IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 7, 2020, accepted March 21, 2020, date of publication April 6, 2020, date of current version April 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985842

Auxiliary-Function-Based Independent Vector Analysis Using Generalized Inter-Clique Dependence Source Models With Clique Variance Estimation

UI-HYEOP SHIN AND HYUNG-MIN PARK, (Senior Member, IEEE)
Department of Electronics Engineering, Sogang University, Seoul 04107, South Korea

Corresponding author: Hyung-Min Park (hpark@sogang.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) under Grant NRF-2017R1A2B4009964, and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korea Government (MSIT), through the Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding, under Grant 2016-0-00564.

ABSTRACT By introducing a frequency dependence source prior including full-band and clique models, independent vector analysis (IVA) has been successfully used for convolutive blind source separation (BSS). In addition, independent low-rank matrix analysis (ILRMA) learns a low-rank approximation of the time-frequency structure of source signals. This paper presents IVA using a clique-based frequency dependence model with time-varying clique variances to combine advantages of both ILRMA and clique-model-based IVA for BSS of speech signals. Although conventional clique models are effective in separating sources with specific spectral structures, the dependency among the cliques is considered by overlaps between cliques or a global clique of all frequency bins if there is. To avoid the permutation problem by strengthening the dependency among the cliques, we develop a generalized probability-density-function (pdf) model imposing a variable exponent on the summed cliques with overlaps and time-varying clique variances, which may include most conventional source models as particular cases. In addition, update rules of the clique variances and demixing matrices are derived by minimization of the cost function of BSS as well as non-negative matrix factorization (NMF) and auxiliary function techniques for fast and robust convergence, respectively. Through experiments on BSS of speech mixtures with various mixing conditions, the proposed IVA showed improved separation performance than the conventional methods. Experimental results consistently demonstrated that the performance of a method could be determined in general by the trade-off between the degree of freedom of source models (as long as model parameters were accurately estimated) and the vulnerability to the permutation problem.

INDEX TERMS Blind source separation, clique, independent vector analysis, time-varying variance.

I. INTRODUCTION

Blind source separation (BSS) recovering source signals from their mixtures without knowing the mixing process has still been one of the most interesting topics in the field of signal processing [1], [2]. If multiple mixtures are available, instead of single-channel separation such as deep clustering [3], independent component analysis (ICA), which is a signal processing method that expresses multivariate data as linear combinations of statistically independent random variables, has attracted considerable interest because of its successful performance in many BSS applications [1], [4]. Because acoustic mixing in real-world situations involves complex reverberations, ICA has been extended to the deconvolution of mixtures in both the time and frequency domains. Although the frequency-domain approach is generally preferred because of the intensive computations and slow convergence of the time-domain approach, it has the permutation problem caused by the random permutation of the separated frequency components [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Pia Addabbo.

Independent vector analysis (IVA) can effectively mitigate this problem and improve the separation performance by introducing a source prior with a full-band radially symmetric joint probability density function (pdf) that assumes the uniform dependency across frequency instead of using an independent prior at each frequency bin [5]–[7]. Then, improved frequency dependence models were presented by assigning the dependency among close frequency bins by using subband local cliques [8] and by introducing a harmonic clique model that was more effective in separating sound sources with strong harmonic structures than the conventional model [9]. Since these full-band or clique models are fixed in advance without considering actual sources, separation performance may be poor by using models different from those of sources. Especially, the performance is significantly degraded due to the block permutation problem when some sources have similar spectral properties [10].

The original IVA adopted the natural-gradient-based optimization [5]. However, a large step size to speed up the convergence may result in diverged parameters whereas a small step size to avoid the divergence may lead to slow convergence. Therefore, the natural gradient update rule for IVA has a trade-off between the convergence speed and the stability like other gradient-based methods. An IVA method using an adaptive step size was proposed to increase the convergence speed [11], and the fixed-point IVA method was derived based on Newton method [12]. Especially, a fast and robust update rule was developed that was based on auxiliary function techniques as an extension of the expectation-maximization algorithm, AuxIVA, which improved the convergence speed further without requiring environment-sensitive parameters such as the step-size parameter [13]–[15]. AuxIVA also employed clique models to obtain improved performance [10], [16]. Recently, independent low-rank matrix analysis (ILRMA), that learns a low-rank approximation of the time-frequency structure of source signals by using nonnegative matrix factorization (NMF) [17], [18] in addition to independence between sources, achieved remarkable performance for separating music signals [19], [20]. Instead of a complex Gaussian distribution as a source model in the conventional ILRMA, a complex Student's t-distribution including a complex Cauchy distribution as a special case was employed in the framework of ILRMA to consider sources with heavy tails [21]. In addition, the separation performance of ILRMA for speech mixtures was improved by imposing strong dependencies between neighboring frequency bins with source models of multivariate complex exponential power distribution [22].

In this paper, we propose IVA using a multivariate clique-based frequency dependence model with time-varying clique variances to combine advantages of both ILRMA and clique-model-based IVA for BSS of speech signals. Although the conventional clique-model-based IVA is effective in separating sources with specific spectral structures, the dependency among the cliques is considered by overlaps between adjacent cliques or a global clique of all frequency bins if there is. By strengthening the dependency among the cliques in the multivariate pdf model of sources and also by introducing time-varying clique variances, the block permutation problem can be effectively addressed. Instead of the bases with a dimension of the number of frequency bins to find the frequency-dependent time-varying variances in the conventional ILRMA, the proposed method requires the bases with a dimension of the number of cliques that is much less than the number of frequency bins. In particular, we develop a generalized pdf model imposing a variable exponent on the summed cliques with overlaps, which may include most of the conventional source pdf models as particular cases. In addition, update rules of the clique variances and demixing matrices are derived by minimization of the cost function of BSS as well as NMF and auxiliary function techniques for fast and robust convergence, respectively.

The remainder of this paper is organized as follows. Section II describes the BSS problem and proposed AuxIVA developed from the conventional methods. The performance of the proposed method is evaluated in Section III, and some concluding remarks are presented in Section IV.

II. PROPOSED AuxIVA

A. PROBLEM FORMULATION

Let us consider M observations that are convolutive mixtures of M mutually independent unknown sources. Assuming the length of the window function for the short-time Fourier transform (STFT) is sufficiently longer than the effective length of the mixing filter, the convolution in the time domain is approximately transformed into multiplication in the frequency domain as follows: [23]

$$\mathbf{x}(k,\tau) \approx \mathbf{A}(k)\mathbf{s}(k,\tau), \tag{1}$$

where $\mathbf{x}(k,\tau) = [X_1(k,\tau),\cdots,X_M(k,\tau)]^T$ and $\mathbf{s}(k,\tau) = [S_1(k,\tau),\cdots,S_M(k,\tau)]^T$ denote vectors composed of the time-frequency segments of mixture and source signals, respectively, at frequency bin k and frame τ. $\mathbf{A}(k)$ represents a mixing matrix at the k-th frequency bin. Like many natural sounds including speech, dependencies are assumed to exist among frequency components of a source corresponding to the elements of $\hat{\mathbf{s}}_m(\tau) = [S_m(1,\tau),\cdots,S_m(K,\tau)]^T$, $m = 1,\cdots,M$, where K is the number of frequency bins.

The aim of BSS for convolutive mixtures is to restore source signals by estimating separating matrices such that

$$\mathbf{y}(k,\tau) = \mathbf{W}(k)\mathbf{x}(k,\tau), \tag{2}$$

where $\mathbf{y}(k,\tau) = [Y_1(k,\tau),\cdots,Y_M(k,\tau)]^T$ denotes a vector composed of the time-frequency segments of estimated source signals at frequency bin k and frame τ, and $\mathbf{W}(k)$ is a demixing matrix at the k-th frequency bin.

Since IVA accomplishes BSS by finding a linear transform of mixtures to obtain statistically independent signals that corresponds to source signals $\mathbf{s}(k,\tau)$, the demixing matrices are estimated by minimizing the dependency between estimated time-frequency segments of source signals measured by the Kullback-Leibler (KL) divergence between an exact

joint pdf $p(\hat{\mathbf{y}}_1(\tau), \cdots, \hat{\mathbf{y}}_M(\tau))$ and the product of hypothesized pdf models of the estimated source $\prod_{m=1}^{M} q(\hat{\mathbf{y}}_m(\tau))$ given as [4], [24]

$$\begin{aligned} &D\left(p(\hat{\mathbf{y}}_1(\tau), \cdots, \hat{\mathbf{y}}_M(\tau))|| \prod_{m=1}^{M} q(\hat{\mathbf{y}}_m(\tau))\right) \\ &= \int p(\hat{\mathbf{y}}_1(\tau), \cdots, \hat{\mathbf{y}}_M(\tau)) \log \frac{p(\hat{\mathbf{y}}_1(\tau), \cdots, \hat{\mathbf{y}}_M(\tau))}{\prod_{m=1}^{M} q(\hat{\mathbf{y}}_n(\tau))} \\ &\quad \times d\hat{\mathbf{y}}_1(\tau), \cdots, d\hat{\mathbf{y}}_M(\tau) \\ &= \text{const.} - \sum_{k=1}^{K} \log|\det \mathbf{W}(k)| - \sum_{m=1}^{M} E\{\log q(\hat{\mathbf{y}}_m(\tau))\}, \end{aligned} \tag{3}$$

where $\hat{\mathbf{y}}_m(\tau) = [Y_m(1, \tau), \cdots, Y_m(K, \tau)]^T$, $m = 1, \cdots, M$ and $E\{\cdot\}$ denotes the expectation operator.

B. CONVENTIONAL MULTIVARIATE PDF MODELS OF SOURCES

In the original IVA [5] and AuxIVA [13], the multivariate pdf model of sources assumes uniform dependency among frequency components by using a full-band radially symmetric joint pdf expressed as

$$\begin{aligned} q(\hat{\mathbf{y}}_m(\tau)) &\propto \exp\{-||\hat{\mathbf{y}}_m(\tau)||_2\} \\ &= \exp\left\{-\sqrt{\sum_{k=1}^{K} |Y_m(k, \tau)|^2}\right\}, \end{aligned} \tag{4}$$

where $||\cdot||_2$ denotes the L_2 norm of a vector. The pdf can be generalized to be given as [14]

$$q(\hat{\mathbf{y}}_m(\tau)) \propto \exp\{-||\hat{\mathbf{y}}_m(\tau)||_2^{\beta}\}, \tag{5}$$

where β is a shape parameter of the generalized Gaussian distribution within $(0,\ 2]$. Improved frequency dependence models for IVA are developed by using subband local cliques to assign the dependency among close frequency bins [8], and by introducing a harmonic clique model to exploit sound sources with strong harmonic structures, such as speech and music signals [9]. The pdf can be expressed as [8], [9], [16]

$$q(\hat{\mathbf{y}}_m(\tau)) \propto \exp\left\{-\sum_{c=1}^{C} \sqrt{\sum_{k \in \Omega_c} |Y_m(k, \tau)|^2}\right\}, \tag{6}$$

where C and Ω_c denote the number of cliques and a set of frequency bins that belongs to the c-th clique. In [8], Ω_c consists of consecutive frequency bins, and the series of cliques have chain-like overlaps. In [9], the fundamental frequency of the c-th harmonic clique, F_c, is defined as

$$F_c = F_1 \times 2^{(c-1)/r_f}, \tag{7}$$

where $F_1 = 55$ Hz and r_f denotes a parameter to determine the resolution of harmonic cliques. For each clique, Ω_c consists of the frequency bins of the first eight to ten multiples of F_c. The bandwidth of the h-th multiple of F_c is $2\delta h F_c$, and δ is a parameter to determine the degree of overlap for two consecutive harmonic cliques. Another clique consisting of all the frequency bins is added to prevent the permutation of frequency bins below 55 Hz and to increase the learning speed of demixing matrices.

Instead of these stationary models, a non-stationary Gaussian distribution is also introduced to model non-stationary acoustic sound sources such as speech, as follows: [14]

$$q(\hat{\mathbf{y}}_m(\tau)) \propto \exp\left\{-\frac{||\hat{\mathbf{y}}_m(\tau)||_2^2}{\lambda'_m(\tau)}\right\}, \tag{8}$$

where $\lambda'_m(\tau)$ represents the time-varying variance at frame τ for the m-th source. As an efficient method combining specific spectral structures, such as the harmonic structures, and non-stationarity of sound sources, in addition, the time-varying variance in ILRMA is estimated by NMF decomposition, which can be formulated as [20], [21], [25]

$$q(\hat{\mathbf{y}}_m(\tau)) \propto \exp\left\{-\left(\sum_{k=1}^{K} \frac{|Y_m(k, \tau)|^2}{\lambda''_m(k, \tau)}\right)^{\beta}\right\} \tag{9}$$

and

$$\lambda''_m(k, \tau) = \sum_{l=1}^{L} t''_m(k, l) v''_m(l, \tau), \tag{10}$$

where $\lambda''_m(k, \tau)$ denotes the frequency-dependent time-varying variance at frequency bin k and frame τ for the m-th source. $t''_m(k, l)$ and $v''_m(l, \tau)$ are the corresponding basis and activation, respectively, and L is the number of bases.

In the previous method, the dimension of bases $t''_m(k, l)$ in (10) for the pdf of non-stationary sources in (9) may be too large to be accurately estimated especially for insufficient input data. In order to reduce the dimension of bases, the frequency range division is introduced to group consecutive frequency bins by using a time-varying variance for a frequency range [22]. Therefore, the source model is changed into

$$q(\hat{\mathbf{y}}_m(\tau)) \propto \exp\left\{-\sum_{d=1}^{D}\left(\frac{\sum_{k=d_f}^{d_g} |Y_m(k, \tau)|^2}{\lambda'''_m(d, \tau)}\right)^{\gamma}\right\} \tag{11}$$

and

$$\lambda'''_m(d, \tau) = \left(\sum_{l=1}^{L} t'''_m(d, l) v'''_m(l, \tau)\right)^{\epsilon}, \tag{12}$$

where D and γ are the number of frequency ranges and a shape parameter, respectively. d_f and d_g denote the first and last frequency bins for frequency range d. $\lambda'''_m(d, \tau)$ denotes the frequency-range-dependent time-varying variance at range d and frame τ for the m-th source while $t'''_m(d, l)$ and $v'''_m(l, \tau)$ are the corresponding basis and activation with a power of ϵ, respectively.

C. PROPOSED MULTIVARIATE PDF MODELS OF SOURCES

Although IVA using the clique models of (6) is more effective in separating sound sources with specific spectral structures than the original method using a full-band model assuming uniform dependency across frequency, the dependency among the cliques is considered by overlaps between adjacent cliques and one additional global clique of all frequency bins in [9]. Since the square root applied to summed frequency components in (4) imposes dependency among the frequency components, we impose the dependency among cliques explicitly by formulating the pdf model as

$$q(\hat{\mathbf{y}}_m(\tau)) \propto \exp\left\{-\sqrt{\sum_{c=1}^{C}\sqrt{\sum_{k\in\Omega_c}|Y_m(k,\tau)|^2}}\right\}. \quad (13)$$

For the sake of simplicity, $\sqrt{\sum_{k\in\Omega_c}|Y_m(k,\tau)|^2}$ is denoted as $\tilde{y}_m(c,\tau)$ from now on. Extending the non-overlapped frequency ranges in (11) to clique models that allow various types including overlaps and harmonics, and imposing the dependency among cliques, another pdf model is derived as

$$q(\hat{\mathbf{y}}_m(\tau)) \propto \exp\left\{-\sqrt{\sum_{c=1}^{C}\left(\frac{\tilde{y}_m^2(c,\tau)}{\lambda_m(c,\tau)}\right)^{\gamma}}\right\} \quad (14)$$

and

$$\lambda_m(c,\tau) = \left(\sum_{l=1}^{L} t_m(c,l)v_m(l,\tau)\right)^{\epsilon}, \quad (15)$$

where $\lambda_m(c,\tau)$ denotes the clique-dependent time-varying variance at clique c and frame τ for the m-th source. $t_m(c,l)$ and $v_m(l,\tau)$ are the corresponding basis and activation, respectively.

Generalizing the pdf models in (13) and (14), we develop a generalized pdf model given as

$$q(\hat{\mathbf{y}}_m(\tau)) = \frac{\alpha}{\prod_{c=1}^{C}\lambda_m^{N_c}(c,\tau)}\exp\left\{-\left[\sum_{c=1}^{C}\left(\frac{\tilde{y}_m^2(c,\tau)}{\lambda_m(c,\tau)}\right)^{\gamma}\right]^{\beta}\right\}, \quad (16)$$

where α and N_c denote a constant and the number of frequency bins in Ω_c, respectively. Shape parameters β and γ are between 0 and 1 in common. If both β and γ are 0.5 and $\lambda_m(c,\tau)$ is fixed to 1, the pdf model of (16) becomes (13). If β is 0.5, the pdf model of (16) corresponds to (14). γ and $\lambda_m(c,\tau)$ impose the intra-clique dependency. Although the inter-clique dependency is implicitly considered by the overlaps between cliques, β imposes the inter-clique dependency explicitly.

D. DERIVATION OF THE PROPOSED AuxIVA

Using the generalized pdf model of (16) with replacement of (15), the KL divergence of (3) without the constant term is used as the cost function for the proposed method, which is

$$\begin{aligned} J = &-\sum_{k=1}^{K}\log|\det\mathbf{W}(k)| \\ &+\sum_{m=1}^{M}E\left\{\left[\sum_{c=1}^{C}\left(\frac{\tilde{y}_m^2(c,\tau)}{\left(\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)\right)^{\epsilon}}\right)^{\gamma}\right]^{\beta}\right. \\ &\left.+\sum_{c=1}^{C}N_c\epsilon\log\left(\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)\right)\right\}. \end{aligned} \quad (17)$$

In order to derive update rules for NMF parameters, the partial derivatives of the cost function with respect to the parameters are written as

$$\begin{aligned} &\frac{\partial J}{\partial t_m(c,l)} \\ &= E\left\{-\epsilon\zeta_m(\tau)\eta_m(c,\tau)\frac{\tilde{y}_m^2(c,\tau)v_m(l,\tau)}{\left(\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)\right)^{\epsilon+1}}\right. \\ &\left.+N_c\epsilon\frac{v_m(l,\tau)}{\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)}\right\} \end{aligned} \quad (18)$$

and

$$\begin{aligned} &\frac{\partial J}{\partial v_m(l,\tau)} \\ &= -\epsilon\zeta_m(\tau)\sum_{c=1}^{C}\eta_m(c,\tau)\frac{\tilde{y}_m^2(c,\tau)t_m(c,l)}{\left(\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)\right)^{\epsilon+1}} \\ &+\sum_{c=1}^{C}N_c\epsilon\frac{t_m(c,l)}{\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)}, \end{aligned} \quad (19)$$

where $\zeta_m(\tau)$ and $\eta_m(c,\tau)$ are as follows:

$$\zeta_m(\tau) = \beta\left[\sum_{c=1}^{C}\left(\frac{\tilde{y}_m^2(c,\tau)}{\left(\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)\right)^{\epsilon}}\right)^{\gamma}\right]^{\beta-1}, \quad (20)$$

$$\eta_m(c,\tau) = \gamma\left(\frac{\tilde{y}_m^2(c,\tau)}{\left(\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)\right)^{\epsilon}}\right)^{\gamma-1}. \quad (21)$$

Therefore, we have the update rules expressed as

$$\begin{aligned} t_m(c,l) = &\, t_m(c,l) \\ &\cdot\left[\frac{E\left\{\epsilon\zeta_m(\tau)\eta_m(c,\tau)\frac{\tilde{y}_m^2(c,\tau)v_m(l,\tau)}{\left(\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)\right)^{\epsilon+1}}\right\}}{E\left\{N_c\epsilon\frac{v_m(l,\tau)}{\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)}\right\}}\right]^{\kappa} \end{aligned} \quad (22)$$

and

$$v_m(l,\tau) = v_m(l,\tau) \cdot \left[\frac{\epsilon\zeta_m(\tau)\sum_{c=1}^{C}\eta_m(c,\tau)\frac{\tilde{y}_m^2(c,\tau)t_m(c,l)}{\left(\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)\right)^{\epsilon+1}}}{\sum_{c=1}^{C}N_c\epsilon\frac{t_m(c,l)}{\sum_{l=1}^{L}t_m(c,l)v_m(l,\tau)}} \right]^{\kappa}, \tag{23}$$

where κ is a learning parameter to be determined for stable and fast convergence.

Without the decomposition of $\lambda_m(c,\tau)$ in (15), $\lambda_m(c,\tau)$ can be directly estimated by minimizing the cost function of (17) as follows:

$$\frac{\partial J}{\partial\lambda_m(c,\tau)} = -\zeta_m(\tau)\eta_m(c,\tau)\frac{\tilde{y}_m^2(c,\tau)}{\lambda_m^2(c,\tau)} + N_c\frac{1}{\lambda_m(c,\tau)} = 0, \tag{24}$$

$$\lambda_m(c,\tau) = \frac{1}{N_c}\zeta_m(\tau)\eta_m(c,\tau)\tilde{y}_m^2(c,\tau), \tag{25}$$

where

$$\zeta_m(\tau) = \beta\left[\sum_{c=1}^{C}\left(\frac{\tilde{y}_m^2(c,\tau)}{\lambda_m(c,\tau)}\right)^{\gamma}\right]^{\beta-1}, \tag{26}$$

$$\eta_m(c,\tau) = \gamma\left(\frac{\tilde{y}_m^2(c,\tau)}{\lambda_m(c,\tau)}\right)^{\gamma-1}. \tag{27}$$

In order to introduce an auxiliary function for the cost function of (17), let us select $-\log q(\hat{\mathbf{y}}_m(\tau))$ for a real-valued continuous and differentiable function $G_R(r)$ satisfying that $G_R'(r)/r$ is continuous everywhere and monotonically decreasing in $r \geq 0$, where a real variable r is equal to

$$\rho_m(\tau) = \sqrt{\sum_{c=1}^{C}\left(\frac{\tilde{y}_m^2(c,\tau)}{\lambda_m(c,\tau)}\right)^{\gamma}}. \tag{28}$$

Then,

$$G_R(r) \leq \frac{G_R'(r_0)}{2r_0}r^2 + G_R(r_0) - \frac{r_0G_R'(r_0)}{2} \tag{29}$$

holds for any r and a real constant r_0. The equality is satisfied if and only if $r_0 = r$ [13], [26]. Using

$$G_R'(\rho_m(\tau)) = 2\beta\rho_m^{2\beta-1}(\tau), \tag{30}$$

$$\begin{aligned} E\{G_R(\rho_m(\tau))\} &\leq E\left\{\frac{G_R'(\rho_m(\tau))}{2\rho_m(\tau)}\rho_m^2(\tau)\right\} + R \\ &= E\left\{\beta\rho_m^{2\beta-2}(\tau)\sum_{c=1}^{C}\left(\frac{\tilde{y}_m^2(c,\tau)}{\lambda_m(c,\tau)}\right)^{\gamma}\right\} + R \\ &= E\left\{\sum_{c=1}^{C}\beta\rho_m^{2\beta-2}(\tau)\tilde{y}_m^2(c,\tau)\frac{\tilde{y}_m^{2\gamma-2}(c,\tau)}{\lambda_m^{\gamma}(c,\tau)}\right\} + R, \end{aligned} \tag{31}$$

where R is a constant term independent of $\mathbf{W}(k)$. Since $\tilde{y}_m^2(c,\tau) = \sum_{k\in\Omega_c}|Y_m(k,\tau)|^2$ and $Y_m(k,\tau) = \mathbf{w}_m^H(k)\mathbf{x}(k,\tau)$ with the m-th row vector $\mathbf{w}_m^H(k)$ of $\mathbf{W}(k)$,

$$E\{G_R(\rho_m(\tau))\} \leq \sum_{c=1}^{C}\sum_{k\in\Omega_c}\mathbf{w}_m^H(k)\mathbf{V}_m(c,k)\mathbf{w}_m(k) + R, \tag{32}$$

where H denotes the Hermitian transpose, and

$$\mathbf{V}_m(c,k) = E\left\{\beta\rho_m^{2\beta-2}(\tau)\frac{\tilde{y}_m^{2\gamma-2}(c,\tau)}{\lambda_m^{\gamma}(c,\tau)}\mathbf{x}(k,\tau)\mathbf{x}^H(k,\tau)\right\}. \tag{33}$$

In (32), minimizing $\sum_{c=1}^{C}\sum_{k\in\Omega_c}\mathbf{w}_m^H(k)\mathbf{V}_m(c,k)\mathbf{w}_m(k) - \log|\det\mathbf{W}(k)|$ with respect to $\mathbf{w}_m(k)$ yields an optimal $\mathbf{w}_m(k)$ for the cost function of (17). Therefore, the equation to find an optimal $\mathbf{w}_m(k)$ is

$$\sum_{c\in\Psi_k}\mathbf{V}_m(c,k)\mathbf{w}_m(k) - \nabla_{\mathbf{w}_m^*(k)}\log|\det\mathbf{W}(k)| = 0, \tag{34}$$

where Ψ_k denotes a set of cliques that includes the k-th frequency bin. Since $(\partial/\partial\mathbf{W}(k))\det\mathbf{W}(k) = \mathbf{W}^{-H}(k)\det\mathbf{W}(k)$, (34) is rearranged to give [13], [14]

$$\mathbf{w}_n^H(k)\sum_{c\in\Psi_k}\mathbf{V}_m(c,k)\mathbf{w}_m(k) = \delta_{nm}, \quad 1\leq n\leq M,\ 1\leq m\leq M. \tag{35}$$

Because a closed-form solution for updating all of $\mathbf{w}_m(k)$ simultaneously is still an open problem, a sequential update of $\mathbf{w}_m(k)$ while fixing the other vectors $\mathbf{w}_n(k)$ $(n \neq m)$ is considered similar to [13], [14]. Then, the update of $\mathbf{w}_m(k)$ can be simply given by

$$\mathbf{w}_m(k) \leftarrow \left(\mathbf{W}(k)\sum_{c\in\Psi_k}\mathbf{V}_m(c,k)\right)^{-1}\mathbf{e}_m, \tag{36}$$

where $\mathbf{e}_m$ denotes the unit vector with the m-th element of unity. Although the updated $\mathbf{w}_m(k)$ should be normalized to hold for $\mathbf{w}_m^H(k)\sum_{c\in\Psi_k}\mathbf{V}_m(c,k)\mathbf{w}_m(k) = 1$ in (35), different numbers of elements in Ψ_k of $\sum_{c\in\Psi_k}\mathbf{V}_m(c,k)$ can cause an imbalance of the scales of $\mathbf{w}_m(k)$, $k = 1,\cdots,K$. The imbalance may be avoided as follows:

$$\mathbf{w}_m(k) \leftarrow \frac{\mathbf{w}_m(k)}{\sqrt{\frac{1}{N_{\Psi_k}^{\nu}}\mathbf{w}_m^H(k)\sum_{c\in\Psi_k}\mathbf{V}_m(c,k)\mathbf{w}_m(k)}}, \tag{37}$$

where N_{Ψ_k} and ν denote the number of elements in Ψ_k and a balancing parameter, respectively. To prevent possible divergence of $\mathbf{w}_m(k)$, furthermore, the scale normalization without the burden of large calculations should be performed by

$$\mathbf{W}(k) \leftarrow \frac{\mathbf{W}(k)}{\sqrt{\frac{1}{MK}\sum_{k=1}^{K}E\left\{||\mathbf{y}(k,\tau)||_2^2\right\}}} \tag{38}$$

for each iteration such that $E\left\{|Y_m(k,\tau)|^2\right\} = 1$ is satisfied [14].

In summary, the overall procedure of the proposed method[1] is as follows:

- Begin
  - Step 1 Transform input data into $\mathbf{x}(k, \tau)$ in the time-frequency domain using the STFT.
  - Step 2 Initialize $\mathbf{W}(k)$, $t_m(c, l)$, and $v_m(l, \tau)$.
  - Step 3 Compute $\mathbf{y}(k, \tau)$ by (2) with the current $\mathbf{W}(k)$.
  - Step 4 Compute $\tilde{y}_m(c, \tau) = \sqrt{\sum_{k \in \Omega_c} |Y_m(k, \tau)|^2}$.
  - Step 5 Compute $\zeta_m(\tau)$ and $\eta_m(c, \tau)$ by (20) and (21), respectively.
  - Step 6 Update $t_m(c, l)$ and $v_m(l, \tau)$ by (22) and (23), respectively.
  - Step 7 Compute $\lambda_m(c, \tau)$ by (15).
  - Step 8 Compute $\rho_m(\tau)$ by (28).
  - Step 9 Compute $\mathbf{V}_m(c, k)$ by (33).
  - Step 10 Update $\mathbf{w}_m(k)$ by (36) and (37).
  - Step 11 Update $\mathbf{W}(k)$ by (38).
  - Step 12 Go to Step 3 until convergence.
  - Step 13 Estimate source signals from $\mathbf{y}(k, \tau)$ using the inverse STFT and the overlap-add method [27].
- End

With the direct estimation of $\lambda_m(c, \tau)$, Step 2 is replaced by "Initialize $\mathbf{W}(k)$ and $\lambda_m(c, \tau)$." Also, Step 5 is replaced by "Compute $\zeta_m(\tau)$ and $\eta_m(c, \tau)$ by (26) and (27), respectively." Steps 6 and 7 are replaced by "Compute $\lambda_m(c, \tau)$ by (25)."[2] Using the pdf model of (16) with $\lambda_m(c, \tau) = 1$, Step 2 is replaced by "Initialize $\mathbf{W}(k)$," and Steps 5, 6, and 7 are skipped.[3] In practice, the expectation $E\{\cdot\}$ in these equations can be replaced by the sample mean.

III. EXPERIMENTAL EVALUATION

To evaluate the performance of the proposed algorithm, we conducted an experiment using the live-recorded speech data obtained from underdetermined BSS tasks in SiSEC2011 [28]. Figs. 1 and 2 describe configurations of sources and microphones in dev1 and dev2 datasets, respectively. In each dev. dataset, one female and one male utterances were used as source signals in each source position. Each utterance was recorded at microphones in Fig. 1 with reverberation times of 130 ms and 250 ms for the dev1 dataset whereas it was recorded in Fig. 2 with a reverberation time of 250 ms for the dev2 dataset. Since the provided data are stereo recordings at a sampling rate of 16 kHz with two pairs of microphones whose distances were 1 m and 5 cm as shown in Figs. 1 and 2, we selected recorded data for two source locations among four locations and summed up the data at the two microphones to obtain stereo mixtures for determined BSS problems. Since there were female and male source signals at each source location, 24 different mixture sets[4] were generated for a configuration at a reverberation time.

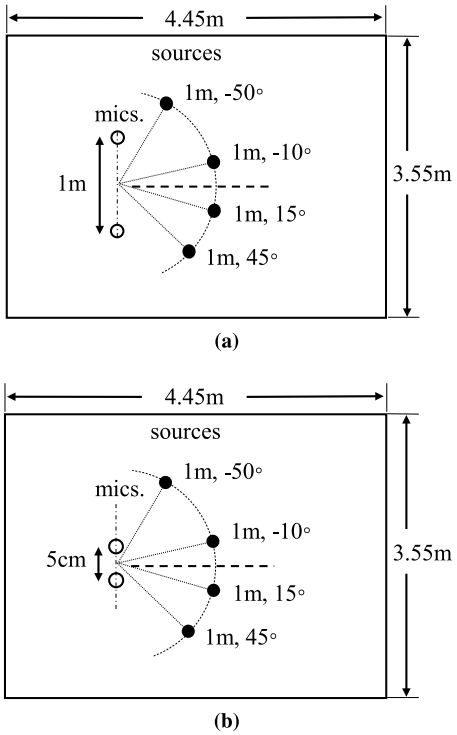

FIGURE 1. Source and microphone positions using (a) 1-m-apart microphones and (b) 5-cm-apart microphones in dev1 dataset of the live-recorded speech data in SiSEC2011 [28]. The common room height is 2.5 m.

The proposed AuxIVA-ICD, ILRMA-ICD, and AuxIVA-TVV-ICD for subband and harmonic cliques were compared with conventional AuxIVA [13]–[15], AuxIVA for subband and harmonic cliques [10], [16], conventional ILRMA [19], [20], and ILRMA based on multivariate complex exponential power distribution (ILRMA-MEPD) [22]. In the ILRMA-MEPD, two and eight frequency range divisions were considered which showed better performance than others in [22], and multivariate complex Gaussian distribution was used as the source pdf. The commonly used subband and harmonic cliques are shown in Fig. 3. The seven subband cliques were the same as in [9] where the first and the last frequency bins of the c-th clique were $128(c-1)+1$ and $128(c+1)$, respectively. Regarding the harmonic cliques, r_f and δ were set to 10 and $1-2^{-1/12}$, respectively,[5] and each harmonic clique consisted of the frequency bins of the first ten multiples of F_c. Since the physical frequency at the k-th frequency bin is $\frac{F_s}{K}k$ with the sampling frequency F_s, the indices k of the frequency bins belonging to the h-th multiple(harmonics) of the c-th clique among 39 cliques except for

[1] Let us refer to this method as ILRMA using source pdf models with inter-clique dependence (ILRMA-ICD) from now on.

[2] Let us refer to this method as AuxIVA using time-varying-variance source pdf models with inter-clique dependence (AuxIVA-TVV-ICD) from now on.

[3] Let us refer to this method as AuxIVA using source pdf models with inter-clique dependence (AuxIVA-ICD) from now on.

[4] Six and twelve sets were generated from the same and different gender source signals, respectively.

[5] Using the selected value for δ, two consecutive harmonic cliques overlapped by less than 50 % to avoid too large cliques obtained when cliques of adjacent multiples were linked due to their large bandwidths.

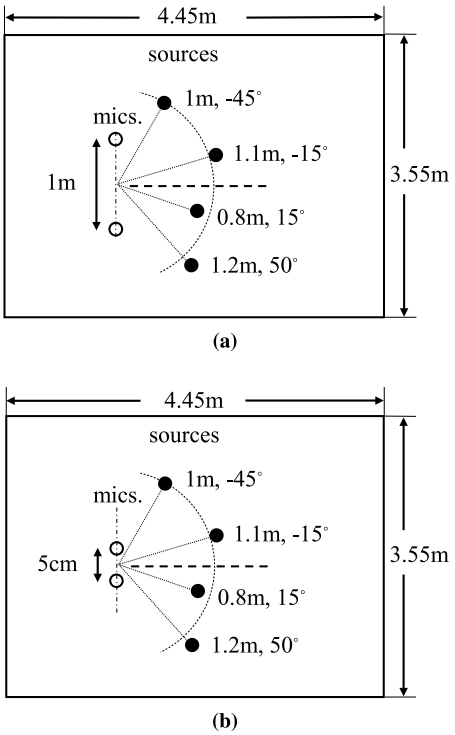

FIGURE 2. Source and microphone positions using (a) 1-m-apart microphones and (b) 5-cm-apart microphones in dev2 dataset of the live-recorded speech data in SiSEC2011 [28]. The common room height is 2.5 m.

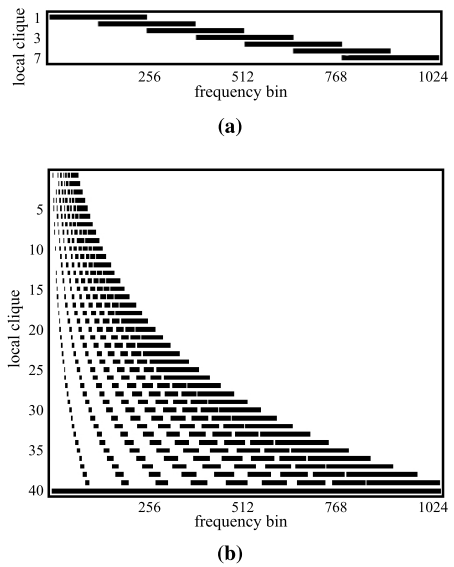

FIGURE 3. Plot of local clique index versus frequency bin for (a) subband and (b) harmonic clique frequency dependence source models.

the last full-band clique should satisfy the following equation:

$$(1-\delta)hF_c \le \frac{F_s}{K}k \le (1+\delta)hF_c. \tag{39}$$

In all ILRMA methods, L, ϵ, and κ were set to 2, 1, and 0.5, respectively. Using the signal-to-distortion ratio (SDR) [29] in decibels, the performance was measured by an SDR

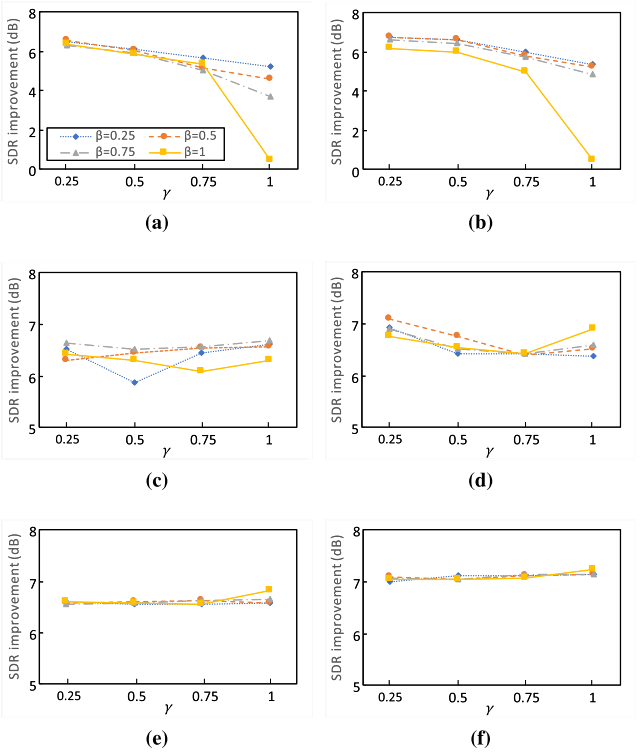

FIGURE 4. SDR improvements averaged over 72 mixture sets with several shape parameter values at configurations in Figs. 1b and 2b with reverberation times of 130 ms and 250 ms using (a) AuxIVA-ICD for the subband clique model, (b) AuxIVA-ICD for the harmonic clique model, (c) ILRMA-ICD for the subband clique model, (d) ILRMA-ICD for the harmonic clique model, (e) AuxIVA-TVV-ICD for the subband clique model, and (f) AuxIVA-TVV-ICD for the harmonic clique model.

improvement between output and input signals. In all experiments, we applied a 2048-point fast Fourier transform (FFT) with a shift size of 512 samples to input signals for frequency analysis. For all the experimented methods, the demixing matrix was initialized by an identity matrix in each frequency bin.

In order to show the dependency of the performances of AuxIVA-ICD, ILRMA-ICD, and AuxIVA-TVV-ICD on the values of shape parameters β and γ, Fig. 4 shows SDR improvements averaged over 72 mixture sets at configurations in Figs. 1b and 2b with reverberation times of 130 ms and 250 ms. The performance of AuxIVA-ICD was much more dependent than the others because the only method to impose the frequency dependence of source signals except overlaps between cliques was shape parameter values with $\lambda_m(c,\tau) = 1$. Inter- and intra-clique dependencies are lost with $\beta = 1$ and $\gamma = 1$, respectively, which resulted in significant performance degradation of AuxIVA-ICD. For the subband clique model, the inter-clique dependence is large enough because there are sufficient overlaps between cliques. Unlike Fig. 4b, therefore, there was little performance improvement in Fig. 4a by making β less than 1 to impose the inter-clique dependence except that $\gamma = 1$ corresponding to the case without the intra-clique dependence.

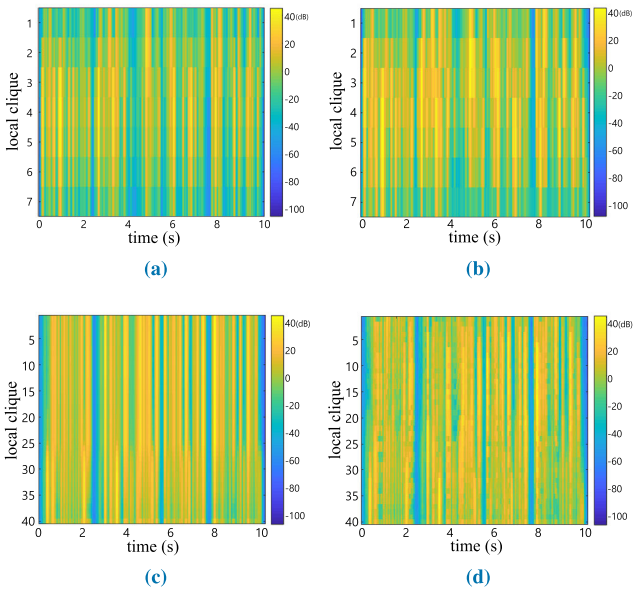

FIGURE 5. Plot of an example of time-varying clique variances at a configuration in Fig. 1b with a reverberation time of 130 ms using (a) ILRMA-ICD for the subband clique model, (b) AuxIVA-TVV-ICD for the subband clique model, (c) ILRMA-ICD for the harmonic clique model, and (d) AuxIVA-TVV-ICD for the harmonic clique model.

SDR improvements of ILRMA-ICD and AuxIVA-TVV-ICD were less dependent on shape parameter values than those of AuxIVA-ICD because time-varying clique variance $\lambda_m(c, \tau)$ with overlaps between cliques in addition to the shape parameters can impose the frequency dependence of source signals. Fig. 5 displays the time-varying clique variances of ILRMA-ICD and AuxIVA-TVV-ICD for a mixture set at a configuration in Fig. 1b with a reverberation time of 130 ms. At frames, the time-varying clique variances of AuxIVA-TVV-ICD had more diverse profiles than those of ILRMA-ICD because the time-varying clique variances of ILRMA-ICD were a weighted sum of limited (or two) bases $t_m(c, l)$.[6] In particular, direct estimation of variances in AuxIVA-TVV-ICD using (25) might provide clique variances optimized for a source model determined by shape parameter values. Therefore, the variation in the performance of AuxIVA-TVV-ICD according to the shape parameter values was less than that of ILRMA-ICD.

Figs. 6 and 7 summarize averaged SDR improvements for 1-m-apart and 5-cm-apart microphone pairs, respectively. For AuxIVA-ICD, orange bars show SDR improvements for $\beta = 0.5$ and $\gamma = 0.5$ corresponding to the source pdf model of (13) while those for ILRMA-ICD and AuxIVA-TVV-ICD represent SDR improvements for $\beta = 1$ and $\gamma = 1$ corresponding to source pdf models without inter-clique dependence. β and γ were also tuned to achieve the best overall performance and shown by red bars. Regardless of the used methods, SDR improvements using input data mixed at a reverberation time of 250 ms were less than those at

[6]More than two bases led to degradation of the overall performance.

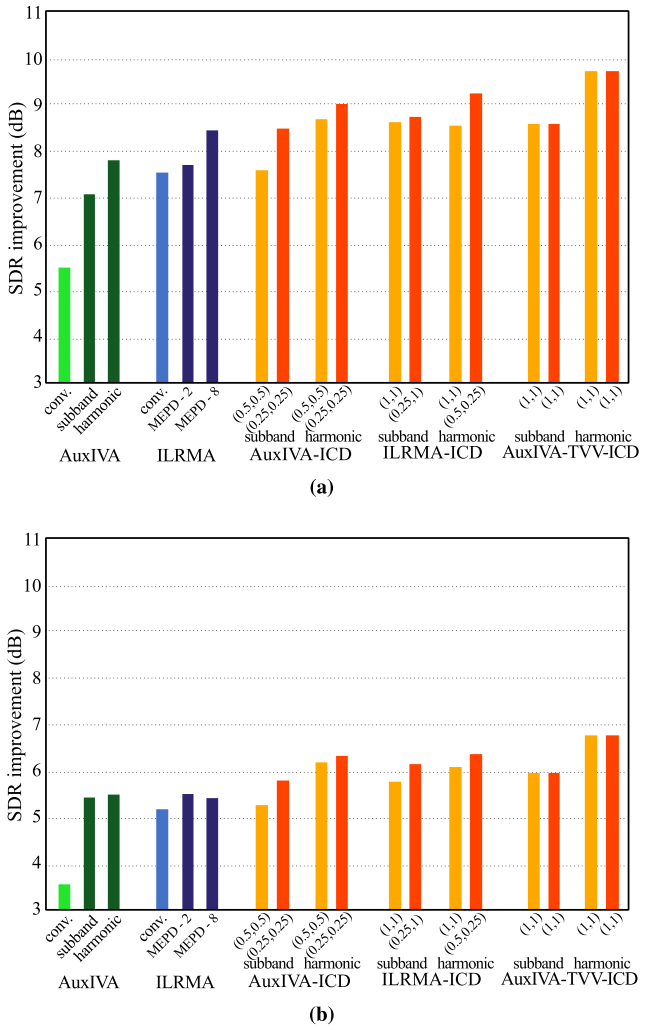

FIGURE 6. SDR improvements averaged over (a) 24 mixture sets at a configuration with two 1-m-apart microphones in Fig. 1a for a reverberation time of 130 ms and (b) 48 mixture sets at configurations with two 1-m-apart microphones in Figs. 1a and 2a for a reverberation time of 250 ms. MEPD - 2 and 8 in ILRMA denote ILRMA-MEPD with two and eight frequency range divisions, respectively. The first and second numbers in the parentheses represent used values of β and γ, respectively.

a reverberation time of 130 ms because mixing systems were more complex with larger reverberation. In addition, the SDR improvements with 5-cm-apart microphones were less than those with 1-m-apart microphones because of vulnerability to the permutation problem of BSS on observations with similar spectral properties acquired at close microphones.

Consistent with [10], [16], the subband or harmonic clique frequency dependence source models achieved better SDR improvements than a full-band model for the conventional AuxIVA. Performance degradation for the subband clique model by changing the microphone distance from 1 m to 5 cm was generally smaller than that for the harmonic clique model since the inter-clique dependence through sufficient overlaps between subband cliques might reduce the vulnerability to the permutation problem. In addition, performance improvements for the harmonic clique model compared with the

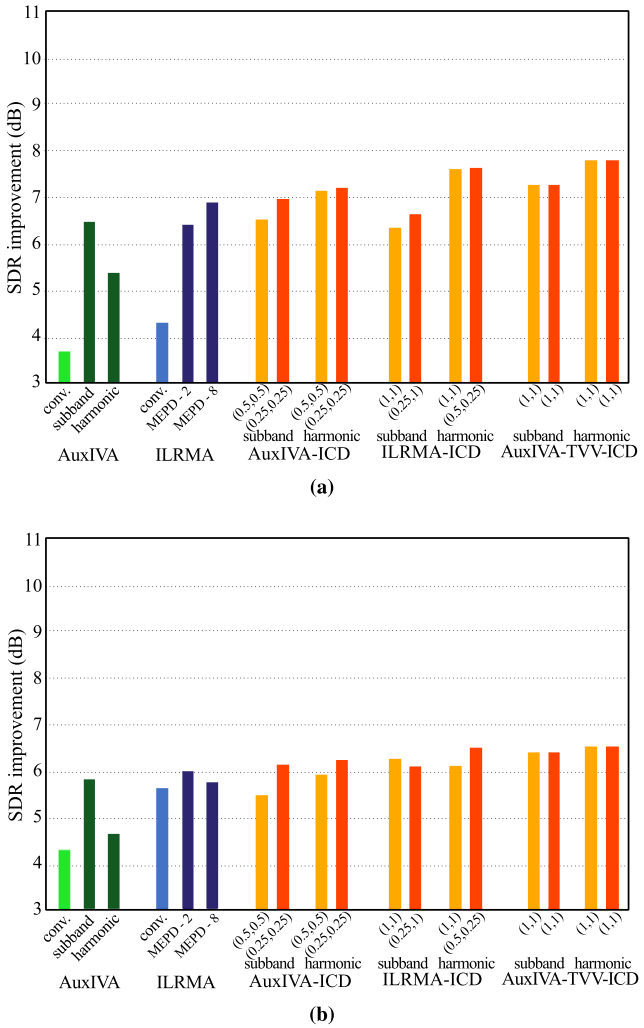

FIGURE 7. SDR improvements averaged over (a) 24 mixture sets at a configuration with two 5-cm-apart microphones in Fig. 1b for a reverberation time of 130 ms and (b) 48 mixture sets at configurations with two 5-cm-apart microphones in Figs. 1b and 2b for a reverberation time of 250 ms. MEPD - 2 and 8 in ILRMA denote ILRMA-MEPD with two and eight frequency range divisions, respectively. The first and second numbers in the parentheses represent used values of β and γ, respectively.

full-band model were higher than that for the subband clique model when the distance between microphones was 1 m. That is because the harmonic clique model may exploit the inherent harmonic structure of speech, as discussed in [9], [16], without concern for the permutation with different observations acquired at distant microphones. Particularly, AuxIVA-ICD using the generalized pdf model in (16) improved the performance regardless of the used clique models by imposing the inter-clique dependence with tuned shape parameters β and γ.

Although the results confirmed that the conventional ILRMA provided more SDR improvements than the conventional AuxIVA and ILRMA-MEPD improved the performance further, ILRMA-ICD for both the subband and harmonic clique models generally achieved better performance than ILRMA-MEPD by imposing the inter-clique dependence through overlapped cliques as well as $\beta \leq 1$. As discussed above, the time-varying clique variances of AuxIVA-TVV-ICD had more diverse profiles than those of ILRMA-ICD due to limited bases of ILRMA-ICD, and direct estimation of variances in AuxIVA-TVV-ICD using (25) might provide clique variances optimized for a source model determined by shape parameter values. Therefore, the variance estimation of AuxIVA-TVV-ICD might be more vulnerable to the permutaton problem than that based on the bases in ILRMA-ICD although AuxIVA-TVV-ICD could provide more accurate variances. This is consistent with the results that ILRMA-ICD showed comparable performance with AuxIVA-TVV-ICD with 5-cm-apart microphones due to the vulnerability to the permutation problem of BSS on similar observations acquired at close microphones. In addition, AuxIVA-TVV-ICD achieved more SDR improvements than ILRMA-ICD for the harmonic clique model with 1-m-apart microphones because of accurate estimation of more variances in the harmonic clique model than in the subband clique model without the concern for the permutation with distant microphones.

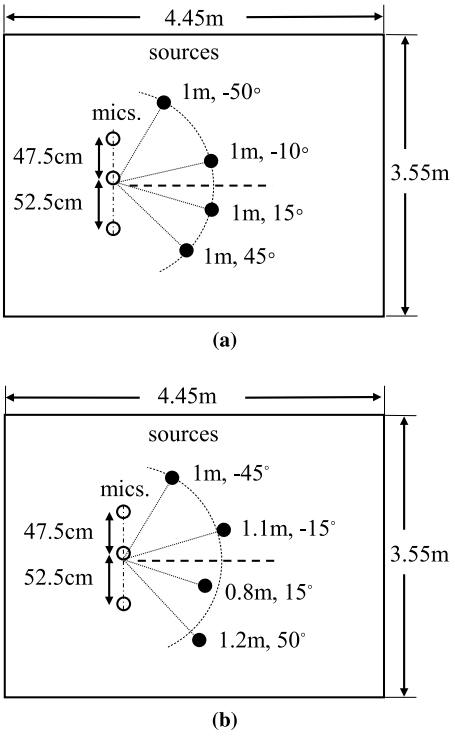

FIGURE 8. Source and microphone positions using three microphones in (a) dev1 and (b) dev2 datasets of the live-recorded speech data in SiSEC2011 [28]. The common room height is 2.5 m.

We used the same live-recorded speech data to conduct an experiment with three sources and mixtures. As shown in Fig. 8, both the 1-m-apart microphones and one of the 5-cm-apart microphones in Figs. 1 and 2 were selected to compose three microphones. To make up determined BSS problems, we selected recorded data for three source locations among four locations and summed up the data at the

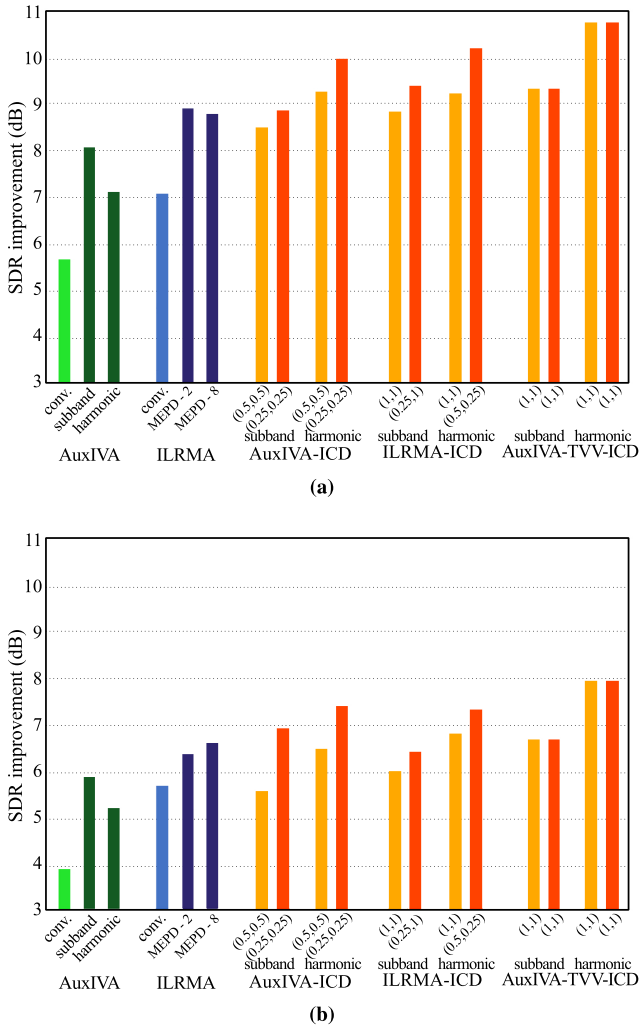

FIGURE 9. SDR improvements averaged over (a) 32 mixture sets at a configuration in Fig. 8a for a reverberation time of 130 ms and (b) 64 mixture sets at configurations in Figs. 8a and 8b for a reverberation time of 250 ms. MEPD - 2 and 8 in ILRMA denote ILRMA-MEPD with two and eight frequency range divisions, respectively. The first and second numbers in the parentheses represent used values of β and γ, respectively.

three microphones. Considering female and male source signals at each source location, 32 different mixture sets[7] were generated for a configuration at a reverberation time. As mentioned above, recordings at microphones were available with both the reverberation times of 130 ms and 250 ms for the dev1 dataset and with a reverberation time of 250 ms for the dev2 dataset.

Fig. 9 summarizes averaged SDR improvements for BSS with three sources and mixtures mentioned above.[8] Most aspects inferred from these results were similar to those of the previous ones. Although the distance between adjacent microphones was about 0.5 m, SDR improvements of the conventional AuxIVA for the harmonic clique model were less than those for the subband clique model, which meant that at least a pair of two adjacent sources were selected when choosing three of the four sources, making similar observations vulnerable to the permutation problem. However, it is noteworthy that SDR improvements of AuxIVA-ICD using the generalized pdf model were significant by imposing the inter-clique dependence with tuned shape parameters especially for the harmonic clique model. Moreover, AuxIVA-TVV-ICD for the harmonic clique frequency dependence source model achieved higher SDR improvements than the others. All the results including the previous cases consistently demonstrated that the performance of a method could be determined in general by the trade-off between the degree of freedom of source models (as long as model parameters were accurately estimated) and the vulnerability to the permutation problem.

[7]Four sets in 32 mixture sets were generated from source signals uttered by either female or male speakers.

[8]A listening demo for a mixture set is available at http://iip.sogang.ac.kr/BSS_AuxIVA.

IV. CONCLUSION

In this paper, we presented IVA using a generalized multivariate source pdf model with inter-clique dependence by imposing a variable exponent on summed cliques with overlaps and time-varying clique variances, which may include most of the conventional source models as particular cases. In addition, update rules of the clique variances and demixing matrices were derived by minimization of the cost function of BSS as well as non-negative matrix factorization (NMF) and auxiliary function techniques for fast and robust convergence, respectively. Through experiments on various mixtures, the proposed methods showed improved separation performances than the conventional methods. In particular, AuxIVA-TVV-ICD for the harmonic clique frequency dependence source model achieved higher overall SDR improvements than the others. Experimental results consistently demonstrated that the performance of a method could be determined in general by the trade-off between the degree of freedom of source models (as long as model parameters were accurately estimated) and the vulnerability to the permutation problem.

REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Kidlington, U.K.: Academic, 2010.
[2] S. Haykin, *Unsupervised Adaptive Filtering, Volume 1: Blind Source Separation*. New York, NY, USA: Wiley, 2000.
[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 31–35.
[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY, USA: Wiley, 2001.
[5] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, Aug. 2007.
[6] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 7, pp. 1431–1438, Jul. 2010.
[7] M. Oh and H.-M. Park, "Blind source separation based on independent vector analysis using feed-forward network," *Neurocomputing*, vol. 74, no. 17, pp. 3713–3715, Oct. 2011.

[8] I. Lee, G.-J. Jang, and T.-W. Lee, "Independent vector analysis using densities represented by chain-like overlapped cliques in graphical models for separation of convolutedly mixed signals," *Electron. Lett.*, vol. 45, no. 13, pp. 710–711, 2009.
[9] C. H. Choi, W. Chang, and S. Y. Lee, "Blind source separation of speech and music signals using harmonic frequency dependent independent vector analysis," *Electron. Lett.*, vol. 48, no. 2, p. 124pp. 124–125, 2012.
[10] Y. Liang, S. M. Naqvi, and J. Chambers, "Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm," *Electron. Lett.*, vol. 48, no. 8, pp. 460–462, 2012.
[11] Y. Liang, S. M. Naqvi, and J. A. Chambers, "Adaptive step size independent vector analysis for blind source separation," in *Proc. 17th Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2011, pp. 1–6.
[12] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Process.*, vol. 87, no. 8, pp. 1859–1871, Aug. 2007.
[13] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*, Oct. 2011, pp. 189–192.
[14] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. Asia Pacific Signal Inf. Process. Assoc. (APSIPA)*, Jun. 2012, pp. 1–4.
[15] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. 4th Joint Workshop Hands-Free Speech Commun. Microphone Arrays (HSCMA)*, May 2014, pp. 107–111.
[16] D. E. Bezanson, "Auxiliary function independent vector analysis using a harmonic clique dependence model," Ph.D. dissertation, Dept. Elect. Eng., UC San Diego, Jolla, CA, USA, 2013.
[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
[18] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, vol. 13, 2001, pp. 556–562.
[19] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 276–280.
[20] 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 Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.
[21] 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 *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.
[22] R. Ikeshita and Y. Kawaguchi, "Independent low-rank matrix analysis based on multivariate complex exponential power distribution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 741–745.
[23] M. Kim and H.-M. Park, "Efficient online target speech extraction using DOA-constrained independent component analysis of stereo data for robust speech recognition," *Signal Process.*, vol. 117, pp. 126–137, Dec. 2015.
[24] T.-W. Lee, *Independent Component Analysis: Theory and Applications*. Boston, MA, USA: Kluwer, 1998.
[25] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*. Cham, Switzerland: Springer, 2018, pp. 125–155.
[26] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. Int. Conf. Latent Variable Anal. Signal Separat. (LVA/ICA)*. Berlin, Germany: Springer, 2010, pp. 165–172.
[27] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. London, U.K.: Pearson, 2014.
[28] 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 *Proc. Int. Conf. Latent Variable Anal. Signal Separat. (LVA/ICA)*. Berlin, Germany: Springer, 2012, pp. 414–422.
[29] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

UI-HYEOP SHIN received the B.S. degree in electronics engineering from Sogang University, Seoul, South Korea, in 2019, where he is currently pursuing the master's degree in electronics engineering. His current research interests include speech processing and deep learning.

HYUNG-MIN PARK (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1997, 1999, and 2003, respectively. From 2003 to early 2005, he held a postdoctoral position at the Department of Biosystems, KAIST. From 2005 to early 2007, he was with the Language Technologies Institute, Carnegie Mellon University. In 2007, he joined the Department of Electronics Engineering, Sogang University, Seoul, South Korea, where he is currently a Professor. His main research interests include multichannel speech processing and robust speech recognition.