INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Unified Bayesian Source Modelling for Determined Blind Source Separation

Chaitanya Narisetty

Data Science Research Laboratories, NEC Corporation, Japan
c-narisetty@cp.jp.nec.com

Abstract

This paper proposes a determined blind source separation (BSS) method with a Bayesian generalization for unified modelling of multiple audio sources. Our probabilistic framework allows a flexible multi-source modelling where the number of latent features required for the unified model is optimally estimated. When partitioning the latent features of the unified model to represent individual sources, the proposed approach helps to avoid over-fitting or under-fitting the correlations among sources. This adaptability of our Bayesian generalization therefore adds flexibility to conventional BSS approaches, where the number of latent features in the unified model has to be specified in advance. In the task of separating speech mixture signals, we show that our proposed method models diverse sources in a flexible manner and markedly improves the separation performance as compared to the conventional methods.

Index Terms: blind source separation, inference, non-negative matrix factorization, Bayesian non-parametrics

1. Introduction

Separation of audio sources from a set of their mixture signals is termed as source separation [1, 2]. The abundance of audio sources surrounding us often interfere with our desired source signal. This is reflected in the classic cocktail party problem, where a listener aims to follow one of the many conversations. Devices replicating this capability of source separation are used in hearing aids, speaker diarization, speech transcription, noise suppression etc [3]. Recent surge of smart devices with AI assistants also necessitate the separation methods to be invariant of each device's configuration of microphones. This lack of any a priori information on how the sources are mixed, the source characteristics, the microphone arrangement etc. are qualified as blind source separation (BSS) [4, 5]. Most fundamental BSS techniques are based on linearly decomposing the matrices containing the complex-valued spectra of mixture signals. One such technique is the frequency-domain independent component analysis (FDICA) [6, 7, 8] which aims to estimate a demixing matrix that can revert the spectra of mixture signals to the spectra of separated source signals. However FDICA suffers from a permutation problem as it necessitates the alignment of demixing parameters in each frequency bin. In case of determined BSS, where number of microphones equal the number of sources, the permutation problem can be overcome. Independent vector analysis (IVA) [9, 10] is one such technique which assumes higher-order dependencies among the frequency bins of each source and iteratively updates the demixing matrix.

Above BSS methods assume statistical independence among the source distributions which is often untrue because of the strong correlations prevalent among the magnitude spectra of audio sources. Non-negative matrix factorization (NMF) [11] is effective in extracting such correlations as a linear combination of relatively few latent features. The set of extracted features (bases) constitute a basis matrix and the set of corresponding coefficients of these bases constitute an activation matrix. The NMF-based modelling of source spectra in addition to IVA's formulation of the demixing matrix was proposed recently as independent low-rank matrix analysis (ILRMA) [12].

Multi-source modelling in ILRMA is further formulated using two approaches. In the first approach, each source is modelled separately using NMF and therefore has its own basis and activation matrix. The second approach, however, has a unified NMF multi-source model wherein all sources share from the single basis and activation matrix. Most common problem with NMF-based BSS methods is that they often over-fit or under-fit the sources, as the number of latent features to be extracted by NMF has to be specified in advance. This motivates us to reformulate these methods with more flexible frameworks. We recently proposed a Bayesian generalization for the first ILRMA approach and showed that it can adaptively model a diverse range of sources [13]. However, a unified multi-source model can capture the correlations among sources, which is not possible when each source is modelled separately.

In this work, we propose a unified Bayesian framework for modelling the multi-source spectra in a flexible manner. Our probabilistic framework enables the unified model to adaptively choose the number of latent features required for optimally modelling the diverse source characteristics. We do by this placing a sparse prior over the reliability of each latent feature in the unified Bayesian model, thereby extracting only the most relevant features and enhancing the overall separation performance. In the following sections, we unpack the above conventional methods, their limitations and proposed method in detail.

2. Conventional Method

The underlying assumption among most frequency-domain based BSS methods is that the signals captured by microphones are convolutive mixtures of source signals [7]. Let M, N denote the number of microphones and sources respectively. The complex-valued spectra of source and mixture signals as estimated by short-time Fourier transform are formulated as

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \tag{1}$$

where $\boldsymbol{x}_{ij} = (x_{ij,1}, \dots, x_{ij,M})^\mathsf{T}$ and $\boldsymbol{s}_{ij} = (s_{ij,1}, \dots, s_{ij,N})^\mathsf{T}$ denote the mixture and source spectra respectively for each index $i \in \{1, 2, \dots, I\}, j \in \{1, 2, \dots, J\}$. $(.)^\mathsf{T}$ denotes a matrix transpose and I, J denote the number of frequency bins, and time frames respectively. $\boldsymbol{A}_i$ is an $M \times N$ mixing matrix comprising of N steering vectors for the N respective sources. For determined BSS, $M = N$ and the square matrix $\boldsymbol{A}_i$ has a valid inverse matrix. Therefore [9] proposed IVA which iteratively estimates a demixing matrix $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} = (\boldsymbol{w}_{i,1}, \dots, \boldsymbol{w}_{i,M})^\mathsf{H}$ by reformulating Eq. (1) as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}, \tag{2}$$

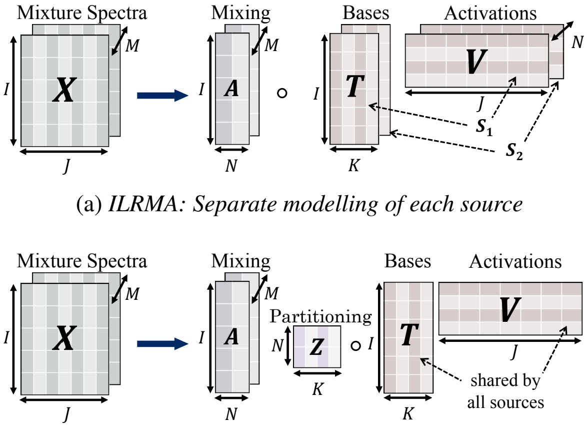
http://dx.doi.org/10.21437/Interspeech.2019-1272

where $\boldsymbol{y}_{ij} = (y_{ij,1}, \ldots, y_{ij,M})^{\mathsf{T}}$ denotes the estimated source spectra and $(.)^{\mathsf{H}}$ denotes the hermitian transpose. AuxIVA is an auxiliary function based IVA which has been proposed by [10] to avoid tuning the IVA's step-size parameter while also providing fast and stable updates.

2.1. ILRMA: Separate and Unified Source Models

IVA based methods do not capitalize on the low-rank nature of typical audio spectra. Considering the acoustic sources to be comprised of a few unique components (latent features), ILRMA extends the above formulation in AuxIVA by modelling the variance of source spectra using NMF [12]. Each source in ILRMA is independently modelled with an isotropic complex Gaussian distribution. Since $M = N$, we denote $r_{ij,m}$ as the distribution variance of each source $m \in \{1, \ldots, M\}$. The cost function Q of ILRMA is given by

$$Q = -2J \sum_i |\det \boldsymbol{W}_i| + \sum_{i,j,m} \left[\log r_{ij,m} + \frac{|y_{ij,m}|^2}{r_{ij,m}} \right]. \quad (3)$$



(a) *ILRMA: Separate modelling of each source*

(b) *U-ILRMA: Partitioning each source from a unified model*

Figure 1: *Two common variations in the multi-source modelling while separating a given mixture spectra.*

As discussed earlier, in ILRMA the multi-source modelling is done using two approaches as shown in Fig. 1. Widely used approach is the separate NMF formulation of $r_{ij,m}$ for each of the M sources [12, 14, 15]. Henceforth, this approach will be referred to as ILRMA and formulated as

$$r_{ij,m} = \sum_{k=1}^{K_m} t_{ik,m} v_{kj,m}, \quad (4)$$

where K_m is the model complexity parameter, $t_{ik,m}$ and $v_{kj,m}$ are the elements of basis and activation matrices respectively. Note that it is possible for the sources in a given mixture signal to have strong correlations among each other. Hence a second approach was given where a single basis and activation matrix is shared among all the sources. We will refer to this approach of unified NMF modelling as U-ILRMA. It is formulated as

$$r_{ij,m} = \sum_{k=1}^{K} z_{mk} t_{ik} v_{kj}, \quad (5)$$

where K is the model complexity parameter, t_{ik} and v_{kj} are elements of the common basis and activation matrix respectively and z_{mk} is called a 'partitioning function' which quantifies the non-negative contribution of k^{th} basis towards the variance of m^{th} source. Note that U-ILRMA constrains the combined contribution of each basis as

$$\sum_{m=1}^{M} z_{mk} = 1, \ \forall\, k \in \{1, \ldots, K\}. \quad (6)$$

Although ILRMA does not capitalize on the existence of correlations among sources, it is shown to be more reliable than the latter approach of U-ILRMA [12]. This is because the additional parameter z_{mk} expands the overall optimization space, thus creating more number of local minima. However in separating music mixtures, U-ILRMA with a large number of bases ($K \sim 30$) is shown to have better separation performance.

2.2. Limitations

One of the main limitations of both ILRMA and U-ILRMA is that the number of bases must be specified in advance. Also, ILRMA specifies equal number of bases for all the sources i.e. $\{K_1 = \cdots = K_M = K\}$. If mixtures of sources with different characteristics like alarm and speech are given, then ILRMA cannot assign different number of basis to different sources. In U-ILRMA, the partitioning function z can assign different overall contributions of the bases towards each source and theoretically overcome ILRMA's limitation. However the combined complexity of all sources K in U-ILRMA must still be specified. As discussed above, U-ILRMA also expands the overall optimization space by adding an additional matrix to the factorization. Therefore it is not guaranteed that bases are optimally partitioned according to each source's characteristics. These limitations motivate the formulation of a Bayesian framework over U-ILRMA and lifting the constraint in Eq. (6) to create a more stable partitioning function.

3. Proposed Method

We overcome the limitations of U-ILRMA by proposing a probabilistic framework for the unified modelling of the source variances. In such techniques, it is common to introduce hidden variables to capture the structure of given observed data, and then inference them to estimate the posterior distribution. We recently proposed a Bayesian generalization of ILRMA [13]. In this work, we extend this generalization to the unified modelling approach of U-ILRMA. We model the multi-source spectra using a unified Bayesian NMF with a large number of basis vectors and place a sparse prior over the partitioning function z to extract the most relevant basis vectors.

3.1. Model Formulation

Contrast to the constrained modelling in Eq. (6), the proposed probabilistic framework assumes the prior distributions for each of t_{ik}, v_{kj} and z_{mk} to be drawn from a random process as

$$\begin{aligned} p(t_{ik}) &\sim \text{Gamma}(a_0, a_0), \\ p(v_{kj}) &\sim \text{Gamma}(b_0, b_0), \\ p(z_{mk}) &\sim \text{Gamma}(c_0, c_m), \end{aligned} \quad (7)$$

where a_0, b_0, c_0 are positive constants, Gamma$(.,.)$ is a gamma distribution defined over a shape parameter and a rate (inverse-scale) parameter. For sparsity, we set $c_0 \ll 1$ so that the partitioning function adaptively models diverse sources [16]. Note that the expectation of t and v are set to 1, while their variance is $1/a_0$ and $1/b_0$ respectively. This implies that a slightly large value of a_0 would sample the basis vectors from a closely packed space, which can help in modelling speech spectra. As each source's expected variance should correspond to the expectation of its power, the choice of prior parameters require that $\mathbb{E}_p[|y_{ij,m}|^2] = \mathbb{E}_p[r_{ij,m}]$,

$$\Rightarrow \mathbb{E}_p[|y_{ij,m}|^2] = \sum_k \mathbb{E}_p[z_{mk}]\mathbb{E}_p[t_{ik}v_{kj}] = \sum_k (c_0/c_m),$$

$$\Rightarrow \qquad c_m = c_0 K \left[\sum_i \sum_j |y_{ij,m}|^2/(IJ)\right]^{-1}. \tag{8}$$

Based on the proposed formulation of t, v and z, we maximize the cost function Q in Eq. (3).

3.2. Update Rules for Demixing Matrix

Under the proposed framework, the partial derivatives of cost function Q over the demixing matrix $\boldsymbol{W}_i$ remain unchanged, and so do their update equations. They are derived in [10] as

$$V_{i,m} = \frac{1}{J}\sum_j \frac{1}{r_{ij,m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^h, \tag{9}$$

$$\boldsymbol{w}_{i,m} \leftarrow (\boldsymbol{W}_i V_{i,m})^{-1}\boldsymbol{e}_m, \tag{10}$$

$$\boldsymbol{w}_{i,m} \leftarrow \boldsymbol{w}_{i,m}(\boldsymbol{w}_{i,m}^h V_{i,m}\boldsymbol{w}_{i,m})^{-1/2}, \tag{11}$$

where $\boldsymbol{e}_m$ is a unit vector whose m^{th} element equals one. The separated source spectra can then be extracted as

$$y_{ij,m} \leftarrow \boldsymbol{w}_{i,m}^h \boldsymbol{x}_{ij}. \tag{12}$$

3.3. Variational Inference

Similar to the Bayesian generalization of ILRMA, we will adopt a fully factorized mean-field variational inference technique and approximate the hidden variables t, v and z from a family of conditional distributions over variational parameters [17]. We choose the Generalized inverse Gaussian (GIG) distributions for our variational family, which are expressed as

$$\mathrm{GIG}(\theta|\gamma,\rho,\tau) = \frac{\exp\{(\gamma-1)\log\theta - \rho\theta - \tau/\theta\}}{2(\tau/\rho)^{\gamma/2}\mathcal{K}_\gamma(2\sqrt{\rho\tau})}, \tag{13}$$

where $\mathcal{K}(.)$ is a modified Bessel function of the second kind and γ, ρ, τ are the variational hyper-parameters. We define the conditional distributions of our unified Bayesian model as

$$\begin{aligned} q(t_{ik}|\Theta_{\setminus t_{ik}}) &\sim \mathrm{GIG}(a_0, \rho_{ik}^{(t)}, \tau_{ik}^{(t)}), \\ q(v_{kj}|\Theta_{\setminus v_{kj}}) &\sim \mathrm{GIG}(b_0, \rho_{kj}^{(v)}, \tau_{kj}^{(v)}), \\ q(z_{mk}|\Theta_{\setminus z_{mk}}) &\sim \mathrm{GIG}(c_0, \rho_{mk}^{(z)}, \tau_{mk}^{(z)}). \end{aligned} \tag{14}$$

The main reason for choosing GIG is that its sufficient statistics are $\theta, (1/\theta)$ and $\log(\theta)$, which are a superset of the sufficient statistics of our Gamma prior distribution [18]. We now derive update equations from the cost function in Eq. (3) using first-order Taylor expansion and Jensen's inequality [19] using their respective auxiliary positive constants $\alpha_{ij,m}$ and $\beta_{ijk,m}$ as

$$Q + 2J\sum_i |\det \boldsymbol{W}_i| = \sum_{i,j,m}\left[\log r_{ij,m} + \frac{|y_{ij,m}|^2}{r_{ij,m}}\right],$$

$$\begin{aligned} &\leq \sum_{i,j,m}\mathbb{E}_q\left[\log r_{ij,m} + \frac{|y_{ij,m}|^2}{r_{ij,m}}\right] + \mathbb{E}_q\left[\log\frac{q(t|\Theta_{\setminus t})}{p(t|a_0)}\right] \\ &\quad + \mathbb{E}_q\left[\log\frac{q(v|\Theta_{\setminus v})}{p(v|b_0)}\right] + \mathbb{E}_q\left[\log\frac{q(z|\Theta_{\setminus z})}{p(z|c_0,c_m)}\right], \\ &\leq \sum_{i,j}\left[\sum_{k,m}|y_{ij,m}|^2\beta_{ijk,m}^2\mathbb{E}_q\left[z_{mk}^{-1}t_{ik}^{-1}v_{kj}^{-1}\right]\right. \\ &\quad \left. - 1 + \log\alpha_{ij,m} + \frac{1}{\alpha_{ij,m}}\sum_k \mathbb{E}_q[z_{mk}t_{ik}v_{kj}]\right] \\ &\quad + \mathbb{E}_q[-\rho_{ik}^{(t)}t_{ik} - \tau_{ik}^{(t)}/t_{ik} + a_0 t_{ik}] \\ &\quad + \mathbb{E}_q[-\rho_{kj}^{(v)}v_{kj} - \tau_{kj}^{(v)}/v_{kj} + b_0 v_{kj}] \\ &\quad + \mathbb{E}_q[-\rho_{mk}^{(z)}z_{mk} - \tau_{mk}^{(z)}/z_{mk} + c_m z_{mk}] + C, \end{aligned} \tag{15}$$

where C is a leftover constant. As mentioned earlier, the pairs $(t, t^{-1}), (v, v^{-1})$ and (z, z^{-1}) are subsets of the GIG sufficient statistics. This allows us to directly derive the analytic coordinate ascent updates for our hyper-parameters by setting their coefficients to zero in the above inequality. Constants $\alpha_{ij,m}$ and $\beta_{ijk,m}$ re-tighten the above inequality (15) when:

$$\alpha_{ij,m} = \sum_k \mathbb{E}_q[z_{mk}]\mathbb{E}_q[t_{ik}]\mathbb{E}_q[v_{kj}], \tag{16}$$

$$\beta_{ijk,m} = \frac{\mathbb{E}_q\left[z_{mk}^{-1}\right]\mathbb{E}_q\left[t_{ik}^{-1}\right]\mathbb{E}_q\left[v_{kj}^{-1}\right]}{\sum_k \mathbb{E}_q\left[z_{mk}^{-1}\right]\mathbb{E}_q\left[t_{ik}^{-1}\right]\mathbb{E}_q\left[v_{kj}^{-1}\right]}. \tag{17}$$

Expectation of source parameters in Eqs. (16) and (17) can be obtained from those of a GIG distribution. The update equations for our hyper-parameters are derived as

$$\rho_{ik}^{(t)} = a_0 + \sum_m \mathbb{E}_q[z_{mk}]\sum_j \mathbb{E}_q[v_{kj}]\alpha_{ij,m}^{-1}, \tag{18}$$

$$\tau_{ik}^{(t)} = \sum_m \mathbb{E}_q\left[z_{mk}^{-1}\right]\sum_j |y_{ij,m}|^2\beta_{ijk,m}^2\mathbb{E}_q\left[v_{kj}^{-1}\right], \tag{19}$$

$$\rho_{kj}^{(v)} = b_0 + \sum_m \mathbb{E}_q[z_{mk}]\sum_i \mathbb{E}_q[t_{ik}]\alpha_{ij,m}^{-1}, \tag{20}$$

$$\tau_{kj}^{(v)} = \sum_m \mathbb{E}_q\left[z_{mk}^{-1}\right]\sum_i |y_{ij,m}|^2\beta_{ijk,m}^2\mathbb{E}_q\left[t_{ik}^{-1}\right], \tag{21}$$

$$\rho_{mk}^{(z)} = c_m + \sum_i\sum_j \mathbb{E}_q[t_{ik}]\mathbb{E}_q[v_{kj}]\alpha_{ij,m}^{-1}, \tag{22}$$

$$\tau_{mk}^{(z)} = \sum_i\sum_j |y_{ij,m}|^2\beta_{ijk,m}^2\mathbb{E}_q\left[t_{ik}^{-1}\right]\mathbb{E}_q\left[v_{kj}^{-1}\right]. \tag{23}$$

In addition to Eqs. (18)-(23), the demixing matrix and the separated source spectra are updated using Eqs. (9)-(12) in each iteration. As there exists a scale ambiguity between the demixing matrix and the source variances, we normalize $\boldsymbol{W}$ and y similar to the normalization suggested in [12]. The difference being that t, v, z are not normalized but instead inferred from y.

4. Simulations and Results

4.1. Experimental Conditions

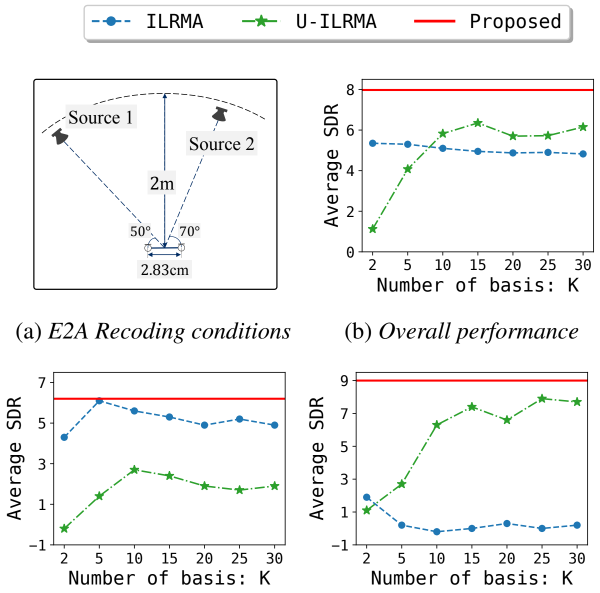
We evaluate our proposed method on the mixtures of speech sources obtained from the SiSEC2011 dataset [20] as given in Table. 1. Using these sources, we create synthetic two-channel reverberant mixtures using the recoding conditions shown in Fig. 2a. The room impulse responses E2A (reverberation time: $T_{60} = 300$ ms) for above recording conditions were obtained from the RWCP Sound Scene Database [21].

Table 1: *Speech sources from SiSEC database*

ID	Class name	Track name	Language
1	dev1_female4	src_1/src_2	English
2	dev1_female4	src_3/src_4	Japanese
3	dev1_male4	src_1/src_2	English
4	dev1_male4	src_3/src_4	Japanese

Mixture spectra are estimated from the time domain signals using a Hamming window of length 512 ms shifted every 128 ms. Each demixing matrix $\boldsymbol{W}_i$ is initialized with an identity matrix. We model each source's variance with $K = 30$ basis and set $a_0 = 0.5$, $b_0 = 0.1$, $c_0 = 1/K$. Hyper-parameters ρ, τ are initialized randomly from gamma distributions with shape and rate parameters set to 100. Number of iterations is 200. A back-projection technique [22] is used to convert the separated spectra y to time domain. Three metrics: signal to distortion ratio (SDR), signal to interference ratio (SIR) and signal to artifacts ratio (SAR) [23] are used to evaluate the quality of separated sources. Each separation is repeated for 10 different random initializations, and the average of above performance metrics are estimated accordingly.

4.2. Results



(a) *E2A Recoding conditions* (b) *Overall performance*

(c) *Mixture ID: 2* (d) *Mixture ID: 3*

Figure 2: *(a) shows the source-microphone setup for the E2A room impulse responses. (b), (c) and (d) compare the average SDR of proposed method with ILRMA and U-ILRMA over all mixture IDs, for ID: 2 and ID: 3 respectively.*

For the given setup, Fig. 2c, 2d show the average SDR for mixture ID: 2 and ID: 3 respectively over different values of K. It can be seen that the performance of ILRMA is higher than that of U-ILRMA for ID: 2 and vice-versa for ID: 3. This could be due to the difference in language and/or gender between IDs 2 and 3. In comparison, our proposed method is able to achieve SDR that matches the better to both of these methods. We use a horizontal line to depict our proposed method's SDR although it is initialized with $K = 30$ bases. This is because over time, it discards the contribution from irrelevant basis and adaptively models the sources with an optimal number of bases.

The overall separation performance averaged over the 4 speech mixtures is compared for all the methods and shown in Fig. 2b. It is evident that the proposed method is able to outperform both ILRMA and U-ILRMA. Further, ILRMA seems to perform better with smaller K as opposed to U-ILRMA which prefers a higher K. We further compare our proposed method with the Bayesian generalization of ILRMA (Bay-ILRMA) [13] with $K = 30$ bases. The comparison of overall SIR, SAR and SDR are summarized in Table. 2. We see that, compared to the best configuration of ILRMA and U-ILRMA, the SDR improvement of the proposed method is at least 1.7dB. For both the Bayesian approaches, we only set $K = 30$ and let them flexibly model the sources in each mixture. As expected, Bay-ILRMA has a higher SDR than the best performing ILRMA. We also see an improvement in all three metrics for the unified multi-source modelling approaches.

Table 2: *Comparison of Separation Performance*

Methods	Best K	**SIR**	**SAR**	**SDR**
ILRMA	2	9.6 dB	10.3 dB	5.4 dB
U-ILRMA	15	12.4 dB	10.4 dB	6.3 dB
Bay-ILRMA	– auto –	11.7 dB	10.2 dB	6.1 dB
Proposed	– auto –	**14.2** dB	**10.9** dB	**8.0** dB

4.3. Discussions

Variational inference methods are often computationally intensive as compared to their deterministic counterparts. Table. 3 shows a comparison of the computational time for 100 and 200 iterations. We observe that the time taken by U-ILRMA is similar to that of the proposed method. This is because, inference is exacting in roughly the first 30 iterations. However, as the optimization continues, number of basis required for source modelling decreases and the iterations become much faster. Note that although ILRMA is faster than the proposed method, the time taken to iterate over different values of K in search of the best performing ILRMA is still considerably higher.

Table 3: *Relative computational time for $K = 30$ bases*

Iterations	**ILRMA**	**U-ILRMA**	**Bay-ILRMA**	**Proposed**
100	1.00	1.45	2.31	1.52
200	1.96	2.85	4.26	2.93

For our proposed method, it is possible to initialize the bases and/or the demixing matrix from conventional approaches for further improvements. Alternatively to the Gamma process prior assumed by our approach, it is also possible to assume a Beta process sparse NMF [24] based unified modelling. Recently, there have also been other generalizations of the source distributions instead of the isotropic complex Gaussian distribution like the Student's-t distribution (t-ILRMA) [25], generalized Gaussian distribution (GGD-ILRMA) [26]. However, they introduce additional latent variables into the cost function Q and hence need parameter tuning. Bayesian extensions of such frameworks will be considered as part of our future work.

5. Conclusions

We propose a determined blind source separation method with a Bayesian generalization for the unified multi-source modelling based on a Gamma process NMF. Our formulation is able to overcome the limitation of conventional methods, whose separation performance changes with the number of latent features extracted by the traditional NMF. We experimentally show on the SiSEC2011 dataset that the proposed approach is flexible in modelling sources of different complexities, allowing it to optimally separate them. We further show that our approach outperforms the state-of-the-art ILRMA based approaches.

6. References

[1] M. Davies, "Audio source separation," in *Institute of mathematics and its applications conference series*, vol. 71, 2002, pp. 57–68.

[2] S. Makino, *Audio Source Separation*, ser. Signals and Communication Technology. Springer International Publishing, 2018.

[3] J. Foote, "An overview of audio information retrieval," *Multimedia systems*, vol. 7, no. 1, pp. 2–10, 1999.

[4] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[5] X. Cao and R. Liu, "General approach to blind source separation," *IEEE Transactions on signal Processing*, vol. 44, no. 3, pp. 562–571, 1996.

[6] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE transactions on speech and audio processing*, vol. 12, no. 5, pp. 530–538, 2004.

[7] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[8] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ica and beamforming," *IEEE Transactions on Audio, speech, and language processing*, vol. 14, no. 2, pp. 666–678, 2006.

[9] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ica to multivariate components," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 165–172.

[10] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 9, pp. 1622–1637, 2016.

[13] C. Narisetty, T. Komatsu, and R. Kondo, "Bayesian non-parametric multi-source modelling based determined blind source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[14] Y. Mitsui, D. Kitamura, S. Takamichi, N. Ono, and H. Saruwatari, "Blind source separation based on independent low-rank matrix analysis with sparse regularization for time-series activity," in *2017 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 21–25.

[15] Y. Mitsui, D. Kitamura, N. Takamune, H. Saruwatari, Y. Takahashi, and K. Kondo, "Independent low-rank matrix analysis based on parametric majorization-equalization algorithm," in *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2017, pp. 1–5.

[16] V. Y. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

[17] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[18] D. M. Blei, P. R. Cook, and M. Hoffman, "Bayesian nonparametric matrix factorization for recorded music," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 439–446.

[19] J. D. Lafferty and D. M. Blei, "Correlated topic models," in *Advances in neural information processing systems*, 2006, pp. 147–154.

[20] S. Araki, F. Nesta, E. Vincent, Z. Koldovskỳ, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (sisec2011):-audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 414–422.

[21] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and Hands-Free speech recognition." in *LREC*, 2000, [Online; accessed 29-Oct-2018] Available: http://www.openslr.org/13/.

[22] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] D. Liang, M. D. Hoffman, and D. P. Ellis, "Beta process sparse nonnegative matrix factorization for music." in *ISMIR*, 2013, pp. 375–380.

[25] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, and N. Ono, "Independent low-rank matrix analysis based on complex student's t-distribution for blind audio source separation," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.

[26] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, p. 28, 2018.