of sources. We have used 18,000 angle segments for evaluating an angular spectrum to achieve an angular resolution of 0.01° because we need enough resolution to plot and compare with the DEMIX-Inst scheme in Fig. 2. For the sake of comparability, we take two decimal places of precision from the DEMIX-Inst estimates. It can be seen that our proposed algorithm considerably outperforms DEMIX-Inst method up to six sources and shows nearly the same performance for more than six sources.

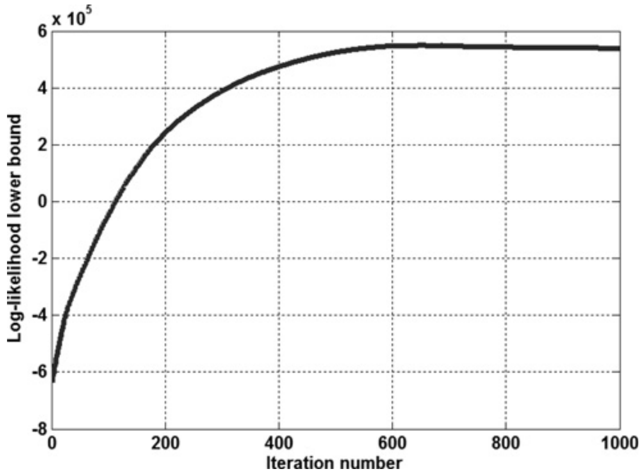

Fig. 4 *Bayesian NTF convergence behaviour*

4.3 Source separation assessment

For the second stage, the parameters of the Bayesian model are initialised by randomly drawing from the prior. The hyperparameters of the priors of $\boldsymbol{W}$, $\boldsymbol{H}$ and $\boldsymbol{Q}$ are chosen as $a^w_{fk} = b^w_{fk} = a^h_{kn} = b^h_{kn} = 1$, $a^q_{ik} = 0.5$, $b^q_{ik} = 1$. We normalise the $\boldsymbol{Y}$ elements to have a mean value of 1. The coupled update equations for variational distributions are iteratively executed as mentioned in Section 3.2. The convergence is verified by calculation of the log-likelihood lower bound relative increase rate in each iteration and checking if it is greater than a tolerance threshold which is taken equal to 10^{-6}. For avoiding local optima, we have executed the algorithm ten times and chosen the results corresponding to the largest likelihood lower bound. The total number of components K is set to 50. For KL-NTF, the initial values of the parameters are taken the same as in the Bayesian approach. The convergence check can also be accomplished as stated above by comparing the relative decrease rate of the KL difference measure with a tolerance threshold equal to 10^{-6} in each iteration.

The convergence behaviour of the KL-NTF and Bayesian NTF approaches derived for nodrums data is illustrated in Figs. 3 and 4, respectively. The optimisation criterion is plotted as a function of the iteration number. The major result of this analysis is that the

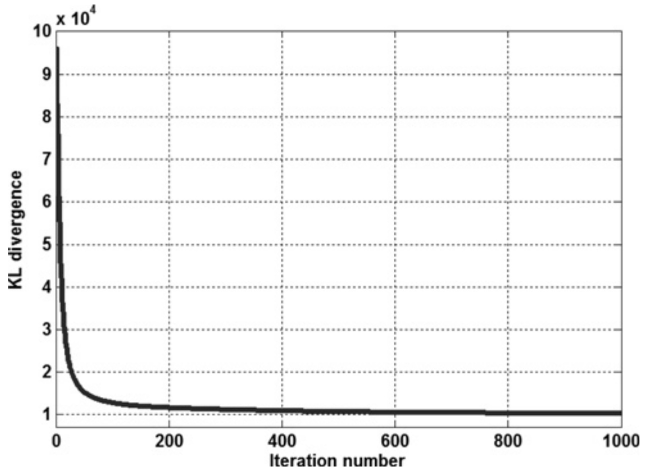

Fig. 3 *KL-NTF convergence behaviour*

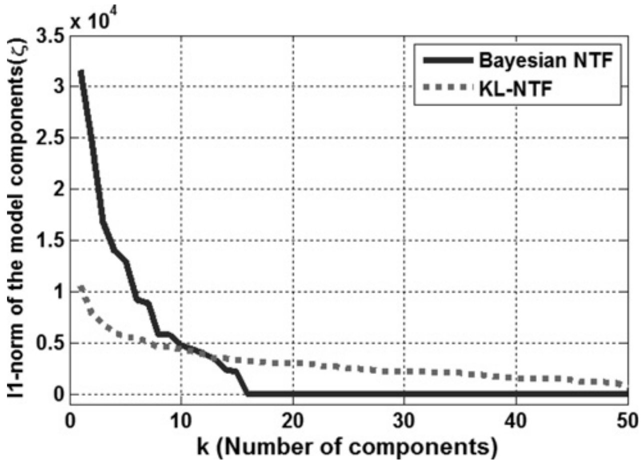

Fig. 5 *Sorted ζ elements for the nodrums mixture*

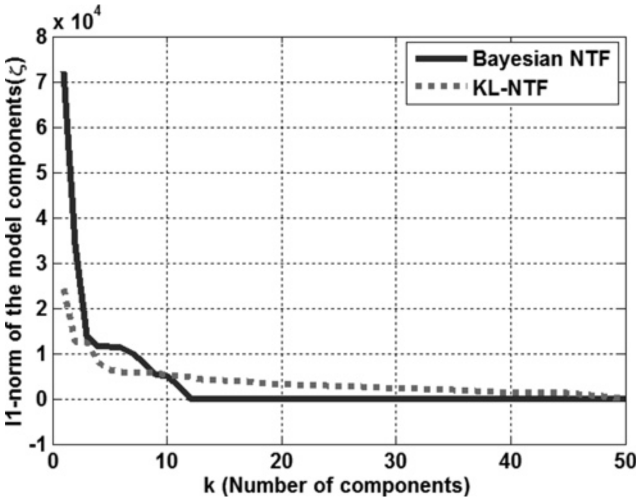

Fig. 6 *Sorted ζ elements for the wdrums mixture*

KL-NTF is sufficiently converged using <1000 iterations and proceeding with more iterations does not lead to better performance.

The sorted values of the ζ metric are depicted in Figs. 5 and 6 corresponding to the nodrums and wdrums data, respectively. As can be seen in Fig. 5, exploiting the proposed Bayesian NTF, 15 out of 50 components are active for the nodrums mixture, which is specified as near zero values of ζ for $k>15$. However using standard NTF, all of the components are utilised in the model. Applying Bayesian NTF, the number of active components for the wdrums mixture is 11 as implied from Fig. 6. Here, the components found by the KL-NTF method are again overfitted to the data and all of the components are exploited. Increasing the total number of components, K, has also led to the same number of active components and the same separation performance for Bayesian NTF. Using this Bayesian approach, we can avoid overfitting. The lower computational load forms another advantage of this non-parametric Bayesian approach over some Bayesian model order selection methods. In our case, we choose a large number as the total number of components and the model order is inferred by executing the variational inference algorithm once, whereas this is not the case for the model selection methods which operate based on likelihood evaluation as they need to run the inference algorithm for multiple values of the number of considered model components.

The active components for each mixture are then classified based on the estimated CID values as described in Section 3.3. For standard KL-NTF, the whole 50 components are classified. The original and separated source spatial image time-domain signals on the first channel are plotted in Fig. 7 for the nodrums mixture applying the proposed Bayesian NTF method. The same graphs are drawn in Fig. 8 demonstrating the separated sources achieved by applying the standard KL-NTF scheme. Similar plots are depicted for the

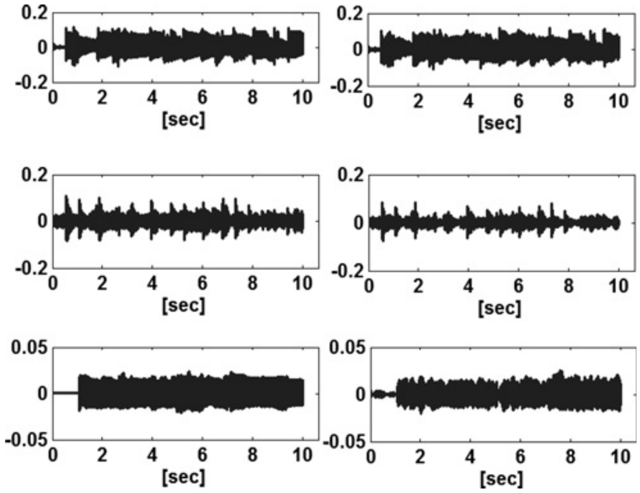

Fig. 7 *Original (left) and separated sources (right) for the nodrums mixture (Bayesian NTF)*

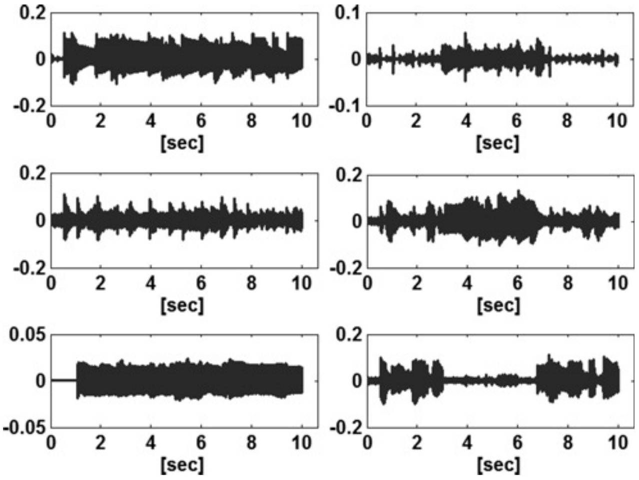

Fig. 8 *Original (left) and separated sources (right) for the nodrums mixture (standard KL-NTF)*

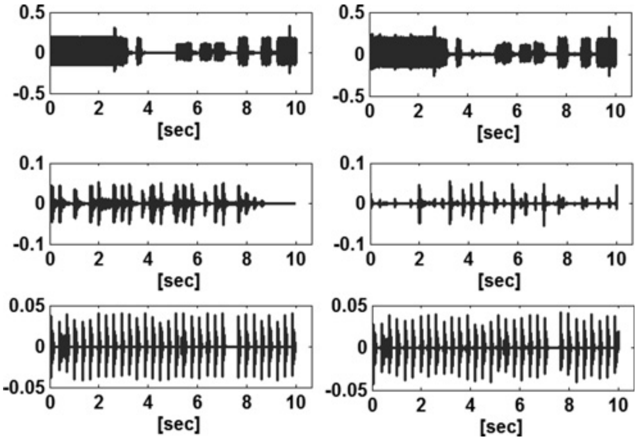

Fig. 9 *Original (left) and separated sources (right) for the wdrums mixture (Bayesian NTF)*

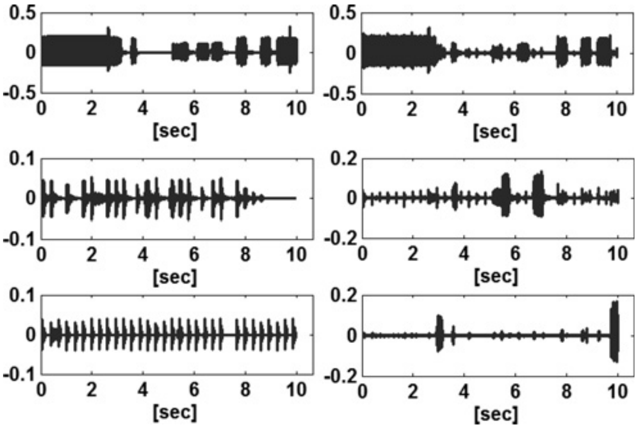

Fig. 10 *Original (left) and separated sources (right) for the wdrums mixture (Bayesian NTF)*

Table 2 BSS evaluation metrics obtained for nodrums

	Nodrums		
	*S*1(Bass)	*S*2(rhythmic guitar)	*S*3(lead guitar)
Bayesian NTF			
SDR, dB	15.8	6.6	4.2
ISR, dB	22.1	9.6	5.9
SIR, dB	17.6	10.5	6.8
SAR, dB	24.4	10.4	9.2
KL-NTF			
SDR, dB	−12.5	3.7	1.0
ISR, dB	4.8	4.4	1.4
SIR, dB	−13.9	14.8	3.0
SAR, dB	8.7	8.8	5.2
KL-NTF [12]			
SDR, dB	13.2	−1.8	1.0
ISR, dB	22.7	1.0	1.2
SIR, dB	13.9	−9.3	6.1
SAR, dB	24.2	7.4	2.6
KL-cNTF [12]			
SDR, dB	5.8	−9.9	3.1
ISR, dB	8.0	0.7	6.3
SIR, dB	13.5	−15.3	2.9
SAR, dB	8.3	2.7	9.9

wdrums mixture in Figs. 9 and 10. We observe that the images resemble the signals a lot more for the Bayesian NTF. Perceptual evaluation by persons not familiar with the work confirms this. Note that the performance differences between Bayesian NF and KL-NTF cannot be attributed to convergence issues since increasing the number of iterations did not result in significant changes.

To measure the quality of source separation objectively, the evaluation metrics for multichannel BSS [36] are calculated. The amount of spatial distortion, interference and artefacts are measured by three energy ratios expressed in decibels (dB): the source image-to-spatial distortion ratio (ISR), the signal-to-interference ratio (SIR) and the signal-to-artefacts ratio (SAR), respectively. The total error is also measured by the signal-to-distortion ratio (SDR). The obtained metrics for the nodrums instantaneous mixture are listed in Table 2. The Bayesian NTF performance is compared with the standard NTF solution. Substantial advantage of Bayesian NTF is revealed over standard KL-NTF. The evaluation criteria for the wdrums mixture are shown in Table 3. Again, the effectiveness of the Bayesian NTF can be observed. Furthermore, significant performance improvement is achieved for the nodrums compared with the results reported in [12], which we reproduce in Tables 1 and 2. In [12], the number of sources as well as the total number of components are assumed known and fixed parameters; KL-NTF [12] uses *k*-means clustering to associate the components to the sources, whereas KL-NTFc [12] assumes known number of components for modelling each source (equal), hence it does not need to classify the components because the clusters have been assumed known in advance.

To provide an additional comparison in terms of the average performance, we have applied the Bayesian NTF method to the instantaneous speech mixtures of the database as well. We typically find eight to nine components to model speech sources, which correspond quite well to the hand-tuned optimum found in [11]. Hence, another advantage of our proposed model order selection scheme over choosing a fixed order is revealed when we deal with mixtures containing both music and speech sources and we are to blindly accomplish source separation. It is clear that in these cases, it is not efficient to fix the number of model components for all sources. The calculated average evaluation metrics over all sources and all mixtures (music and speech) are

Table 3 BSS evaluation metrics obtained for wdrums

	wdrums		
	*S*1(hi-hat)	*S*2(drums)	*S*3(bass)
Bayesian NTF			
SDR, dB	9.8	7.1	17.7
ISR, dB	15.9	7.5	35.8
SIR, dB	14.1	13.9	18.2
SAR, dB	13.2	8.0	28.6
KL-NTF			
SDR, dB	9.5	−9.7	−4.1
ISR, dB	11.9	−0.3	1.3
SIR, dB	16.9	−16.5	−14.3
SAR, dB	14.4	−0.3	2.6
KL-NTF [12]			
SDR, dB	−0.2	0.4	17.9
ISR, dB	15.5	0.7	31.5
SIR, dB	1.4	−0.9	18.9
SAR, dB	7.4	−3.5	25.7
KL-cNTF [12]			
SDR, dB	−.02	−14.2	1.9
ISR, dB	15.3	2.8	2.1
SIR, dB	1.5	−15.0	18.9
SAR, dB	7.8	13.2	9.2

Table 4 Average evaluation metrics

Method	Bayesian NTF	SASSEC2007	SiSEC 2008	SiSEC 2010
SDR	16.6	10.3	14.0	13.4
ISR	20.3	19.2	23.3	23.4
SIR	19.2	16.0	20.4	20.0
SAR	18.3	12.2	15.4	14.9

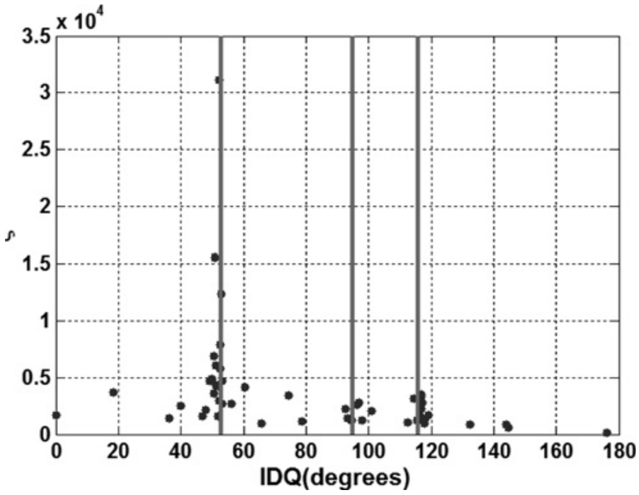

Fig. 11 *Scatter plot of ζ against IDQ for the wdrums data using KL-NTF (the true CID values are indicated by red vertical lines)*

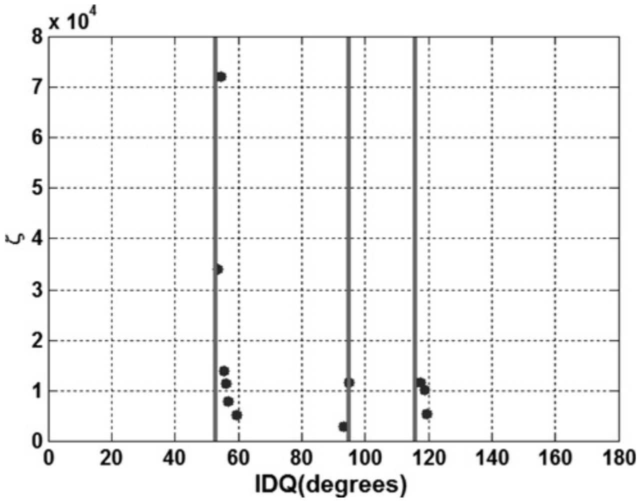

Fig. 12 *Scatter plot of ζ against IDQ for the wdrums data using Bayesian NTF (the true CID values are indicated by red vertical lines)*

listed in Table 4. They can be compared with the best results obtained in the campaigns of 2007, 2008 and 2010 [36]. Our average results are not better for all criteria, but we have to stress that the prior information used in all methods is not the same. For instance, we estimate the CID values, the number of sources and the number of components for each source, while these are taken known by the experimenters in the competing algorithms. Besides, we are using random initialisation instead of using the conventional BSS methods outcomes for initialising the parameters.

To analyse where the superiority of Bayesian NTF stems from, we provide a scatter plot of ζ versus the IDQ value of each component for KL-NTF (Fig. 11) and Bayesian NTF (Fig. 12) for the wdrums data. For Bayesian NTF, we only include the retained components. We observe that for Bayesian NTF, the IDQ values are close to one of the true CID values of each of the sources, even for the weaker components. For KL-NTF, this is not the case, especially for the weaker components. This explains the poor image quality for KL-NTF.

5 Conclusion

We have introduced novel approaches to the tasks of source counting and separation for stereo instantaneous audio mixtures. In the first stage, the channel intensity values and the number of sources are estimated. This is accomplished by evaluating an angular spectrum based on the magnitude spectrogram ratios of the two channels. It has been shown that the proposed approach can lead to accurate estimates.

The source separation task is regarded in the second stage knowing the CID values corresponding to each source. A non-parametric Bayesian model is proposed for estimating the channel mixing coefficients, spectral components and time activations. A large value is taken for the total number of components. After inference through a variational Bayesian procedure, a limited number of these components remain active. This can be specified by observing the ζ metric. Grouping the remaining active components to the individual source magnitude spectrograms is then achieved by classification of the inferred channel coefficient ratios to the classes with known centroids given by the previous stage. For time-domain derivation of the source signals, the original phase of the mixture STFT is exploited to obtain the complex STFT matrix corresponding to individual source images on the two channels. It has been shown that a Bayesian extension of NTF can provide more powerful modelling due to its ability to discard irrelevant components. This is manifested by observing the corresponding BSS evaluation metrics and also the separated signals in the time domain. In contrast to the standard NTF, our proposed method is not prone to overfitting issues. Moreover, the major advantage of the proposed non-parametric modelling is that the number of spectral components required for modelling each source is automatically obtained based on data.

6 Acknowledgment

This research was funded by the KULeuven research grant GOA/14/005 (CAMETRON).

7 References

1 Reju, V.G., Koh, S.N., Soon, I.Y.: 'Underdetermined convolutive blind source separation via time–frequency masking', *IEEE Trans. Audio Speech Lang. Process.*, 2010, **18**, (1), pp. 101–116

2 Bofill, P., Zibulevsky, M.: 'Underdetermined blind source separation using sparse representations', *Signal Process.*, 2001, **81**, (11), pp. 2353–2362

3 Vincent, E.: 'Complex nonconvex *l p* norm minimization for underdetermined source separation'. Independent Component Analysis and Signal Separation, 2007, pp. 430–437

4 Saab, R., Yilmaz, O., McKeown, M.J., *et al.*: 'Underdetermined anechoic blind source separation via *lq*-basis-pursuit', *IEEE Trans. Signal Process.*, 2007, **55**, (8), pp. 4004–4017

5 Smaragdis, P., Brown, J.: 'Non-negative matrix factorization for polyphonic music transcription'. 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2003, pp. 177–180

6 Virtanen, T.: 'Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (3), pp. 1066–1074

7 Jaiswal, R., FitzGerald, D., Barry, D., *et al.*: 'Clustering NMF basis functions using shifted NMF for monaural sound source separation'. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), May 2011, pp. 245–248

8 Lee, D.D., Seung, H.S.: 'Learning the parts of objects by non-negative matrix factorization', *Nature*, 1999, **401**, (6755), pp. 788–791

9 Virtanen, T., Cemgil, A., Godsill, S.: 'Bayesian extensions to non-negative matrix factorisation for audio signal modelling'. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008, March 2008, pp. 1825–1828

10 Cemgil, A.T.: 'Bayesian inference in non-negative matrix factorisation models'. Technical Report, CUED/F-INFENG/TR.609, University of Cambridge, July 2008

11 Ozerov, A., Fevotte, C.: 'Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation', *IEEE Trans. Audio Speech Lang. Process.*, 2010, **18**, (3), pp. 550–563

12 Févotte, C., Ozerov, A.: 'Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues'. Exploring Music Contents, 2011, pp. 102–115

13 FitzGerald, D., Cranitch, M., Coyle, E.: 'Extended nonnegative tensor factorisation models for musical sound source separation', *Comput. Intell. Neurosci.*, 2008

14 Mitsufuji, Y., Roebel, A.: 'Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge'. 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), May 2013, pp. 71–75

15 Blei, D.M., Cook, P.R., Hoffman, M.: 'Bayesian nonparametric matrix factorization for recorded music'. Proc. 27th Int. Conf. on Machine Learning (ICML-10), 2010, pp. 439–446

16 Nakano, M., Le Roux, J., Kameoka, H., *et al.*: 'Infinite-state spectrum model for music signal analysis'. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), May 2011, pp. 1972–1975

17 Nakano, M., Le Roux, J., Kameoka, H., *et al.*: 'Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model'. 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), October 2011, pp. 325–328

18 Nagira, K., Otsuka, T., Okuno, H.G.: 'Nonparametric Bayesian sparse factor analysis for frequency domain blind source separation without permutation ambiguity', *EURASIP J. Audio Speech Music Process.*, 2013, **2013**, (1), pp. 1–14

19 Arberet, S., Ozerov, A., Gribonval, R., *et al.*: 'Blind spectral-GMM estimation for underdetermined instantaneous audio source separation'. Independent Component Analysis and Signal Separation, 2009, pp. 751–758

20 Barkat, B., Sattar, F., Abed-Meraim, K.: 'Sources separation of instantaneous mixtures using a linear time–frequency representation and vectors clustering'. 2006 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proc., 2006, vol. 3, p. III

21 Sandiko, C.M., Magsino, E.R.: 'A blind source separation of instantaneous acoustic mixtures using natural gradient method'. 2012 IEEE Int. Conf. on Control System, Computing and Engineering (ICCSCE), 2012, pp. 124–129

22 Parvaix, M., Girin, L.: 'Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. audio, speech, and language processing', *IEEE Trans.*, 2011

23 Sanchis, J., Rieta, J.: 'Computational cost reduction using coincident boundary microphones for convolutive blind signal separation', *Electron. Lett.*, 2005, **41**, (6), pp. 374–376

24 Sanchis, J.M., Castells, F., Rieta, J.J.: 'Convolutive acoustic mixtures approximation to an instantaneous model using a stereo boundary microphone configuration'. Independent Component Analysis and Blind Signal Separation, 2004, pp. 816–823

25 Gunel, B., Hacihabiboglu, H., Kondoz, A.M.: 'Acoustic source separation of convolutive mixtures based on intensity vector statistics', *IEEE Trans. Audio Speech Lang. Process.*, 2008, **16**, (4), pp. 748–756

26 Mirzaei, S., Van hamme, H., Norouzi, Y.: 'Bayesian non-parametric matrix factorization for discovering words in spoken utterances'. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), October 2013, pp. 1–4

27 Yilmaz, O., Rickard, S.: 'Blind separation of speech mixtures via time–frequency masking', *IEEE Trans. Signal Process.*, 2004, **52**, (7), pp. 1830–1847

28 Arberet, S., Gribonval, R., Bimbot, F.: 'A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture'. Independent Component Analysis and Blind Signal Separation, 2006, pp. 536–543

29 Pavlidi, D., Griffin, A., Puigt, M., *et al.*: 'Real-time multiple sound source localization and counting using a circular microphone array', *IEEE Trans. Audio Speech Lang. Process.*, 2013, **21**, (10), pp. 2193–2206

30 Loesch, B., Yang, B.: 'Blind source separation based on time–frequency sparseness in the presence of spatial aliasing'. Latent Variable Analysis and Signal Separation, 2010, pp. 1–8

31 FitzGerald, D., Cranitch, M., Coyle, E.: 'Non-negative tensor factorisation for sound source separation' (Dublin Institute of Technology, 2005)

32 Bishop, C.M., *et al.*: 'Pattern recognition and machine learning' (Springer, New York, 2006)

33 Vincent, E., Araki, S., Bofill, P.: 'The 2008 signal separation evaluation campaign: a community-based approach to large-scale evaluation'. Independent Component Analysis and Signal Separation, 2009, pp. 734–741

34 Arberet, S., Gribonval, R., Bimbot, F.: 'A robust method to count and locate audio sources in a multichannel underdetermined mixture', *IEEE Trans. Signal Process.*, 2010, **58**, (1), pp. 121–133

35 Demix-inst software reference website. Available at https://www.sites.google.com/site/simonarberet/codes/

36 Vincent, E., Araki, S., Theis, F., *et al.*: 'The signal separation evaluation campaign (2007–2010): achievements and remaining challenges', *Signal Process.*, 2012, **92**, (8), pp. 1928–1936