MADRID
inter.noise 2019
June 16 - 19
NOISE CONTROL FOR A BETTER ENVIRONMENT

Generalized-Gaussian-Distribution-Based Independent Deeply Learned Matrix Analysis for Multichannel Audio Source Separation

Makishima, Naoki[1]
The Graduate School of Information Science and Technology,
The University of Tokyo, Tokyo 113-8656, Japan

Takamune, Norihiro[2]
The Graduate School of Information Science and Technology,
The University of Tokyo, Tokyo 113-8656, Japan

Kitamura, Daichi[3]
Department of Electrical and Computer Engineering,
National Institute of Technology, Kagawa College, Kagawa 761-8058, Japan

Saruwatari, Hiroshi[4]
The Graduate School of Information Science and Technology,
The University of Tokyo, Tokyo 113-8656, Japan

Takahashi, Yu[5]
Yamaha Corporation, Shizuoka 430-8650, Japan

Kondo, Kazunobu[6]
Yamaha Corporation, Shizuoka 430-8650, Japan

Nakajima, Hiroaki[7]
Yamaha Corporation, Shizuoka 430-8650, Japan

ABSTRACT

In this paper, generalization of a statistical generative model in independent deeply learned matrix analysis (IDLMA) is addressed to achieve higher audio

[1] naoki_makishima@ipc.i.u-tokyo.ac.jp
[2] norihiro_takamune@ipc.i.u-tokyo.ac.jp
[3] kitamura-d@t.kagawa-nct.ac.jp
[4] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp
[5] yu.takahashi@music.yamaha.com
[6] kazunobu.kondo@music.yamaha.com
[7] hiroaki.nakajima@music.yamaha.com

source separation performance. Audio source separation is the task of extracting source signals from multichannel mixtures observed using a microphone array, which can be applied to many systems including noise reduction, speech recognition, and music analysis. IDLMA is a state-of-the-art separation method exploiting statistical independence between sources and deep neural network (DNN) inference of source models, where a time-frequency-varying complex Gaussian distribution is assumed as a source generative model. This paper presents the generalization of the source generative model in IDLMA: a time-frequency-varying complex generalized Gaussian distribution (GGD) is exploited as a new source generative model in IDLMA. From theoretical and experimental results, both numerical stability in parameter estimation and improved separation performance are confirmed.

Keywords: Audio source separation, Deep neural networks
I-INCE Classification of Subject Number: 74

1. INTRODUCTION

Audio source separation aims to recover original source signals from an observed multichannel mixture. This technique can be applied to almost all audio systems, e.g., noise reduction, speech recognition, and music analysis, as a front-end system. In particular, blind source separation (BSS) estimates the sources without any prior knowledge such as locations of sources and microphones. The most commonly used algorithms for BSS in the (over)determined case (number of microphones $\geq$ number of sources) are independent component analysis (ICA) [1] and its extended algorithms such as independent vector analysis (IVA) [2], which assume statistical independence between the sources and estimate the demixing system. Recently, independent low-rank matrix analysis (ILRMA) [3, 4], which is a unification of IVA and nonnegative matrix analysis (NMF) [5], was proposed to achieve high separation accuracy. The original ILRMA proposed in [3] employs a time-frequency-varying complex Gaussian distribution as a source generative model. Then, this model was generalized to a time-frequency-varying complex generalized Gaussian distribution (GGD) [6, 7], where the GGD can represent a more heavy-tailed (super-Gaussian) distribution. It is reported that the heavy-tailed source generative model improves the separation performance in ILRMA.

In the underdetermined case (number of microphones $<$ number of sources), on the other hand, algorithms that estimate the mixing system have been proposed, and many state-of-the-art algorithms are based on the Duong model proposed in [8]. In the Duong model, the spatial covariance matrix, which decodes the sourcewise spatial information (relative locations of the source and microphones and their spatial spread), is estimated by the expectation-maximization (EM) algorithm. Multichannel NMF (MNMF) [9, 10] is a technique for underdetermined BSS combining the Duong model [8] and NMF-based source modeling. The models assumed in MNMF and ILRMA are equivalent in the determined case if the rank of the spatial covariance matrix is restricted to one. However, it has been experimentally confirmed that the estimation of a demixing matrix in ILRMA is more stable than the estimation of a mixing system (covariance matrix) in MNMF, resulting in higher separation accuracy in ILRMA [3].

In supervised (informed) source separation, deep neural networks (DNNs) have shown promising performance [11, 12]. Nugraha et al. proposed a unified approach of a Duong-model-based spatial model and DNN-based source models [11] (hereafter referred to as Duong+DNN). Since Duong+DNN estimates the mixing system (spatial covariance matrix) for the separation, this method can even be applied to the underdetermined case. A unification of the independence-based demixing model and DNN source model was also proposed as independent deeply learned matrix analysis (IDLMA) [12], and it was reported that IDLMA achieves higher separation performance than Duong+DNN in the overdetermined case, just as ILRMA outperforms MNMF.

In this paper, we generalize the source generative model assumed in IDLMA from the time-frequency-varying complex Gaussian distribution to the time-frequency-varying complex GGD, which is called GGD-IDLMA. We also reveal the relationship between the numerical stability of the estimation and the separation performance, showing the efficacy of GGD-IDLMA.

2. CONVENTIONAL METHOD

2.1. Formulation

Let M and N be the numbers of microphones and sound sources, respectively. We assume the determined case where $M = N$. The short-time Fourier transform (STFT) of the observed mixtures, estimated signals, and source signals are defined as

$$\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijN})^{\mathrm{T}}, \tag{1}$$

$$\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijN})^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{s}_{ij} = (s_{ij1}, \ldots, s_{ijN})^{\mathrm{T}}, \tag{3}$$

where $^{\mathrm{T}}$ denotes the transpose and $i = 1, \ldots, I$ and $j = 1, \ldots, J$ denote the indexes of frequency bins and time frames, respectively. In the determined case, the estimated signals $\boldsymbol{y}_{ij}$ can be represented as $\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}$, where $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iN})^{\mathrm{H}} \in \mathbb{C}^{N \times N}$ is the demixing matrix, $\boldsymbol{w}_{in}^{\mathrm{H}}$ is the demixing filter for the nth source, and $^{\mathrm{H}}$ denotes the Hermitian transpose.

2.2. ILRMA and its extension to complex GGD

In ILRMA [3], the following time-frequency-varying complex Gaussian distribution is assumed as a source model:

$$\prod_{i,j} p(y_{ijn}) = \prod_{i,j} \frac{1}{\pi {r_{ijn}}^2} \exp\left(-\frac{|y_{ijn}|^2}{{r_{ijn}}^2} \right), \tag{4}$$

$${r_{ijn}}^2 = \sum_k t_{ikn} v_{kjn}, \tag{5}$$

where r_{ijn} is the scale parameter of the Gaussian distribution, $t_{ikn} > 0$ and $v_{kjn} > 0$ are the NMF bases and activation parameters, respectively, and $k = 1, \ldots, K$ is the index of the NMF bases. We denote the scale parameter matrix as $\boldsymbol{R}_n \in \mathbb{R}_{\geq 0}^{I \times J}$, whose elements are r_{ijn}.

The source model in Equation 4 is generalized to the following time-frequency-varying complex GGD [6]:

$$\prod_{i,j} p(y_{ijn}) = \prod_{i,j} \frac{\beta^{1-\frac{2}{\beta}}}{2^{1-\frac{2}{\beta}} \pi {r_{ijn}}^2 \Gamma\left(\frac{2}{\beta}\right)} \exp\left(-\frac{2}{\beta} \frac{|y_{ijn}|^\beta}{{r_{ijn}}^\beta} \right), \tag{6}$$

$${r_{ijn}}^p = \sum_k t_{ikn} v_{kjn}, \tag{7}$$

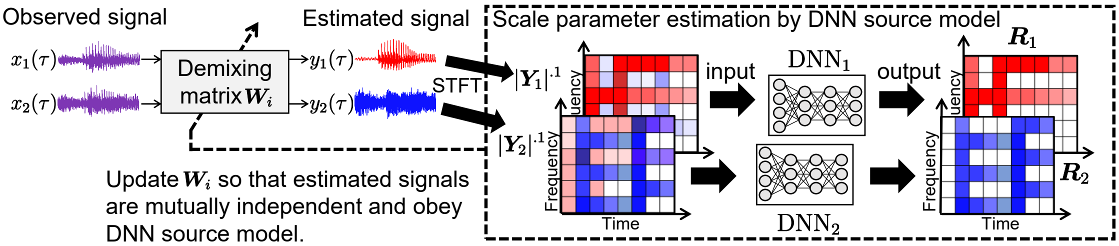

Figure 1: Principle of IDLMA.

where β is the shape parameter and p is the parameter that defines the domain of NMF decomposition. When $\beta = 2$ and $p = 2$, Equations 6 and 7 become identical to Equations 4 and 5, respectively. The demixing matrix $\boldsymbol{W}_i$ can be obtained by maximizing the likelihood of the observed signals [3, 6].

2.3. IDLMA

IDLMA is a technique of multichannel audio source separation that estimates the source model $\boldsymbol{R}_n$ using DNNs. An overview of IDLMA is shown in Fig. 1. In [12], similar to the original ILRMA [3], the time-frequency-varying complex Gaussian distribution is assumed as a source model (hereafter referred to as Gauss-IDLMA). On the basis of Equation 4, the cost function (negative log-likelihood of the observed signal $\boldsymbol{x}_{ij}$) of Gauss-IDLMA is obtained as

$$\mathcal{L} = \sum_{i,j,n} \left[\frac{|\boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{x}_{ij}|^2}{{r_{ijn}}^2} + \log {r_{ijn}}^2 \right] - 2J \sum_i \log |\det \boldsymbol{W}_i| + \mathrm{const.} \tag{8}$$

While ILRMA expresses the scale parameter r_{ijn} with NMF, Gauss-IDLMA estimates r_{ijn} using DNNs. Since an NMF-based low-rank source model is not always valid, ILRMA sometimes fails to separate the sources, especially in the case of a speech-speech mixture. In IDLMA, the appropriate source models for each source are trained in advance using solo-recorded source signals, and the trained DNNs are used as the estimator of $\boldsymbol{R}_n$.

Gauss-IDLMA iteratively updates the demixing matrix $\boldsymbol{W}_i$ and the source model $\boldsymbol{R}_n$. Equation 8 consists of a negative log-determinant term of $\boldsymbol{W}_i$ and a quadratic form of $\boldsymbol{w}_{in}$. The minimization of Equation 8 leads to the solution that maximizes the independence between sources taking the source model (the DNN estimates) $\boldsymbol{R}_n$ into account. Similar to ILRMA, $\boldsymbol{W}_i$ is updated by iterative projection (IP) [13], which is a convergence-guaranteed fast algorithm. By applying IP, we can derive the update rule of $\boldsymbol{W}_i$ as follows:

$$\boldsymbol{U}_{in} = \frac{1}{J} \sum_j \frac{1}{{r_{ijn}}^2} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{H}}, \tag{9}$$

$$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i \boldsymbol{U}_{in})^{-1} \boldsymbol{e}_n, \tag{10}$$

$$\boldsymbol{w}_{in} \leftarrow \frac{\boldsymbol{w}_{in}}{\sqrt{\boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{U}_{in} \boldsymbol{w}_{in}}}, \tag{11}$$

where $\boldsymbol{e}_n \in \mathbb{R}^N$ denotes the unit vector with the nth element equal to unity. To fix the scales of $\boldsymbol{y}_{ij}$ among the frequency bins, the following back-projection technique is applied:

$$\hat{y}_{ijn} \leftarrow [\boldsymbol{W}_i^{-1} (\boldsymbol{e}_n \circ \boldsymbol{y}_{ij})]_{n_{\mathrm{ref}}}, \tag{12}$$

where $\hat{y}_{ijn}$ is the scale-fitted estimated signal, $\circ$ is the Hadamard product (element-wise product), $[\cdot]_n$ is the nth element of the vector, and n_{ref} is the index of the reference channel.

Let DNN$_n$ be the DNN source model that enhances the nth source component from a mixture signal, namely, the scale parameter matrix $\boldsymbol{R}_n$ is estimated by DNN$_n$. DNN$_n$ is pretrained so that $|\boldsymbol{R}_n|^{\cdot 1}$ is predicted from an input mixture spectrogram $|\tilde{\boldsymbol{X}}|^{\cdot 1}$, where $|\cdot|^{\cdot 1}$ denotes the element-wise absolute operation and $\tilde{\boldsymbol{X}} \in \mathbb{C}^{I \times J}$ is a mixture spectrogram in the training data. In the inference for open data, the variance matrix is updated by the pretrained DNN$_n$ as

$$\boldsymbol{R}_n \leftarrow \mathrm{DNN}_n(|\boldsymbol{Y}_n|^{\cdot 1}), \tag{13}$$

$$r_{ijn} \leftarrow \max(r_{ijn}, \epsilon), \tag{14}$$

where DNN$_n(|\boldsymbol{Y}_n|^{\cdot 1})$ is the DNN$_n$ output when the input is $|\boldsymbol{Y}_n|^{\cdot 1}$, ϵ is a small value used to increase the stability of IP, and $\boldsymbol{Y}_n \in \mathbb{C}^{I \times J}$ is the spectrogram of the estimated signal whose elements are y_{ijn}, temporally obtained through the update of $\boldsymbol{W}_i$.

3. PROPOSED METHOD

3.1. Motivation

In ILRMA, the separation performance can be improved by generalizing its source model from the complex Gaussian distribution to the complex GGD [6]. Motivated by this fact, we generalize the source model in IDLMA using the complex GGD, which is expected to improve the performance of source separation, as well as BSS based on ILRMA. To consistently maximize the likelihood in GGD-IDLMA, we derive the maximum likelihood estimate based on the GGD and combine it with the source model inferred by DNN.

3.2. Update rule of spatial model

As in Gauss-IDLMA, GGD-IDLMA combines the blind estimation of $\boldsymbol{W}_i$ and DNN inference of the source model $\boldsymbol{R}_n$. On the basis of Equation 6, the negative log-likelihood of the observed signal is obtained as follows:

$$\mathcal{L}_{\mathrm{GGD}} = \frac{2}{\beta} \sum_{i,j,n} \left[\frac{|\boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{x}_{ij}|^{\beta}}{{r_{ijn}}^{\beta}} + \log {r_{ijn}}^{\beta} \right] - 2J \sum_{i} \log |\det \boldsymbol{W}_i| + \mathrm{const.} \tag{15}$$

When $\beta = 2$, Equation 15 reduces to Equation 8 and IP can be applied. However, when $\beta \neq 2$, IP cannot be applied to Equation 15. This is because IP is only applicable to the sum of a negative log-determinant of $\boldsymbol{W}_i$ and a quadratic form of $\boldsymbol{w}_{in}$, and Equation 15 includes $|\boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{x}_{ij}|^{\beta}$, which is not a quadratic form of $\boldsymbol{w}_{in}$ when $\beta \neq 2$.

To solve this problem, we employ the majorization-minimization (MM) algorithm [14] to minimize Equation 15. The MM algorithm is an optimization technique that minimizes a specially designed majorization instead of minimizing the original cost function. Equation 15 can be transformed into an IP-applicable form by designing its majorization function.

We use the following inequality to design a majorization:

$$|z|^{\beta} \leq \frac{\beta}{2\alpha^{2-\beta}} |x|^2 + \left(1 - \frac{\beta}{2}\right) \alpha^{\beta}, \tag{16}$$

where $0 \le \beta \le 2$, $z \ge 0$ is the variable of the function, and $\alpha \ge 0$ is the auxiliary variable. We derive the majorization function of Equation 15 as follows by applying Equation 16:

$$\begin{aligned}\mathcal{L}_{\mathrm{GGD}} \le \frac{2}{\beta}\sum_{i,j,n}\left[\frac{\beta}{2\alpha_{ijn}{}^{2-\beta}}\frac{|\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2}{r_{ijn}{}^{\beta}} + \frac{2-\beta}{2r_{ijn}{}^{\beta}}\alpha_{ijn}{}^{\beta} + \log r_{ijn}{}^{\beta}\right] - 2J\sum_{i}\log|\det\boldsymbol{W}_i| \\ + \text{const.} \qquad (17) \\ := \mathcal{L}^{+}_{\mathrm{GGD}}, \qquad (18)\end{aligned}$$

where α_{ijn} is an auxiliary variable. The inequality holds if and only if

$$\alpha_{ijn} = |\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij}|. \qquad (19)$$

By applying IP to $\mathcal{L}^{+}_{\mathrm{GGD}}$, the update rule of $\boldsymbol{W}_i$ to minimize Equation 15 is obtained as follows:

$$c_{ijn} = r_{ijn}{}^{\beta}|y_{ijn}|^{2-\beta}, \qquad (20)$$

$$\boldsymbol{U}_{in} = \frac{1}{J}\sum_{j}\frac{1}{c_{ijn}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}}, \qquad (21)$$

$$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i\boldsymbol{U}_{in})^{-1}\boldsymbol{e}_n, \qquad (22)$$

$$\boldsymbol{w}_{in} \leftarrow \frac{\boldsymbol{w}_{in}}{\sqrt{\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{U}_{in}\boldsymbol{w}_{in}}}. \qquad (23)$$

Note that these update rules are valid only for $\beta \le 2$. When $\beta = 2$, Equations 21–23 become identical to Equations 9–11. In the case of Gauss-IDLMA, since the variance $r_{ijn}{}^2$ inferred by the DNN source model often includes an excessively small value, $\boldsymbol{U}_{in}$ obtained by Equation 9 tends to be a rank-deficient matrix, resulting in the numerical instability of optimization. In GGD-IDLMA, on the other hand, the geometric mean of r_{ijn} and $|y_{ijn}|$ weighted by $\beta : 2-\beta$ is used instead of $r_{ijn}{}^2$ in Equation 21, which makes the DNN output smooth and improves the numerical stability of IP. Although a small β improves the stability, it reduces the convergence speed of the optimization in GGD-IDLMA because the inference of the DNN source model is discounted. For this reason, it is expected that there exists a trade-off between the numerical stability and the optimization speed w.r.t. β. We experimentally find an appropriate β value in the experimental section.

3.3. Architecture and training of DNN source model

In this paper, we only focus on the simplest networks, i.e., fully connected DNNs as in conventional Gauss-IDLMA. This is because the aim of IDLMA is to build a framework of DNN-based BSS under a consistent ML criterion and appropriate utilization of DNNs in terms of parameter optimization.

An outline of DNN training is depicted in Fig. 2. To prepare the training data of mixed signals, we define the following vectors:

$$\vec{\boldsymbol{s}}_{jn} = (\tilde{\boldsymbol{s}}^{\mathrm{T}}_{(j-2c)n}, \tilde{\boldsymbol{s}}^{\mathrm{T}}_{(j-2c+2)n}, \cdots, \tilde{\boldsymbol{s}}^{\mathrm{T}}_{(j+2c)n})^{\mathrm{T}}, \qquad (24)$$

$$\vec{\boldsymbol{x}}_j = \frac{\sum_n \alpha_{jn}\vec{\boldsymbol{s}}_{jn}}{\|\sum_n \alpha_{jn}\vec{\boldsymbol{s}}_{jn}\|_2 + \delta_1}, \qquad (25)$$

$$\bar{\boldsymbol{s}}_{jn} = \frac{\alpha_{jn}\tilde{\boldsymbol{s}}_{jn}}{\|\sum_n \alpha_{jn}\vec{\boldsymbol{s}}_{jn}\|_2 + \delta_1}, \qquad (26)$$

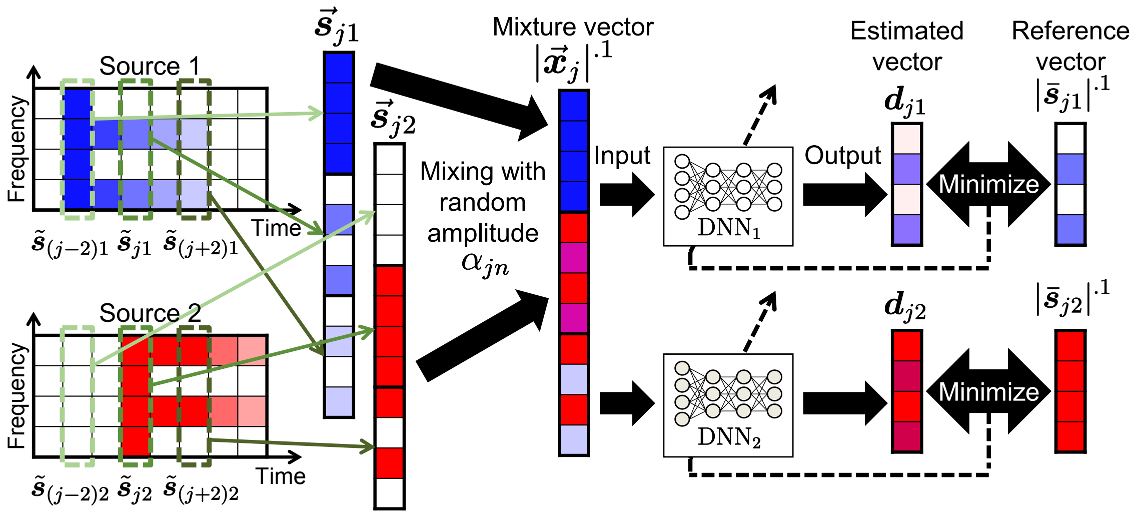

Figure 2: Outline of DNN training when $I = 4$, $J = 8$, $N = 2$, and $c = 1$.

where $\|\cdot\|_2$ denotes the Euclidean norm, $\tilde{\boldsymbol{s}}_{jn} \in \mathbb{C}^I$ is the STFT vector of the nth source at j, $\vec{\boldsymbol{s}}_{jn} \in \mathbb{C}^{I(2C+1)}$ is a vector that vertically concatenates $\tilde{\boldsymbol{s}}_{jn}$ for $2c$ frames around j as shown in Fig. 2, $\vec{\boldsymbol{x}}_j \in \mathbb{C}^{I(2C+1)}$ is the normalized mixture vector whose amplitude $|\vec{\boldsymbol{x}}_j|^{.1}$ is an input vector for all DNN$_n$, $\bar{\boldsymbol{s}}_{jn} \in \mathbb{C}^I$ is the reference vector for each source, α_{jn} is a random variable in the range [0.05, 1], which controls the SNR in $\vec{\boldsymbol{x}}_j$, and δ_1 is a small value to avoid division by zero. DNN$_n$ is optimized so that the following loss function between the output vector $\boldsymbol{d}_{jn} \in \mathbb{R}^I_{\geq 0}$ and the reference vector $|\bar{\boldsymbol{s}}_{jn}|^{.1}$ is minimized:

$$L_{\mathrm{GGD}} = \frac{1}{IJ} \sum_{i,j} \left[\frac{|\bar{s}_{ijn}|^{\beta} + \delta_2}{{d_{ijn}}^{\beta} + \delta_2} - \log \frac{|\bar{s}_{ijn}|^{\beta} + \delta_2}{{d_{ijn}}^{\beta} + \delta_2} \right], \tag{27}$$

where δ_2 is a small value for numerical stability and $\bar{s}_{ijn}$ and d_{ijn} are the elements of $\bar{\boldsymbol{s}}_{jn}$ and $\boldsymbol{d}_{jn}$, respectively. Since minimizing Equation 27 is equivalent to the maximum likelihood estimation of r_{ijn} in Equation 15, DNN$_n$ trained with this cost function can be interpreted as an appropriate source model. After the training, the scale parameter $\boldsymbol{R}_n$ is inferred by Equations 13 and 14 as in Gauss-IDLMA.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

To confirm the validity of the proposed GGD-IDLMA, we conducted an experiment on music source separation. We compared six methods: MNMF with 20 NMF bases (BSS), ILRMA with 20 NMF bases (BSS), DNN+WF, Duong+DNN, Gauss-IDLMA, and the proposed GGD-IDLMA, where DNN+WF is a combined method employing the pretrained DNN source model and Wiener filtering (monaural source separation) [15]. For Duong+DNN and IDLMA, the scale parameter matrix $\boldsymbol{R}_n$ was updated by DNN$_n$ after every 10 iterations of the spatial parameter optimization (IP). Note that the DNNs employed in this paper were different from those of the original Duong+DNN [11] as follows: (a) each DNN was prepared for each single source, and (b) each DNN was trained under multiple-SNR conditions using the random amplitude α_{jn}.

We used the DSD100 dataset of SiSEC2016 [16] as the dry sources and the training datasets of each DNN. The 50 songs in the dev data were used to train DNN$_n$ and the top

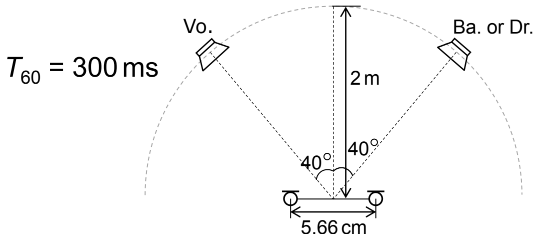

Figure 3: Recording conditions of impulse responses obtained from RWCP database.

25 songs in alphabetical order in the `test` data were used for performance evaluation. The test songs were trimmed to the interval of 30 to 60 s. To simulate reverberant mixtures, we produced two-channel observed signals by convoluting the impulse response E2A (T_{60} = 300 ms) obtained from the RWCP database [17] with each source, and mixtures of bass (Ba.) and vocals (Vo.) (Ba./Vo.) and mixtures of drums (Dr.) and Vo. (Dr./Vo.) ware produced. The recording conditions of E2A are shown in Fig. 3. All the signals were downsampled to 8 kHz. An STFT was performed using a 512-ms-long Hamming window with a 256-ms-long shift for Ba./Vo. separation and a 256-ms-long Hamming window with a 128-ms-long shift for Dr./Vo. separation. We used the signal-to-distortion ratio (SDR) [18] to evaluate the total separation performance. The number of hidden layers in the DNN was set to four. Each layer had 1024 units, and a rectified linear unit was used for the output of each layer. To optimize the DNN, we added the term $(\lambda/2)\sum_q {g_q}^2$ to Equation 27 for regularization, where g_q is the weight coefficient in DNN, and ADADELTA [19] with a 128-size minibatch was performed for 2000 epochs. The parameter ϵ was experimentally optimized and set to $0.1 \times (IJ)^{-1} \sum_{i,j} {r_{ijn}}^2$ for Gauss-IDLMA and $0.35 \times (IJ)^{-1} \sum_{i,j} {r_{ijn}}^2$ for GGD-IDLMA. The other parameters were set to $\delta_1 = \delta_2 = 10^{-5}$, $c = 3$, and $\lambda = 10^{-5}$.

4.2. Results

Figures 4 and 5 show the average SDR improvements of Ba./Vo. and Dr./Vo., respectively. The results show the efficacy of a time-frequency-varying complex GGD as a source model. In particular, although the DNN separation performance (DNN+WF) when $\beta = 1.94$ is not the highest, GGD-IDLMA with $\beta = 1.94$ achieves the best separation performance of all the methods in both Ba./Vo. and Dr./Vo. separation, which supports the trade-off between the numerical stability and optimization speed when the demixing matrix is updated as discussed in Sect. 3.2.

5. CONCLUSIONS

In this paper, we generalized the source model of IDLMA to a time-frequency-varying complex GGD. The update rule of the demixing matrix in GGD-IDLMA shows that the proposed GGD-IDLMA decreases the numerical instability caused by the DNN chasm while also reducing the optimization speed. Experimental evaluation showed the efficacy of the proposed GGD-IDLMA compared with state-of-the-art separation methods including conventional IDLMA.

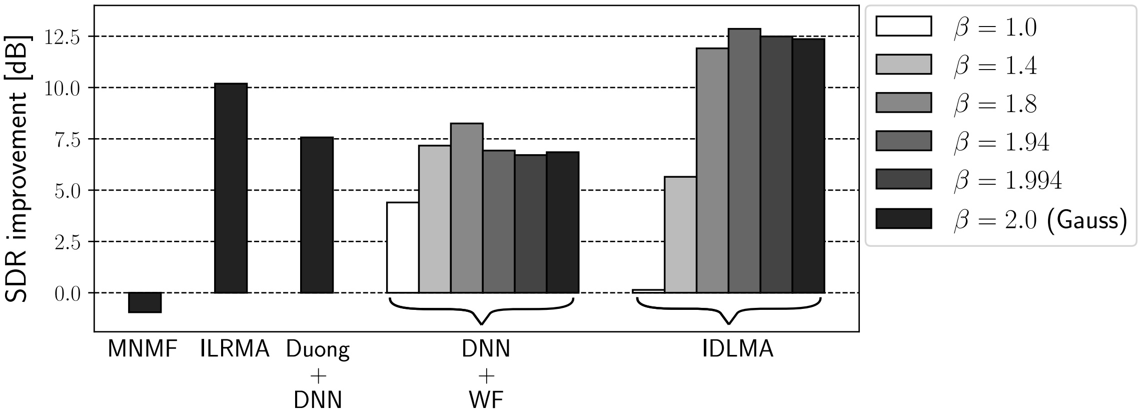

Figure 4: Average SDR improvements of 25 Ba./Vo. songs.

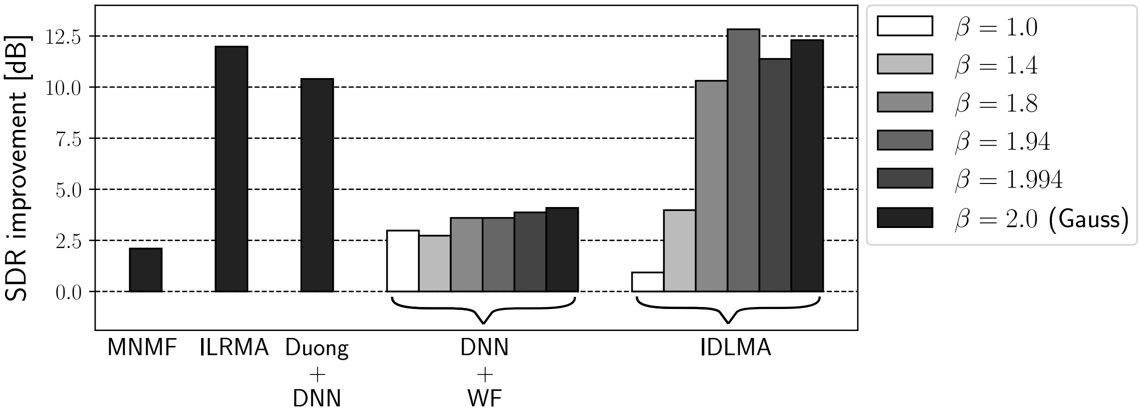

Figure 5: Average SDR improvements of 25 Dr./Vo. songs.

6. ACKNOWLEDGEMENTS

This work was partly supported by SECOM Science and Technology Foundation and JSPS KAKENHI Grant Numbers JP17H06101 and JP17H06572.

7. REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

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

[3] 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. ASLP*, vol. 14, no. 9, pp. 1626–1641, 2016.

[4] 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*, S. Makino, Ed. Springer, Cham, 2018, ch. 6, pp. 125–155.

[5] 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, 1999.

[6] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 28, pp. 1–25, 2018.

[7] R. Ikeshita and Y. Kawaguchi, "Independent low-rank matrix analysis based on multivariate complex exponential power distribution," in *Proc. ICASSP*, 2018, pp. 741–745.

[8] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[9] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[10] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[11] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.

[12] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for multichannel audio source separation," in *Proc. EUSIPCO*, 2018, pp. 1571–1575.

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

[14] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *J. Comput. Graph. Stat.*, vol. 9, no. 1, pp. 60–77, 2000.

[15] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. ICASSP*, 2015, pp. 2135–2139.

[16] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. LVA/ICA*, 2012, pp. 323–332.

[17] S. Nakamura, K. Hiyane, F. Asano, T. Nishimura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. LREC*, 2000, pp. 965–968.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.