Independent Vector Analysis with Frequency Range Division and Prior Switching

Rintaro Ikeshita*, Yohei Kawaguchi*, Masahito Togami†, Yusuke Fujita*, Kenji Nagamatsu*,
* Hitachi, Ltd., Research & Development Group, Japan,
† Hitachi America, Ltd., USA

***Abstract*—A novel source model is developed to improve the separation performance of independent vector analysis (IVA) for speech mixtures. The source model of IVA generally assumes the same amount of statistical dependency on each pair of frequency bins, which is not effective for speech signals with strong correlations among neighboring frequency bins. In the proposed model, the set of all frequency bins is divided into frequency bands, and the statistical dependency is assumed only within each band to better represent speech signals. In addition, each source prior is switched depending on the source states, active or inactive, since intermittent silent periods have totally different priors from those of speech periods. The optimization of the model is based on an EM algorithm, in which the IVA filters, states of sources, and permutation alignments between each pair of bands are jointly optimized. The experimental results show the effectiveness of the proposed model.**

***Index Terms*—Blind source separation, independent vector analysis (IVA), independent component analysis (ICA)**

I. Introduction

Blind source separation (BSS) is a key technique for recovering unknown source signals from only their mixtures, and has been applied to, for instance, preprocessing for speech recognition tasks for multiple speakers. Among frequency-domain approaches for BSS, independent vector analysis (IVA [1]–[3]), multivariate extension of frequency-domain independent component analysis (FD-ICA [4], [5]), is appealing since it is not affected by the permutation ambiguity owing to its modeling of each source. IVA generally assumes a spherical multivariate distribution as a source prior, i.e., the same amount of statistical dependency on each pair of frequency bins. As reported in a variety of researches [6]–[8], however, this assumption is not effective for speech signals with stronger correlations among neighboring frequency bins than between distant bins.

As a generalization of IVA to improve dependency models of sources, a family of non-spherical distributions is introduced as a source prior for IVA [6], [7], [9]. These distributions have strong correlations within neighboring frequency bins, but weak ones between distant bins. In the presence of speech, the priors in [6], [7], [9] fit the distributions of speech signals more accurately than those that treat all frequencies equally. However, during silent periods of speech signals, the dependency models are not appropriate because they have zero powers in all frequency bins, i.e., the same amount of dependency between each pair of frequency bins. Since speech signals have many intermittent silent intervals, the model proposed in [6], [7], [9] is not sufficient to achieve high separation performance.

On the other hand, motivated by research for ICA [10], [11], signal states, active or inactive, were incorporated into IVA models in [12], [13] since intermittent silent periods should be modeled in principle as a delta function that is totally different from those of speech periods. In the source models of [12], [13], however, it is assumed that, given signal states, the distributions of sources are independent between each pair of frequency bins. In other words, the dependencies of active speech signals in the frequency axis direction are disregarded, which causes separation performance degradation.

Recently, as another extension of IVA, variances of sources are modeled by using nonnegative matrix factorization (NMF) in [8], [14]. This approach is called an independent low-rank matrix analysis (ILRMA, [15]). By modeling sources with NMF, ILRMA can express the co-occurrence of frequency components such as the harmonic structure commonly seen in speech and music signals. In particular, since the co-occurrence in music signals is remarkable, ILRMA demonstrates very high separation performance for them. However, for speech signals that cannot be solely represented by the co-occurrence of frequency components, modeling them with NMF is not appropriate.

In this paper, we propose a novel source model where the set of all frequency bins is divided into frequency bands and the statistical dependency is assumed only within each band while at the same time intermittent silent periods are captured by switching signal distributions. In the proposed model, statistical dependencies that are only within neighboring frequency bins can be represented for signals in the presence of speech, while all frequency components can be modeled to have zero powers for signals during silent periods. This makes it possible to demonstrate higher separation performance than in the conventional method [6], [7], [9], [12], [13]. In addition, since the dependency of each band is treated uniformly and the signal model by NMF is not assumed, the proposed model is considered to better represent distributions of speech signals than ILRMA.

Besides developing the new model, we derive an optimization algorithm for it, where the states of sources, IVA filters, and permutation alignments between each pair of frequency bands are jointly optimized by combining an EM algorithm with a fast and stable auxiliary function method for IVA proposed in [16]–[18]. Owing to the simultaneous optimization,

ISBN 978-0-9928626-7-1

the proposed algorithm can robustly fit the prior of the model to the distribution of each source. The experimental results suggest that the proposed model with the robust optimization algorithm is more effective than the conventional methods when separating speech mixtures.

II. Proposed method

A. Formulation

Assume that N sources are observed by N microphones. Let us denote the source signals and the microphone observations in each time-frequency bin (f,t) as

$$\boldsymbol{s}_{f,t} = [s_{1,f,t},\dots,s_{N,f,t}]^\top \tag{1}$$

$$\boldsymbol{x}_{f,t} = [x_{1,f,t},\dots,x_{N,f,t}]^\top, \tag{2}$$

where $\cdot^\top$ means the matrix transpose, and $f \in [N_F] := \{1,\dots,N_F\}$ and $t \in [N_T] := \{1,\dots,N_T\}$ are the indexes of frequency bins and time frames, respectively. As in the case of the conventional FD-ICA and IVA, we assume the linear mixing model

$$\boldsymbol{x}_{f,t} = A_f \boldsymbol{s}_{f,t}, \tag{3}$$

where A_f is the $N \times N$ mixing matrix. The sources are recovered by

$$\boldsymbol{s}_{f,t} = W_f^* \boldsymbol{x}_{f,t}, \quad W_f = [\boldsymbol{w}_{1,f},\dots,\boldsymbol{w}_{N,f}], \tag{4}$$

where $\cdot^*$ represents the conjugate transpose and W_f is the demixing matrix whose columns consist of the separation filters $\boldsymbol{w}_{n,f}$ for each source $n \in [N] := \{1,\dots,N\}$.

B. Generative model and objective function

Given a frequency range $[N_F]$, $[N_F]$ is divided into frequency bands by introducing

$$\mathcal{F} \subseteq 2^{[N_F]} \quad \text{s.t.} \quad \sqcup_{F\in\mathcal{F}} F = [N_F], \tag{5}$$

where $\sqcup$ denotes the disjoint union of sets. Then, let us assume the decomposition of

$$p(\{\boldsymbol{s}_{n,F,t}\}_{n,F,t}) = \prod_{n\in[N]} \prod_{F\in\mathcal{F}} \prod_{t\in[N_T]} p(\boldsymbol{s}_{n,F,t}), \tag{6}$$

where we define

$$\boldsymbol{s}_{n,F,t} := [s_{n,f_1,t},\dots,s_{n,f_k,t}]^\top \tag{7}$$

for $F = \{f_1,\dots,f_k\}$. Decomposition (6) allows statistical dependency within each frequency band $F \in \mathcal{F}$ but assumes that each source is independent between frequency bands. We call the above $\mathcal{F}$ a *frequency range division*.

To represent the source states (active/inactive), we introduce hidden variables $\{z_{n,F,t}\}_{n,F,t}$ defined as

$$z_{n,F,t} = \begin{cases} 1 & \text{if active,} \\ 0 & \text{if inactive.} \end{cases} \tag{8}$$

Then, for each $(F,t) \in \mathcal{F} \times [N_T]$ the probability density function (p.d.f.) of source n is expressed as

$$p(\boldsymbol{s}_{n,F,t}) = \sum_{c\in\{0,1\}} \pi_{n,t,c} \cdot p(\boldsymbol{s}_{n,F,t} \mid z_{n,F,t} = c), \tag{9}$$

where $\pi_{n,t,c} = p(z_{n,F,t} = c)$ is the mixing coefficient. In principle, the conditional p.d.f. under the state $z_{n,F,t} = 0$ shall be defined as the delta function since the power of the silent signals is equal to zero. We use a Dirichlet prior for $\{\pi_{n,t,c}\}_c$ as

$$p(\{\pi_{n,t,c}\}_c) \propto \prod_{c\in\{0,1\}} (\pi_{n,t,c})^{\phi_c - 1}, \tag{10}$$

where ϕ_c is the hyperparameter of the Dirichlet distribution. Note that $\pi_{n,t,c}$ is assumed to be independent of frequency $F \in \mathcal{F}$, which makes the proposed method permutation-free even under decomposition (6) (see Subsection II-D).

The set of parameters in the model is

$$\Theta := \{W_f,\, \pi_{n,t,c}\}_{n,f,t,c}, \tag{11}$$

which in this paper will be optimized based on a MAP criterion that is equivalent to the following minimization problem (we use relation (4)):

$$\begin{aligned} \min_\Theta \; & -\frac{1}{N_T} \sum_{n,F,t} \log p(\boldsymbol{s}_{n,F,t}) - 2\sum_f \log|\det W_f| \\ & - \sum_{n,t} \log p(\{\pi_{n,t,c}\}_c). \end{aligned} \tag{12}$$

C. EM algorithm for parameter estimation

In this subsection, an EM algorithm to solve (12) is developed. After the convergence of the EM, the separated signals are obtained by (4), and the signal scale can be restored based on the minimal distortion principle [19], [20], in which the microphone observation of source n is calculated by

$$s_{n,f,t} A_f e_n = (\boldsymbol{w}_{n,f}^* \boldsymbol{x}_{f,t})(W_f^*)^{-1} e_n \in \mathbb{C}^N, \tag{13}$$

where e_n is a unit vector with the n-th element equal to one and the other elements equal to zero.

1) E-step of EM algorithm: In the E-step, for each source $n \in [N]$ and each time-frequency $(F,t) \in \mathcal{F} \times [N_T]$, the a posteriori probability of $z_{n,F,t} = c \in \{0,1\}$ given an estimated source signal $\boldsymbol{s}'_{n,F,t}$ is expressed as

$$\begin{aligned} q_{n,F,t,c} &\equiv p(z_{n,F,t} = c \mid \boldsymbol{s}'_{n,F,t}\,;\, \Theta') \\ &= \frac{\pi'_{n,t,c} \cdot p(\boldsymbol{s}'_{n,F,t} \mid z_{n,F,t} = c)}{\sum_{c\in\{0,1\}} \pi'_{n,t,c} \cdot p(\boldsymbol{s}'_{n,F,t} \mid z_{n,F,t} = c)}, \end{aligned} \tag{14}$$

where $\pi'_{n,t,c}, \boldsymbol{w}'_{n,f} \in \Theta'$ denotes the model parameters estimated in the last iteration of the EM algorithm and

$$\boldsymbol{s}'_{n,f,t} := (\boldsymbol{w}'_{n,f})^* \boldsymbol{x}_{f,t}. \tag{15}$$

This $q_{n,F,t,c}$ has a similar meaning to the conventional time-frequency mask for source n, and in what follows it is called a time-frequency mask as well.

ISBN 978-0-9928626-7-1 © EURASIP 2017

By using the time-frequency mask $q_{n,F,t,c}$, the first term of (12) is transformed as follows:

$$\begin{aligned}
&-\log p(\boldsymbol{s}_{n,F,t}\,;\,\Theta)\\
&=-\log\sum_{z_{n,F,t}} p(\boldsymbol{s}_{n,F,t},z_{n,F,t}\,;\,\Theta)\\
&\le-\sum_{c\in\{0,1\}} q_{n,F,t,c}\log\pi_{n,t,c}\\
&\quad-\sum_{c\in\{0,1\}} q_{n,F,t,c}\log p(\boldsymbol{s}_{n,F,t}\mid z_{n,F,t}=c\,;\,\Theta)\\
&\quad+\sum_{c\in\{0,1\}} q_{n,F,t,c}\log q_{n,F,t,c}, \qquad (16)
\end{aligned}$$

where the inequality holds if and only if

$$q_{n,F,t,c}=p(z_{n,F,t}=c\mid \boldsymbol{s}_{n,F,t}\,;\,\Theta). \qquad (17)$$

Inequality (16) gives an upper bound on the cost function of (12) that we will try to minimize:

$$\begin{aligned}
\min_{\Theta}\ &-\frac{1}{N_T}\sum_{n,F,t,c} q_{n,F,t,c}\log p(\boldsymbol{s}_{n,F,t}\mid z_{n,F,t}=c\,;\,\Theta)\\
&-2\sum_f\log|\det W_f|\\
&-\sum_{n,t,c}\Big(\sum_{F\in\mathcal{F}} q_{n,F,t,c}+\phi_c-1\Big)\log\pi_{n,t,c}. \qquad (18)
\end{aligned}$$

2) M-step of EM algorithm: In the M-step, we update the model parameters according to the problem (18). The update rules for $\{\pi_{n,t,c}\}_{n,t,c}$ are easily obtained as

$$\pi_{n,t,c}\leftarrow\frac{\sum_{F\in\mathcal{F}} q_{n,F,t,c}+\phi_c-1}{|\mathcal{F}|+\sum_{c\in\{0,1\}}(\phi_c-1)}, \qquad (19)$$

where $|\mathcal{F}|$ denotes the cardinality of set $\mathcal{F}$.

As for the separation filters $\{\boldsymbol{w}_{n,f}\}_{n,f}$, we derive fast and stable update rules based on an auxiliary function method for IVA proposed by [16], [17]. To do that, we need to suppose in the following that the conditional p.d.f. of signal $\boldsymbol{s}_{n,F,t}$ given its state $z_{n,F,t}\in\{0,1\}$ is characterized by

$$r_{n,F,t}:=\|\boldsymbol{s}_{n,F,t}\|_2=\sqrt{\sum_{f\in F}|s_{n,f,t}|^2}. \qquad (20)$$

To simplify the notation, we define

$$g_c^{(n,F)}(r_{n,F,t}):=-\log p(\boldsymbol{s}_{n,F,t}\mid z_{n,F,t}=c), \qquad (21)$$

which we call a contrast function for the state $c\in\{0,1\}$. It will sometimes be abbreviated as $g_c(r_{n,F,t})$. Suppose also that g_c satisfies the following two conditions as in [16]–[18]:

(C1) $g_c\colon\mathbb{R}_{>0}\to\mathbb{R}$ is continuously differentiable;

(C2) $\frac{g_c'(r)}{r}$ is positive and monotonically decreasing, where g_c' is the first derivative of g_c.

Functions that satisfy (C1) and (C2) include (30) below.

Under the above assumptions, we can derive the following auxiliary function $J(\Theta)$ of the cost (18) in the same way as described in [16]–[18]:

$$J(\Theta)=\sum_{n,f}\boldsymbol{w}_{n,f}^* R_{n,f}\boldsymbol{w}_{n,f}-2\sum_f\log|\det W_f|+C, \qquad (22)$$

where C is independent of $\{\boldsymbol{w}_{n,f}\}_{n,f}$, and

$$R_{n,f}:=\frac{1}{N_T}\sum_t\left[\phi(r'_{n,F,t})\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^*\right],\ f\in F \qquad (23)$$

$$\phi(r'_{n,F,t}):=\frac{\sum_{c\in\{0,1\}} q_{n,F,t,c}\cdot g_c'(r'_{n,F,t})}{2r'_{n,F,t}} \qquad (24)$$

$$r'_{n,F,t}:=\sqrt{\sum_{f\in F}|(\boldsymbol{w}'_{n,f})^*\boldsymbol{x}_{f,t}|^2}. \qquad (25)$$

The minimization of (22) will be iteratively performed by a block coordinate descent method, i.e., for each $n\in[N]$, update $\boldsymbol{w}_{n,f}$ in order using

$$\boldsymbol{w}_{n,f}\leftarrow\left(W_f^*R_{n,f}\right)^{-1}e_n \qquad (26)$$

$$\boldsymbol{w}_{n,f}\leftarrow\frac{\boldsymbol{w}_{n,f}}{\sqrt{\boldsymbol{w}_{n,f}^*R_{n,f}\boldsymbol{w}_{n,f}}} \qquad (27)$$

D. Permutation alignment

While the independence between divided frequency bands is assumed in (6), the proposed algorithm is not affected by the permutation ambiguity after separating mixtures owing to the frequency independent $\pi_{n,t,c}$. This permutation-free technique was first proposed by Ito *et al.* [21] to make BSS methods based on time-frequency clustering permutation-free. This technique can also be adopted in the proposed algorithm, i.e., at each step in the EM algorithm, the permutation alignments $\sigma_F\colon[N]\to[N]$ $(F\in\mathcal{F})$ are obtained to minimize (12) by permuting the separation filters and the contrast functions as follows:

$$\boldsymbol{w}_{n,f}\leftarrow\boldsymbol{w}_{\sigma_F(n),f}\quad\text{for } f\in F, \qquad (28)$$

$$g_c^{(n,F)}\leftarrow g_c^{(\sigma_F(n),F)}. \qquad (29)$$

E. Contrast function

When we obtain the time-frequency mask (14), we need to calculate the constant terms of contrast function g_c for the state c. In this paper, we use as the p.d.f. for $\{g_c(r)\}_c$ a complex-valued multivariate exponential power (MEP) distribution (see e.g., [22] for real-valued MEP) given by

$$g_c(r)=\left(\frac{r^2}{\alpha_c}\right)^{\beta_c}-\log\frac{\Gamma(1+d)}{(\alpha_c\pi)^d\cdot\Gamma(1+\frac{d}{\beta_c})} \qquad (30)$$

where d denotes the dimension of the complex-valued random variables $\boldsymbol{s}_{n,F,t}\in\mathbb{C}^d$ (recall the relation of (20)), and $\Gamma(\cdot)$ denotes the gamma function. Note that this contrast function is the same as that in [17], under which a fast and stable IVA based on an auxiliary function method can easily be established.

ISBN 978-0-9928626-7-1

F. Summary of proposed algorithm

The following is the overall procedure of our algorithm:

1) Initialize the model parameters (11).
2) Iterate the following steps until convergence.
   - Update $\{q_{n,F,t,c}\}$ by using (14).
   - Update $\{\pi_{n,t,c}\}$ by using (19).
   - Iterate the following steps until convergence.
     - Calculate $J(\Theta)$ by using (22)-(25).
     - Update $\{W_f\}$ by using (26)-(27) iteratively.
   - Solve the permutation by using (28) and (29).
3) Calculate the separated signals by using (4) and (13).

III. Relation to prior work

Let us assume $\phi_c = 1$ $(c \in \{0,1\})$ in (10) and compare the proposed method with the conventional FD-ICA and IVA. Assume also that the source states are always active, i.e., $\pi_{n,t,1} = 1$ for $(n,t) \in [N] \times [N_T]$. Then, it holds that $q_{n,F,t,1} = 1$ for each $(n,F,t) \in [N] \times \mathcal{F} \times [N_T]$. Hence, in the case where $\mathcal{F} = \{[N_F]\}$ the proposed method is the same as the conventional (auxiliary function based) IVA [16], [17]. On the other hand, in the case where $\mathcal{F} = \{\{f\} \colon f \in [N_F]\}$ it is nothing but the conventional (auxiliary function based) FD-ICA [4], [18]. In this sense, the proposed method is regarded as an extension of FD-ICA and IVA.

Let us move on to explain the difference between the proposed model and that in [6], [7], [9], both of which can consider the statistical dependencies only within neighboring frequency bins. In our model, we further introduce the idea of switching priors according to the states of sources in a statistical sense while switching priors cannot be performed in the model of [6], [7], [9]. From the optimization viewpoint, to switch the priors, we need the normalization terms of contrast function g_c for each state c as described in Subsection II-E. However, since the source priors supposed in [6], [7], [9] have the frequency-overlapped form (while our model has the frequency-divided form), there are no closed-form expressions of the normalization constants for them. In that mean, our model is superior to that in [6], [7], [9].

IV. Experiment

A. Conditions

To confirm the effectiveness of the proposed method, we conducted an experiment using speech signals of two Japanese males recorded in a meeting room. In the experiment, we compared the following methods: AuxIVA1 (an auxiliary function based IVA with spherical distribution [16], [17]), AuxIVA2 (an auxiliary function based IVA with non-spherical distribution [9]), ILRMA [8], [14], [15], and three proposed methods indexed by $k \in \{1,2,4\}$. The evaluation data was created by adding microphone observation signals of each speaker recorded using a circular array with a diameter of 75 mm and eight microphones. In the experiment, we prepared mixtures ($N = 2$) of all the microphone combinations (${}_8C_2 = 28$ samples in total). The following four pairs of source directions were examined: $\{-30°, 30°\}$, $\{-60°, 60°\}$, $\{-90°, 90°\}$, $\{0°, 90°\}$.

TABLE I
Experimental conditions

Sampling rate	16 kHz
Frame length (points)	4069 (256 ms) or 8192 (512 ms)
Frame shift	Half of frame length
Window function	Hanning
Signal length	10 s
Source-microphone distance	1 m
Reverberation time (RT_{60})	710 ms

In the proposed method $k \in \{1,2,4\}$, the frequency range division $\mathcal{F}_k$ is set

$$\mathcal{F}_k = \{F_1, \ldots, F_k\} \tag{31}$$

$$F_i = \{\lfloor N_F - \frac{N_F}{2^{i-1}} \rfloor + 1, \ldots, \lfloor N_F - \frac{N_F}{2^i} \rfloor\} \quad \text{for } i = 1, \ldots, k-1 \tag{32}$$

$$F_k = [N_F] \setminus \sqcup_{i=1}^{k-1} F_i. \tag{33}$$

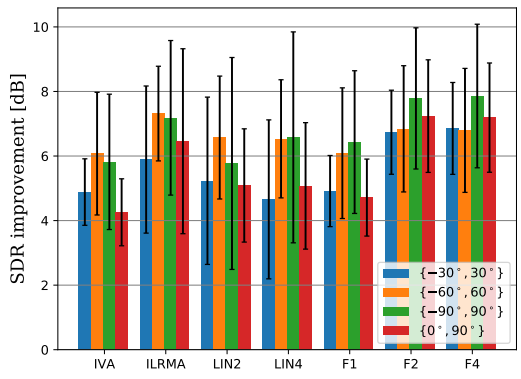
Also, we set $(\alpha_0, \beta_0) = (0.09, 0.1)$ and $(\alpha_1, \beta_1) = (1, 0.1)$ in (30) and $\phi_c = 5$ $(c \in \{0,1\})$ in (10). In the ILRMA, the partitioning function was not used, and the number of bases was set to two for each source, where ILRMA demonstrates the best performance for speech signals according to [8]. In AuxIVA1 and AuxIVA2, the contrast functions are set by (30) with $(\alpha, \beta) = (1, 0.1)$ (note that AuxIVA1 and AuxIVA2 do not consider source states). In AuxIVA2, we chose LIN2 and LIN4 as source models as defined in [7, Table 1]. In each method, the separation filters $\{W_f\}_f$ are initialized by the Identity matrix while the other parameters are identified by the value drawn from the uniform distribution on $(0,1) \subseteq \mathbb{R}$. In our method, the time-frequency masks $\{q_{n,F,t,c}\}_{n,F,t,c}$ of signals are updated as described in Subsection II-F, except that we set (not update) them by $1 - q_{n,F,t,0} = q_{n,F,t,1} = 0.999$ only for the first time.

In the experiment for each method, we optimized $\{W_f\}_f$ by an auxiliary function method, whose iteration number was set to 200 throughout the experiment. In the proposed method, the time-frequency mask was updated every 10 iterations. We used the average signal-to-distortion ratio (SDR [23]) of two speakers as an evaluation criterion. The other experimental conditions are described in Table I.

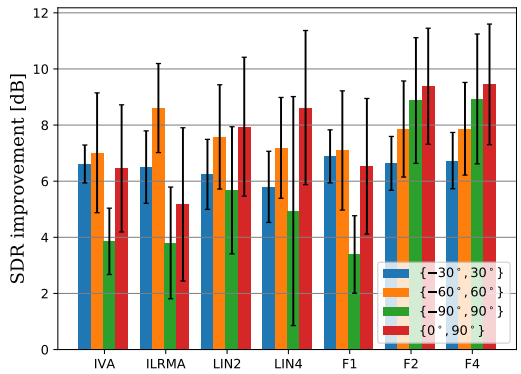
B. Results

Figure 1 shows the average scores and their deviations among the 28 samples. In the figure, models LIN2 and LIN4 show slightly better results than IVA as a whole. Also, the proposed model with the frequency range division $\mathcal{F}_1$ shows almost the same results as IVA. On the other hand, the proposed models with $\mathcal{F}_2$ and $\mathcal{F}_4$ show higher separation performance than IVA. The deviation of the scores is rather large, but these results suggest that our model better represents the distribution of speech signals by simultaneously considering (i) the statistical dependencies within neighboring frequency bins such as in LIN2 and LIN4 and (ii) the state of each source such as in the proposed model with $\mathcal{F}_1$.

ISBN 978-0-9928626-7-1 © EURASIP 2017



(a) Frame length of 4096 points



(b) Frame length of 8192 points

Fig. 1. Average SDR improvement and its deviation in four pairs of source directions when frame length is (a) 4096 points and (b) 8192 points. IVA corresponds to AuxIVA1, and LIN2 and LIN4 correspond to AuxIVA2. Also, F1, F2, and F4 denote proposed method with frequency range division $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_4$, respectively.

Compared with ILRMA, which models sources by using NMF, our models with $\mathcal{F}_2$ and $\mathcal{F}_4$ give similar scores when (a) the frame length is 4096 points. However, when (b) the frame length is 8192 points and the pair of source directions is $\{-90°, 90°\}$ or $\{0°, 90°\}$, our models show rather better results. The performance degradation of ILRMA in (b) is considered to occur because speech signals cannot be captured by co-occurrence of frequency components alone, which implies that our model is superior to that of ILRMA for speech signals.

V. Conclusion

This paper proposed a new generative model for IVA by introducing frequency range division and source prior switching to better represent the distribution of speech signals. In addition, we derived an optimization algorithm for the model, where the states of sources (active/inactive), IVA filters, and permutation alignments between each pair of frequency bands are jointly optimized. The experimental results suggest the effectiveness of the proposed model for speech signals.

References

[1] T. Kim, T. Eltoft, and T. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.

[2] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[3] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

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

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, vol. 46, John Wiley & Sons, 2004.

[6] G. J. Jang, I. Lee, and T. W. Lee, "Independent vector analysis using non-spherical joint densities for the separation of speech signals," in *Proc. ICASSP*, 2007, pp. II–629–II–632.

[7] I. Lee and G.-J. Jang, "Independent vector analysis based on overlapped cliques of variable width for frequency-domain blind signal separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, 2012.

[8] 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*, vol. 24, no. 9, pp. 1626–1641, 2016.

[9] Y. Liang, S. M. Naqvi, and J. Chambers, "Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm," *Electronics letters*, vol. 48, no. 8, pp. 460–462, 2012.

[10] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[11] J. Hirayama, S. Maeda, and S. Ishii, "Markov and semi-markov switching of source appearances for nonstationary independent component analysis," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1326–1342, 2007.

[12] A. Masnadi-Shirazi and B. Rao, "Independent vector analysis incorporating active and inactive states," in *Proc. ICASSP*, 2009, pp. 1837–1840.

[13] A. Masnadi-Shirazi, W. Zhang, and B. D. Rao, "Glimpsing IVA: A framework for overcomplete/
complete/undercomplete convolutive source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1841–1855, 2010.

[14] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. ICASSP*, 2015, pp. 276–280.

[15] D. Kitamura, "Algorithms for independent low-rank matrix analysis," http://d-kitamura.sakura.ne.jp/pdf/misc/AlgorithmsForIndependentLowRankMatrixAnalysis.pdf.

[16] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[17] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. APSIPA*, 2012, pp. 1–4.

[18] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. LVA/ICA*, 2010, pp. 165–172.

[19] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proc. SICE*, 2002, vol. 4, pp. 2138–2143.

[20] 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, pp. 1–24, 2001.

[21] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," *in Proc. ICASSP*, pp. 3238–3242, 2013.

[22] E. Gómez, M.A. Gómez-Viilegas, and J.M. Marín, "A multivariate generalization of the power exponential family of distributions," *Communications in Statistics - Theory and Methods*, vol. 27, no. 3, pp. 589–600, 1998.

[23] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1462–1469, 2006.

ISBN 978-0-9928626-7-1