# Independent Vector Analysis for Blind Speech Separation Using Complex Generalized Gaussian Mixture Model with Weighted Variance

Xinyu Tang*†, Rilin Chen*, Xiyuan Wang‡, Yi Zhou† and Dan Su*
* Tencent AI Lab, Beijing 100193, China
E-mail: rilinchen@tencent.com
† School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
E-mail: xinyutangs@outlook.com
‡ School of Information and Communications Engineering, Beijing Information Science and Technology University, Beijing 100101, China
E-mail: wangxiyuan@bistu.edu.cn

***Abstract*—In this paper, we propose using complex generalized Gaussian mixture distribution with weighted variance for speech modelling and devise an improved independent vector analysis (IVA) algorithm for blind speech separation (BSS). Capable of capturing both non-Gaussianity and non-stationarity, the proposed complex generalized Gaussian mixture model (CGGMM) allows for a much flexible characterization of practical speech signals. The majorization minimization (MM) framework is adopted for the IVA algorithm design. Each iteration of the algorithm is comprised of the updates of demixing matrices and mixture model parameters. For demixing matrices, the update operates in a manner similar to that of the auxiliary function based IVA (AuxIVA) method, and for mixture parameters, the expectation maximization (EM) update is performed. As both updates are in closed form and pre-whitening is not a prerequisite, the IVA algorithm under CGGMM is of low complexity and can be carried out efficiently. Experimental results show that the proposed algorithm outperforms existing ones in terms of separation accuracy and also enjoys a fast convergence rate in both simulated and real environments.**

## I. Introduction

Maximizing the independence of the outputs of linear demixing systems, independent vector analysis (IVA) is an efficient blind source separation (BSS) technique for extracting acoustic sources from mixtures [1]. As the IVA algorithms do not require precise knowledge of the mixing system, their performance relies heavily on the proper modelling of acoustic sources. In order to achieve an interference-free separation, the distribution adopted in IVA should match the exact source distribution as closely as possible.

In conventional IVA methods [2–6], the spherical distribution is used as the joint distribution of source spectral coefficients. Simple as it is, the spherical distribution could be far from sufficient in modelling the variations in complex speech signals. To remedy the shortcomings of spherical distribution, various mixture models for IVA have been proposed, which include Gaussian mixture model (GMM) [7] and Student's t mixture model (SMM) [8]. With multiple density components, IVA methods based on mixture models can cater for multimodal distributions, which are common for the non-stationary speech signals. Both mixture model parameters and demixing matrices can be updated iteratively using the expectation maximization (EM) algorithm [7]. Nevertheless, for IVA algorithms based on GMM and SMM, pre-whitening is needed to stabilize the IVA iterations, and careful initialization of the EM algorithm for mixture model parameter estimation is also essential to guarantee the separation performance. Recently, Gu et al. [9] incorporated an amplitude adjusting factor into GMM to obtain an amplitude-variable GMM-based IVA algorithm (AV-GMM-IVA) whose performance is less affected by the EM initialization. In the AV-GMM-IVA, the amplitude adjusting factor is used to adapt to the temporal power fluctuation inherent to the non-stationary speech signals and then the speech source could be separated efficiently under the random initialization. Still, AV-GMM-IVA runs on signals after pre-whitening, whose error due to limited sample size could cause performance degradation. Besides, there is a disparity between the Gaussian distribution adopted by AV-GMM-IVA and the actual distribution of speech, which is in general non-Gaussian.

Inspired by AV-GMM-IVA, we propose using mixture model with variable variance in IVA algorithms. But rather than Gaussian distribution, the complex generalized Gaussian distribution (CGGD) is employed as mixture component. As a large family of bivariate symmetric distributions from super-Gaussian to sub-Gaussian distributions, the CGGD is mathematically flexible in capturing the statistical behavior of speech signals [10–12]. Therefore, the proposed speech model could capture both non-stationarity and non-Gaussianity of speech signals. Based on the majorization minimization (MM) framework, the EM algorithm is used to estimate the mixture parameters and a new cost function using the inequality from the auxiliary function based IVA (AuxIVA) [3, 13] is derived to update the demixing matrix. In this way,

the proposed algorithm does not require the pre-whitening process of observations, which reduces the computation and could be implemented online conveniently. The separation performance of the proposed algorithm is investigated and compared with the other four well-known IVA methods in the following experiments.

*Notations:* Vectors and matrices are boldface italic. $[\cdot]^{\mathrm{T}}$ and $[\cdot]^{\mathrm{H}}$ denotes non-conjugate transpose and conjugate transpose, respectively. $E[\cdot]$ is the expectation operator and $\det$ is matrix determinant.

## II. Probabilistic Model for IVA

### A. BSS in Frequency Domain

Consider an acoustic scenario where $K$ sources in an enclosure are captured by $K$ microphones. The short-time Fourier transform (STFT) representations of multiple source signals and multichannel microphone observations are denoted as $\boldsymbol{s}_{ft} = [s^1_{ft}, \ldots, s^k_{ft}, \ldots, s^K_{ft}]^{\mathrm{T}} \in \mathbb{C}^{K\times 1}$ and $\boldsymbol{x}_{ft} = [x^1_{ft}, \ldots, x^k_{ft}, \ldots, x^K_{ft}]^{\mathrm{T}} \in \mathbb{C}^{K\times 1}$ respectively where $f \in \boldsymbol{\mathcal{F}} = \{1, \ldots, F\}$ is the frequency bin index and $t \in \boldsymbol{\mathcal{T}} = \{1, \ldots, T\}$ is the frame index. The superscript $k \in \boldsymbol{\mathcal{K}} = \{1, \ldots, K\}$ denotes the source or channel index. In a noise-free system, the instantaneous mixing in the frequency domain can be expressed as [14]:

$$\boldsymbol{x}_{ft} = \boldsymbol{A}_f \boldsymbol{s}_{ft} \tag{1}$$

where $\boldsymbol{A}_f \in \mathbb{C}^{K\times K}$ is the linear mixing matrix. The original sources can be estimated by a matrix multiplication between the demixing matrix and observed mixtures. Let $\boldsymbol{y}_{ft} = [y^1_{ft}, \ldots, y^k_{ft}, \ldots, y^K_{ft}]^{\mathrm{T}} \in \mathbb{C}^{K\times 1}$ be the vector of the estimated source signals and $\boldsymbol{W}_f = [\boldsymbol{w}^1_f, \ldots, \boldsymbol{w}^k_f, \ldots, \boldsymbol{w}^K_f]^{\mathrm{H}} \in \mathbb{C}^{K\times K}$ be the demixing matrix where $\boldsymbol{w}^k_f$ is the separation filter for the $k$th source. The demixing process can be written as:

$$\boldsymbol{y}_{ft} = \boldsymbol{W}_f \boldsymbol{x}_{ft}. \tag{2}$$

### B. The Statistical Model for Source Priors

The statistical model for source priors is proposed in this section and the index $k$ is omitted for simplicity. The CGGD is adopted as a source prior at each frequency bin. Given the shape parameter $\gamma$, the variance vector $\boldsymbol{\Lambda} = [\lambda_1, \ldots, \lambda_f, \ldots, \lambda_F] \in \mathbb{R}^F$ and the frame-wise weight $\rho_t$, the joint PDF of $\boldsymbol{s}_t = [s_{1t}, \ldots, s_{ft}, \ldots, s_{Ft}] \in \mathbb{C}^F$ is given by:

$$p(\boldsymbol{s}_t \,|\, \rho_t \boldsymbol{\Lambda}, \gamma) = \prod_f \frac{1}{\pi \Gamma\left(\frac{2}{\gamma}+1\right) \rho_t \lambda_f} \exp\left(-\left|\frac{s_{ft}}{\sqrt{\rho_t \lambda_f}}\right|^{\gamma}\right) \tag{3}$$

where $\Gamma(\cdot)$ is the Gamma function. $\rho_t$ is the time-varying weight factor of variances over all frequency bins, which partially preserves the dependencies over frequency components and allows each frame to be treated differently. It acts as the temporal power compensation between the estimated $\lambda_f$ and the output signal $y_{ft}$ and was first proposed in [9]. So the non-stationarity of speech signals caused by the temporal power fluctuation can be captured. For another, the non-Gaussian statistical properties of speech signals are also considered. The variable shape parameter $\gamma$ determines the decay rate of the density function, whose smaller value corresponds to heavier-tailed distribution and vice versa. The CGGD is mathematically flexible in capturing the statistical behavior of speech signals [10], from super-Gaussian ($\gamma < 2$) to sub-Gaussian ($\gamma > 2$) including specific densities such as Gaussian ($\gamma = 2$) and Laplacian ($\gamma = 1$) distributions [12].

However, using a single PDF in (3) as source priors can not adapt to the statistical properties of different sources. Therefore, as the IVA methods using mixture models [7–9], we further derive the complex generalized Gaussian mixture model (CGGMM) with $I$ components as follows:

$$p(\boldsymbol{s}_t \,|\, \boldsymbol{\Omega}) = \sum_i \pi_i p\left(\boldsymbol{s}_t \,|\, \rho_{t,i} \boldsymbol{\Lambda}_i, \gamma_i\right) \tag{4}$$

where $\boldsymbol{\Omega} = \{\pi_1, \ldots, \pi_I, \rho_{t,1}\boldsymbol{\Lambda}_1, \ldots, \rho_{t,I}\boldsymbol{\Lambda}_I, \gamma_1, \ldots, \gamma_I\}$ and the subscript $i \in \boldsymbol{\mathcal{I}} = \{1, \ldots, I\}$ indicates the $i$th mixture component. $\pi_1, \ldots, \pi_I$ are the mixture coefficients satisfying $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. The proposed statistical model in (4) generalizes the GMM in the IVA method and is able to capture both non-Gaussianity and non-stationarity of speech signals.

### C. Objective Function of IVA

For the purpose of maximizing the statistical independence of sources and avoiding the permutation problem, IVA measures the independence from the entire spectrogram of each source signal. Let $\boldsymbol{Y}^k = \{\boldsymbol{y}^k_1, \ldots, \boldsymbol{y}^k_t, \ldots, \boldsymbol{y}^k_T\}$ be the estimated $k$th source data where $\boldsymbol{y}^k_t = [y^k_{1t}, \ldots, y^k_{ft}, \ldots, y^k_{Ft}]$. $\boldsymbol{P}^k_i = \{\rho^k_{1,i}, \ldots, \rho^k_{t,i}, \ldots, \rho^k_{T,i}\}$ are the weights associated with the $i$th mixture component of $\boldsymbol{Y}^k$. Using the Kullback-Leibler (KL) divergence between $p(\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^K)$ and $\prod_k p(\boldsymbol{Y}^k)$, the separated process can be realized by minimizing the objective function [15]:

$$\mathcal{J}(\boldsymbol{W}, \boldsymbol{\Theta}) = \sum_k E\left[G\left(\boldsymbol{Y}^k\right)\right] - \sum_f \log\left|\det \boldsymbol{W}_f\right| \tag{5}$$

where $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ represent the sets of demixing matrices and mixture model parameters respectively, i.e. $\boldsymbol{W} = \{\boldsymbol{W}_f\}_{f=1}^F$, $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}^k\}_{k=1}^K$ where $\boldsymbol{\Theta}^k = \{\pi^k_1, \ldots, \pi^k_I, \boldsymbol{P}^k_1, \ldots, \boldsymbol{P}^k_I, \boldsymbol{\Lambda}^k_1, \ldots, \boldsymbol{\Lambda}^k_I\}$. Note that shape parameters $\{\gamma^k_1, \ldots, \gamma^k_I\}_{k=1}^K$ would be set as priors. $G\left(\boldsymbol{Y}^k\right)$ is the contrast function with a relationship of $G\left(\boldsymbol{Y}^k\right) = -\log p\left(\boldsymbol{Y}^k\right)$. In this research, we consider the statistical model in (4) for the contrast function and (5) can then be written as:

$$\begin{aligned}\mathcal{J}(\boldsymbol{W}, \boldsymbol{\Theta}) = &-\frac{1}{T}\sum_{k,t} \log\left(\sum_i \pi^k_i p(\boldsymbol{y}^k_t \,|\, \rho^k_{t,i} \boldsymbol{\Lambda}^k_i, \gamma^k_i)\right) \\ &- \sum_f \log\left|\det \boldsymbol{W}_f\right|.\end{aligned} \tag{6}$$

## III. Optimization Based on Auxiliary Function

As minimizing (6) is a nonlinear optimization problem, the majorization minimization (MM) framework is adopted to find the closed form solutions of mixture model parameters $\boldsymbol{\Theta}$ and demixing matrices $\boldsymbol{W}$. The MM algorithm is to find an auxiliary function $Q$, which satisfies $Q\Big(\theta,\hat{\theta}\Big) \geq \mathcal{J}(\theta)$ where the equality sign is satisfied if and only if $\theta = \hat{\theta}$. The alternative updates in terms of $\theta$ and $\hat{\theta}$ guarantee $\mathcal{J}(\theta)$ monotonically decreases to a stationary value [16].

The mixture parameters are estimated via the EM algorithm and the upper bound of $\mathcal{J}(\boldsymbol{W},\boldsymbol{\Theta})$ can first be obtained by Jensen's inequality [17]:

$$Q(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{q}) = -\frac{1}{T}\sum_{k,t}\Bigg(\sum_i q_{t,i}^k \log\big(\pi_i^k p\big(\boldsymbol{y}_t^k \mid \rho_{t,i}^k \boldsymbol{\Lambda}_i^k, \gamma_i^k\big)\big) - \sum_i q_{t,i}^k \log q_{t,i}^k\Bigg) - \sum_f \log|\det \boldsymbol{W}_f|. \quad (7)$$

$q_{t,i}^k$ is the posterior probability of the $i$th mixture component at the $t$th frame for the $k$th source, given observations and the estimated parameters from the last iteration. The calculation of $q_i^k$ is the expectation-step (E-step) and can be derived as:

$$q_{t,i}^k = \frac{\pi_i^k p\left(\boldsymbol{y}_t^k \mid \rho_{t,i}^k \boldsymbol{\Lambda}_i^k, \gamma_i^k\right)}{\sum\limits_{j\in\boldsymbol{\mathcal{I}}} \pi_j^k p\left(\boldsymbol{y}_t^k \mid \rho_{t,j}^k \boldsymbol{\Lambda}_j^k, \gamma_j^k\right)}. \quad (8)$$

The expansion of (7) is

$$Q(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{q}) = \frac{1}{T}\sum_{k,t}\Bigg(-\sum_i q_{t,i}^k \log\pi_i^k + \sum_i q_{t,i}^k \left(F\log\Gamma\left(\frac{2}{\gamma_i^k}+1\right) + F\log\rho_{t,i}^k + \sum_f \log\lambda_{f,i}^k + \sum_f \left|\frac{y_{ft}^k}{\sqrt{\rho_{t,i}^k\lambda_{f,i}^k}}\right|^{\gamma_i}\right) + \sum_i q_{t,i}^k \log q_{t,i}^k\Bigg) + KF\log\pi - \sum_f \log|\det \boldsymbol{W}_f|. \quad (9)$$

Then, in the maximization step (M-step), $\boldsymbol{\Theta}$ is updated by setting the derivatives of $Q(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{q})$ to zero. With some straightforward mathematical manipulations, the following formulas are obtained for the mixture coefficient, the variance and the weight, respectively.

$$\pi_i^k = \frac{1}{T}\sum_t q_{t,i}^k, \quad (10)$$

$$\lambda_{f,i}^k = \left(\frac{\gamma_i^k \sum\limits_t q_{t,i}^k \left|\frac{y_{ft}^k}{\sqrt{\rho_{t,i}^k}}\right|^{\gamma_i^k}}{2\sum\limits_t q_{t,i}^k}\right)^{2/\gamma_i^k}, \quad (11)$$

and

$$\rho_{t,i}^k = \left(\frac{\gamma_i^k \sum\limits_f \left|\frac{y_{ft}^k}{\sqrt{\lambda_{f,i}^k}}\right|^{\gamma_i^k}}{2F}\right)^{2/\gamma_i^k}. \quad (12)$$

To further find the closed form solution of $\boldsymbol{W}_f$, an inequality is derived from the theorem proven in original AuxIVA [3, 13] and can be stated as:

$$|y|^\gamma \leq |\hat{y}|^\gamma + \frac{\gamma}{2}|\hat{y}|^{\gamma-2}\left(|y|^2 - |\hat{y}|^2\right) \quad (13)$$

where the equality sign is satisfied if and only if $|y| = |\hat{y}|$. Thus, (9) can be modified to a new upper bound in terms of $\boldsymbol{W}$ and $\boldsymbol{V}$ by applying (13):

$$Q_W(\boldsymbol{W},\boldsymbol{V}) = \sum_f\left(\frac{1}{2}\sum_k \boldsymbol{w}_f^{k\mathrm{H}} \boldsymbol{V}_f^k \boldsymbol{w}_f^k - \log|\det \boldsymbol{W}_f|\right) + R_\Theta, \quad (14)$$

$$\boldsymbol{V}_f^k = \frac{1}{T}\sum_{t,i} q_{t,i}^k \gamma_i^k \left(\rho_{t,i}^k \lambda_{f,i}^k\right)^{-\gamma_i^k/2} \left|\hat{y}_{ft}^k\right|^{\gamma_i^k-2} \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^{\mathrm{H}}, \quad (15)$$

where $\hat{y}_{ft}^k$ is the estimated source signal in the last iteration. $R_\Theta$ contains the constant and the terms with parameters of the mixture model but independent of $\boldsymbol{W}$. The auxiliary variable $\boldsymbol{V}$ represents a series of $\boldsymbol{V}_f$ for any $f$ where $\boldsymbol{V}_f = \{\boldsymbol{V}_f^k\}_{k=1}^K$. Resemblance to the original AuxIVA, the demixing matrix updates for any $f$ and $k$ can then be expressed as:

$$\boldsymbol{w}_f^k = \left(\boldsymbol{W}_f \boldsymbol{V}_f^k\right)^{-1} \boldsymbol{e}_k, \quad (16)$$

$$\boldsymbol{w}_f^k = \boldsymbol{w}_f^k \Big/ \sqrt{\boldsymbol{w}_f^{k\mathrm{H}} \boldsymbol{V}_f^k \boldsymbol{w}_f^k}, \quad (17)$$

where $\boldsymbol{e}_k$ is the unit vector which has a single non-zero element 1 in the $k$th position.

Based on the above auxiliary function approach, $Q$ is iteratively minimized over $\boldsymbol{q},\boldsymbol{\Theta}$ and $\boldsymbol{W}$ and still obtains a monotonic decrease until the convergence. In each iteration, the mixture parameters are updated based on (8) and (10)$\sim$(12) with a complexity of $\mathcal{O}(KIFT)$, while the demixing matrices are estimated via (15)$\sim$(17) with a complexity of $\mathcal{O}\big(K^4F\big)$.

## IV. Experiments

In this section, the performance of the proposed algorithm (**CGGMM-IVA**) is evaluated and compared with the following four well-known IVA algorithms: (1) **CL-AuxIVA**, original AuxIVA with a time-invariant circular Laplace distribution [3]; (2) **SCGG-AuxIVA**, AuxIVA with a spherical complex-valued generalized Gaussian distribution [18]; (3) Independent low-rank matrix analysis (**ILRMA**), estimating a spatial model using IVA and a source model by low-rank decomposition using the nonnegative matrix factorization (NMF) [19]; (4) **AV-GMM-IVA**, a recently proposed algorithm with an amplitude variable Gaussian mixture model using IVA [9].

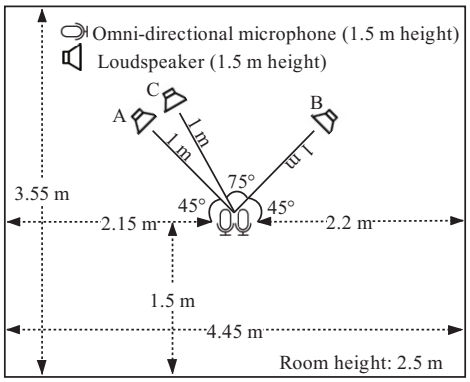


Fig. 1. The layout of simulated experimental setup.

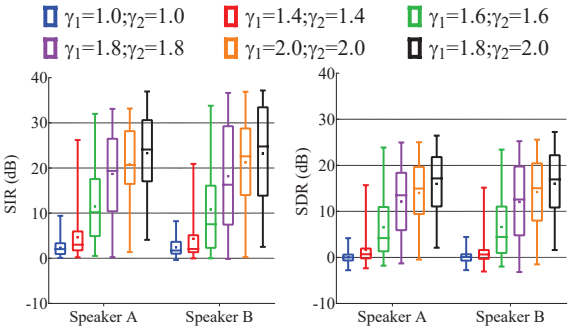


Fig. 2. Box-plots of SIR (left) and SDR (right) improvements for different shape parameters over 50 mixtures in Exp. 1. The dot shows the mean.

The number of mixture states for both CGGMM-IVA and AV-GMM-IVA is set to 2, while the number of bases for ILRMA is set to 10. The shape parameter in SCGG-AuxIVA is set to 0.4 for each source . For CGGMM-IVA, $\lambda_{f,i}^{k}$ is initialized to 1 and $\{\gamma_i^k\}_{i=1}^{I}$ is the same for any $k$ but may set to the different value for different state $i$. All the algorithms run until the decrement of the cost function between adjacent iterations is less than or equal to $10^{-6}$. Moreover, the data pre-whitening is implemented merely in ILRMA and AV-GMM-IVA. The minimal distortion principle [20] is utilized in the post-processing for all the algorithms. A 4096-point FFT, 4096-tab Hanning window with half-overlap are used in STFT domain. The results are evaluated by the signal-to-interference ratio (SIR) and signal-to-distortion ratio (SDR) in decibels using the BSS_EVAL toolbox [21]. Some audio samples are available online at https://github.com/shelly-tang/CGGMM-IVA.

### *A. Separation Results in the Simulated Environment*

Live-recorded speech segments from SiSEC2018 database [22] have been used as sources, which are ten-second-long and sampled at 16 kHz. All source signals are convoluted with room impulse responses (RIRs) obtained by the image method [23] and totally 100 mixed speech signals are simulated where

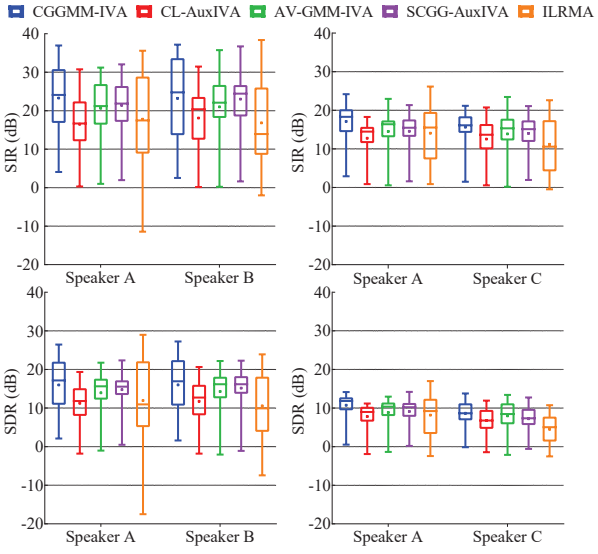


Fig. 3. Box-plots of SIR (top) and SDR (bottom) improvements for five IVA models, with 100 sample data for each algorithm. The left column is for Exp. 1 while the right is for Exp. 2. The dot shows the mean.

the target sources have similar energy levels. Fig. 1 depicts the experimental environment. Reverberation time is set to 130 ms in a room of size $4.45 \times 3.55 \times 2.5$ m. Two omnidirectional microphones are configured with 10 cm spacing while the distance between sources and microphones is 1 m. For the 2×2 case, the experiments are conducted using two different source location settings. **Exp. 1**: The first simulation mixes speaker A (from location 'A', 45°) and speaker B (from location 'B', 135°). **Exp. 2**: The second simulation mixes speaker A and speaker C (from location 'C', 60°).

As the shape parameter could affect the performance of the proposed algorithm, we first test algorithm under various values of $\gamma$. Fig. 2 shows the box-plots of SIR and SDR improvements for different shape parameters in Exp. 1. $\gamma_1$ and $\gamma_2$ are the shape parameters of two mixture components for each source, i.e. $\gamma_1^1 = \gamma_1^2 = \gamma_1$, $\gamma_2^1 = \gamma_2^2 = \gamma_2$. In the case of $\gamma_1 = \gamma_2$, the best separation performance is generally achieved when $\gamma_1 = \gamma_2 = 2$. It also can be observed that $\gamma_1 = \gamma_2 = 1.8$ or 1.6 performs better than 2 in some trials. Additionally, we evaluate the improvements when using different $\gamma_1$ and $\gamma_2$, i.e., $\gamma_1 = 1.8$ and $\gamma_2 = 2$. It shows significant superiority compared with the cases of $\gamma_1 = \gamma_2$. These results reveal that the PDF using different shape parameters for different mixture components matches better with speech signals and $\gamma = 1.6 \sim 2$ would be a good choice.

Fig. 3 shows the comparison of separation performance between the proposed algorithm $(\gamma_1 = 1.8, \gamma_2 = 2)$ and the other four algorithms. The improvements in Exp. 1 and Exp. 2 are presented in the left and right columns, respectively. It can be found that CGGMM-IVA performs the best of all the mentioned algorithms. The performance of the AV-GMM-IVA and

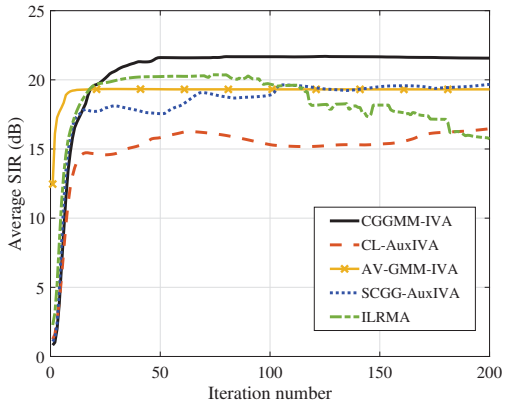


Fig. 4. Average SIR convergence over 30 trials in Exp. 1.

SCGG-AuxIVA is slightly inferior to that of CGGMM-IVA. ILRMA shows the excellent performance in music signals separation [19] but is relatively ineffective in separating some speech mixtures in this research. As sources get closer, both SIR and SDR metrics decrease for all algorithms mentioned, but the proposed CGGMM-IVA still retain the competitive speech separation performance.

Fig. 4 demonstrates the average SIR convergence over 30 trials mixing speaker A and speaker B. The proposed algorithm requires approximately 50 iterations to convergence, more than AV-GMM-IVA but distinctly fewer than CL-AuxIVA and SCGG-AuxIVA whose iteration numbers are about 115 and 153 respectively. ILRMA has achieved convergence after 56 iterations but the SIR improvement exhibits fluctuations when the algorithm iterates more than 100 times. Both CGGMM-IVA and AV-GMM-IVA are more stable than other three algorithms.

### *B. Separation Results in the Real Environment*

The proposed CGGMM-IVA algorithm has shown competitive speech separation performance in the simulated environments. In this section, we record speech utterances of speakers in different directions separately by dual microphones in a meeting room and then obtain the mixtures of 2 sources by summing the recorded source signals. The experimental conditions in the real environment are summarized in Table I. The real recording of source is used as ground-truth signal.

Table II shows separation performances in averaged SIR and SDR values over 24 trials (6 trials for each combination of source direction). The proposed CGGMM-IVA algorithm is first evaluated under different values of shape parameters ranging from 1.6 to 2 and the best improvement for each combination of source directions has been listed. Different from the experiments in simulated environment where the setting of $\gamma_1 = 1.8$ and $\gamma_2 = 2$ always attains the best separation performance, it achieves better improvements to set $\gamma_1 = \gamma_2 = 1.6$ for the case of $(45°, 60°)$ and $\gamma_1 = \gamma_2 = 2$ for the case of $(45°, 90°)$ in this experiment. Overall, though similar

TABLE I
EXPERIMENTAL CONDITIONS IN THE REAL ENVIRONMENT

| | |
|---|---|
| Room size | $10.78 \times 7.58 \times 3$ m |
| Reverberation time | 300 ms |
| Microphone spacing | 10 cm |
| Source-microphone distance | 1m |
| Direction of source 1 | $45°$ |
| Direction of source 2 | $60°, 90°, 135°$ |
| Signal length | 65 s |
| Sampling frequency | 16 kHz |
| Frame length | 4096 |
| Frame shift | 2048 |
| Window function | Hanning |
| Iteration number | 100 times |

to AV-GMM-IVA, the proposed CGGMM-IVA shows the superior separation performance compared with CL-AuxIVA, SCGG-AuxIVA and ILRMA. Moreover, the SIR values of all algorithms are relatively low in the case of $(45°, 60°)$, and even negative SDR values are observed. It is consistent with the conclusion from simulated results in Fig. 3 that separation performance could deteriorate when the positions of sources are close.

Besides, the permutation problem common to IVA algorithms has been observed in this experiment. Fig. 5 shows the spectrograms of source signals and processed signals of five IVA algorithms in one trial under the combination of $(45°, 135°)$. In this case, although frequency components under around 2 kHz can be well separated by all IVA methods, the obvious permutation problem in high frequency domain still exists for CL-AuxIVA, AV-GMM-IVA, SCGG-AuxIVA and ILRMA. In contrast, the permutation problem for the proposed CGGMM-IVA is much alleviated when setting suitable shape parameter ($\gamma_1 = 1.8$ and $\gamma_2 = 2$ in this case) to match the source model.

## V. CONCLUSIONS

This paper introduced a complex generalized Gaussian mixture model with weighted variance as source priors for the IVA method to increase flexibility in modelling various statistical properties of non-stationary speech signals. The auxiliary function approach based on the MM framework was effective to realize optimization and did not require the data pre-whitening. The experimental results in both simulated and real environments revealed that the proposed algorithm attained best performance when the shape parameter was within the range of 1.6 and 2 and the flexibility in modelling various statistical properties made the proposed algorithm outperform conventional IVA ones by setting the suitable shape parameters to match the source model.

## REFERENCES

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007, vol. 615.

[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, 2006.

TABLE II
SEPARATION PERFORMANCES IN AVERAGED SIR AND SDR (DB)

| Algorithm | SIR | | | SDR | | |
|---|---|---|---|---|---|---|
| | $(45°, 60°)$ | $(45°, 90°)$ | $(45°, 135°)$ | $(45°, 60°)$ | $(45°, 90°)$ | $(45°, 135°)$ |
| proposed CGGMM-IVA $\gamma_1 = 1.6, \gamma_2 = 1.6$ | **4.0479** | 8.4987 | 9.9250 | -2.6435 | 3.7807 | 4.1499 |
| proposed CGGMM-IVA $\gamma_1 = 2.0, \gamma_2 = 2.0$ | 2.1174 | **13.1070** | 11.7231 | -1.0149 | **6.4742** | 5.4778 |
| proposed CGGMM-IVA $\gamma_1 = 1.8, \gamma_2 = 2.0$ | 2.1386 | 12.9601 | **15.5446** | -1.9822 | 6.3398 | **7.3882** |
| CL-AuxIVA | 2.6423 | 10.1911 | 10.7625 | -0.7807 | 5.1380 | 4.9414 |
| AV-GMM-IVA | 1.2458 | **13.1628** | 15.0681 | -3.9355 | 6.4165 | **7.3982** |
| SCGG-AuxIVA | 2.3769 | 12.5489 | 11.0812 | -0.8535 | 6.2756 | 5.1401 |
| ILRMA | 1.8758 | 9.5692 | 11.4433 | -2.9594 | 5.0017 | 5.7498 |

$(x°, y°)$ The combination of source directions. Directions of source 1 and source 2 are $x°$ and $y°$, respectively.

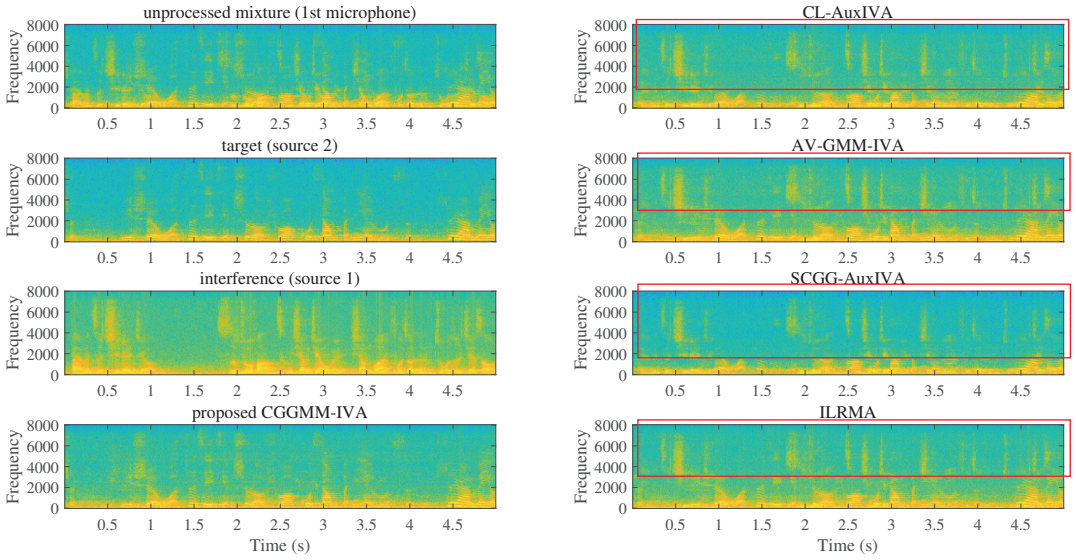


Fig. 5. Observed spectrograms of source signal and processed signals of five IVA algorithm in one trial mixing sources from $45°$ and $135°$. They are the separated results for target source 2. Since both 'target' and 'interference' are real recordings of sources, the background noise caused by interior circuit is unavoidable, which slightly pollutes spectrograms.

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

[4] Y. Liang, J. Harris, G. Chen, S. M. Naqvi, C. Jutten, and J. Chambers, "Auxiliary function based iva using a source prior exploiting fourth order relationships," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.

[5] 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 *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE, 2014, pp. 107–111.

[6] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Processing*, vol. 87, no. 8, pp. 1859–1871, 2007.

[7] J. Hao, I. Lee, T.-W. Lee, and T. J. Sejnowski, "Independent vector analysis for source separation using a mixture of gaussians prior," *Neural computation*, vol. 22, no. 6, pp. 1646–1673, 2010.

[8] W. Rafique, J. Chambers, and A. I. Sunny, "An expectation–maximization-based iva algorithm for speech source separation using student's t mixture model based source priors," in *Acoustics*, vol. 1, no. 1. Multidisciplinary Digital Publishing Institute, 2019, pp. 117–136.

[9] Z. Gu, J. Lu, and K. Chen, "Speech separation using independent vector analysis with an amplitude variable gaussian mixture model," *Proc. Interspeech 2019*, pp. 1358–1362, 2019.

[10] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[11] K. Sharifi and A. Leon-Garcia, "Estimation of shape pa-

rameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, 1995.

[12] M. Novey, T. Adali, and A. Roy, "A complex generalized gaussian distribution—characterization, generation, and estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1427–1433, 2009.

[13] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-gaussian sources," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 165–172.

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

[15] 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.

[16] D. R. Hunter and K. Lange, "A tutorial on mm algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[17] F. Dellaert, "The expectation maximization algorithm," Georgia Institute of Technology, Tech. Rep., 2002.

[18] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.

[19] 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.

[20] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proceedings of the 41st SICE Annual Conference. SICE 2002.*, vol. 4. IEEE, 2002, pp. 2138–2143.

[21] 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.

[22] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 293–305.

[23] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.