Auxiliary-function-based Independent Vector Analysis with Power of Vector-norm Type Weighting Functions

Nobutaka Ono*
* National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
E-mail: onono@nii.ac.jp

***Abstract*—In this paper, we present an auxiliary-function-based independent vector analysis (AuxIVA) based on the Generalized super Gaussian source model or Gaussian source model with time-varying variance. AuxIVA is a convergence-guaranteed iterative algorithm for independent vector analysis (IVA) with a spherical and super Gaussian source model, and the source model can be characterized by a weighting function. We show that both of the generalized Gaussian source models with the shape parameter $0 < \beta \leq 2$ and the Gaussian source model with time-varying variance unifiedly yield a power of vector-norm type weighting functions. A scaling and a clipping technique for numerical stability are discussed. The dependency of the separation performance on the source model is also investigated.**

I. Introduction

Blind source separation (BSS) is one of the important signal processing techniques for extracting a desired source from mixtures, and it has still been one of the most interesting topics in the signal processing field. In recent years, multivariate-type independent component analysis (ICA), which can be referred to as independent vector analysis (IVA), was developed [1], [2] and applied to the frequency-domain approach for convolutive mixtures [3]. In IVA, the whole frequency components of a source are modeled as a stochastic vector variable. Thanks to modeling the dependencies over frequency components, IVA is theoretically not affected by the permutation ambiguity, unlike conventional frequency-domain ICA.

Recently, a fast and robust update rule for IVA was developed [4] that is based on auxiliary function technique [5], which is denoted as auxiliary-function-based IVA (AuxIVA), and its implementation in the iPhone was also investigated [6]. In AuxIVA, a spherical and super Gaussian source model is used, and the separation performance should depend on the source model, and this source model can be characterized by a weighting function.

In this paper, as typical source models in AuxIVA, the generalized Gaussian source model with the shape parameter $0 < \beta \leq 2$ and the Gaussian source model with time-varying variance are focused on. We show that both of them unifiedly yield a power of vector-norm type weighting functions. The dependency of the separation performance on the source model is also investigated.

II. Independent Vector Analysis

A. BSS in Frequency Domain

Assume here that K sources are observed by K microphones and that their short-time Fourier transform (STFT) representations are obtained. Let $\boldsymbol{s}(\omega,\tau)$, $\boldsymbol{x}(\omega,\tau)$, and $\boldsymbol{y}(\omega,\tau)$ be the frequency-wise vector representation of the sources, the observations, and the estimated sources, respectively, which are defined as

$$\boldsymbol{s}(\omega,\tau) = (s_1(\omega,\tau)\ \cdots\ s_K(\omega,\tau))^t, \quad (1)$$

$$\boldsymbol{x}(\omega,\tau) = (x_1(\omega,\tau)\ \cdots\ x_K(\omega,\tau))^t, \quad (2)$$

$$\boldsymbol{y}(\omega,\tau) = (y_1(\omega,\tau)\ \cdots\ y_K(\omega,\tau))^t, \quad (3)$$

where ^{t} denotes the vector transpose, and the size of each vector is $K \times 1$. In the frequency-domain approach for a convolutive mixture, a linear mixing model,

$$\boldsymbol{x}(\omega,\tau) = A(\omega)\boldsymbol{s}(\omega,\tau), \quad (4)$$

is assumed, where $A(\omega)$ is a $K \times K$ mixing matrix. The sources are estimated by a linear demixing process,

$$\boldsymbol{y}(\omega,\tau) = W(\omega)\boldsymbol{x}(\omega,\tau), \quad (5)$$

where

$$W(\omega) = (\boldsymbol{w}_1(\omega)\ \cdots\ \boldsymbol{w}_K(\omega))^h \quad (6)$$

is a $K \times K$ demixing matrix, and ^{h} denotes Hermitian transpose.

B. Objective Function of IVA

In IVA, assuming a multivariate p.d.f. for sources to exploit the dependencies over frequency components, the demixing matrices are estimated by minimizing the following objective function.

$$J(\boldsymbol{W}) = \sum_{k=1}^{K} \frac{1}{N_\tau} \sum_{\tau=1}^{N_\tau} G(\boldsymbol{y}_k(\tau)) - \sum_{\omega=1}^{N_\omega} \log|\det W(\omega)|, \quad (7)$$

where $\boldsymbol{W}$ denotes a set of $W(\omega)$, N_ω is the number of frequency bins, N_τ is the number of time frames, $\boldsymbol{y}_k(\tau)$ is the source-wise vector representation with the size $N_\omega \times 1$ defined as

$$\boldsymbol{y}_k(\tau) = (y_k(1,\tau)\ \cdots\ y_k(N_\omega,\tau))^t, \quad (8)$$

and $G(\boldsymbol{y}_k(\tau))$ is called a contrast function. When $G(\boldsymbol{y}_k(\tau)) = -\log p(\boldsymbol{y}_k(\tau))$, where $p(\boldsymbol{y}_k(\tau))$ represents a multivariate p.d.f. of a source, the minimization of eq. (7) is equivalent to the maximum likelihood (ML) estimation.

III. Overview of AuxIVA [4]

A. Conditions for contrast function

In AuxIVA, the following two conditions are assumed for the contrast function.

Spherical symmetry $G(\boldsymbol{y}_k(\tau))$ is assumed to be a function of only the L_2 norm of $\boldsymbol{y}_k(\tau)$. This means that $G(\boldsymbol{y}_k(\tau))$ can be represented as

$$G(\boldsymbol{y}_k(\tau)) = G_R(r_k(\tau)), \tag{9}$$

$$r_k(\tau) = ||\boldsymbol{y}_k(\tau)||_2, \tag{10}$$

where $G_R(r)$ is a function of a real-valued scalar variable, r.

Super Gaussianity $G_R(r)$ is assumed to be a continuous and differentiable function of r, satisfying the condition that $G'_R(r)/r$ is positive and continuous everywhere and is monotonically decreasing in the wider sense in $r \geq 0$. Taking the relationship $G(\boldsymbol{y}_k(\tau)) = -\log p(\boldsymbol{y}_k(\tau))$ into account, this means that a multivariate p.d.f. of a source, $p(\boldsymbol{y}_k(\tau))$, is a Gaussian or a super Gaussian distribution.

B. Auxiliary function for IVA

When $G(\boldsymbol{y}_k(\tau))$ satisfies these two conditions,

$$G(\boldsymbol{y}_k(\tau)) \leq \frac{G'_R(r_0)}{2r_0}||\boldsymbol{y}_k(\tau)||_2^2 + \left(G_R(r_0) - \frac{r_0 G'_R(r_0)}{2}\right) \tag{11}$$

holds for any $\boldsymbol{y}_k(\tau)$ and r_0. The equality sign is satisfied if and only if $r_0 = ||\boldsymbol{y}_k(\tau)||_2$.

On the basis of this inequality, the following auxiliary function can be derived.

$$Q(\boldsymbol{W}, \boldsymbol{r}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{\omega=1}^{N_\omega} \boldsymbol{w}_k^h(\omega)V_k(\omega)\boldsymbol{w}_k(\omega) - \sum_{\omega=1}^{N_\omega} \log|\det W(\omega)| + R, \tag{12}$$

$$V_k(\omega) = \frac{1}{N_\tau}\sum_{\tau=1}^{N_\tau}\left[\frac{G'_R(r_k(\tau))}{r_k(\tau)}\boldsymbol{x}(\omega,\tau)\boldsymbol{x}^h(\omega,\tau)\right], \tag{13}$$

where $\boldsymbol{r}$ denotes a set of auxiliary variables, $r_k(\tau)$, and R is a constant independent of $\boldsymbol{W}$. For any $\boldsymbol{W}$ and $\boldsymbol{r}$,

$$J(\boldsymbol{W}) \leq Q(\boldsymbol{W}, \boldsymbol{r}) \tag{14}$$

holds. The equality sign holds if and only if

$$r_k(\tau) = ||\boldsymbol{y}_k(\tau)||_2 = \sqrt{\sum_{\omega=1}^{N_\omega}|\boldsymbol{w}_k^h(\omega)\boldsymbol{x}(\omega,\tau)|^2}. \tag{15}$$

C. Update Rules

Auxiliary variable updates: Update the weighted covariance matrices $V_k(\omega)$ for all ω as follows.

$$r_k(\tau) = \sqrt{\sum_{\omega=1}^{N_\omega}|\boldsymbol{w}_k^h(\omega)\boldsymbol{x}(\omega,\tau)|^2}, \tag{16}$$

$$\phi(r_k(\tau)) = \frac{G'_R(r_k(\tau))}{r_k(\tau)}, \tag{17}$$

$$V_k(\omega) = \frac{1}{N_\tau}\sum_{\tau=1}^{N_\tau}\left[\phi(r_k(\tau))\boldsymbol{x}(\omega,\tau)\boldsymbol{x}^h(\omega,\tau)\right]. \tag{18}$$

Demixing matrix updates: Apply the following updates in order for all ω.

$$\boldsymbol{w}_k(\omega) \leftarrow (W(\omega)V_k(\omega))^{-1}\boldsymbol{e}_k, \tag{19}$$

$$\boldsymbol{w}_k(\omega) \leftarrow \boldsymbol{w}_k(\omega)/\sqrt{\boldsymbol{w}_k^h(\omega)V_k(\omega)\boldsymbol{w}_k(\omega)}. \tag{20}$$

In AuxIVA, $\phi(r)$ is a key function rather than the contrast function $G(\boldsymbol{y}_k(\tau))$ or $G_R(r)$. Hereafter, we denote $\phi(r)$ as a weighting function because it works as a weight for calculating a weighted covariance matrix, $V_k(\omega)$.

IV. Power of Vector-norm Type Weighting Functions

A. Generalized Gaussian Source Models

One of the typical super Gaussian distributions can be given by generalized Gaussian distribution with an appropriate shape parameter [7]. If we assume a spherical complex-valued generalized Gaussian distribution such as

$$p(\boldsymbol{y}_k(\tau)) \propto \exp\left\{-\left(\frac{||\boldsymbol{y}_k(\tau)||_2}{\alpha}\right)^\beta\right\}, \tag{21}$$

as the p.d.f of the source, the corresponding contrast function and the weighting function can be represented as follows.

$$G_R(r) = \left(\frac{r}{\alpha}\right)^\beta, \tag{22}$$

$$\phi(r) = \beta\alpha^{-\beta}r^{\beta-2}, \tag{23}$$

where α and β denote a scale and a shape parameter, respectively, and $0 < \beta \leq 2$ is necessary for super Gaussianity.

B. Gaussian Source Model with Time-varying Variance

Instead of a stationary super Gaussian model, we can assume that a source-wise vector $\boldsymbol{y}_k(\tau)$ follows a non-stationary Gaussian distribution such as

$$p(\boldsymbol{y}_k(\tau); \sigma_k^2(\tau)) \propto \exp\left(-\frac{||\boldsymbol{y}_k(\tau)||_2^2}{2\sigma_k^2(\tau)}\right), \tag{24}$$

where $\sigma_k^2(\tau)$ denotes the time-varying variance [8][1]. In this case, the corresponding contrast function and the weighting

[1] In this paper, we focus on only Gaussian distribution as the time-varying model because a *generalized* Gaussian model with time-varying variance does not yield the weighting function with the unified form of eq. (29)

function can be represented as follows.

$$G_R(r) = \frac{r^2}{2\sigma_k^2(\tau)}, \tag{25}$$

$$\phi(r) = \frac{1}{\sigma_k^2(\tau)}. \tag{26}$$

By replacing the unknown $\sigma_k^2(\tau)$ by its ML estimation, we have

$$\hat{\sigma}_k^2(\tau) = \frac{1}{N_\omega}\sum_\omega ||y_k(\omega,\tau)||_2^2 = \frac{r^2}{N_\omega}, \tag{27}$$

which is equivalent to alternatively minimizing the objective function $J(\boldsymbol{W}, \boldsymbol{\sigma}^2)$ in terms of $\boldsymbol{W}$ and $\boldsymbol{\sigma}^2$, where $\boldsymbol{\sigma}^2$ is a set of $\sigma_k^2(\tau)$. Then, we have

$$\phi(r) = N_\omega r^{-2}. \tag{28}$$

C. Unified Weighting Function Form

Both eq. (23) and eq. (28) can be represented as the power of a vector norm r such as

$$\phi(r) = \gamma r^{\beta-2}, \tag{29}$$

where γ is a scale parameter. Note that $0 \le \beta \le 2$ is necessary for AuxIVA. If $\beta < 0$, minimizing the objective function loses the meaning of ML estimation because the integral of the corresponding probability function $p(\boldsymbol{y}_k(\tau))$ does not converge, while, if $\beta > 2$, AuxIVA is invalid because eq. (11) is not satisfied.

V. Scaling and Clipping of Weighting Function for Numerical Stability

A. Scaling

Let $\hat{W}(\omega)$ be the demixing matrix at the convergence point of AuxIVA when the weighting function defined in eq. (29) with $\gamma = 1$ is used. $\hat{W}(\omega)$ should satisfy

$$\hat{\boldsymbol{w}}_l^h(\omega)V_k(\omega)\hat{\boldsymbol{w}}_k(\omega) = \delta_{kl}, \tag{30}$$

where

$$V_k(\omega) = \frac{1}{N_\tau}\sum_{\tau=1}^{N_\tau}\left[\gamma\hat{r}_k^{\beta-2}(\tau)\boldsymbol{x}(\omega,\tau)\boldsymbol{x}^h(\omega,\tau)\right], \tag{31}$$

$$\hat{r}_k(\tau) = \sqrt{\sum_{\omega=1}^{N_\omega}|\hat{\boldsymbol{w}}_k^h(\omega)\boldsymbol{x}(\omega,\tau)|^2}, \tag{32}$$

and $\gamma = 1$. Then, we can easily confirm that when a scale parameter, $\gamma = \gamma_0$, is used, $C\hat{W}(\omega)$ becomes the convergence point where $C^\beta\gamma_0 = 1$. Therefore, the scale parameter γ does not matter theoretically because it only determines a scale of $\boldsymbol{W}$, and it will be adjusted by the followed projection back operation.

However, a simple setting, $\gamma = 1$, may cause a very huge magnitude of $W(\omega)$, especially when β is close to or equal to 0. Let us assume that $\boldsymbol{y}_k(\tau)$ actually follows eq. (21). Then, the variance of $\boldsymbol{y}_k(\tau)$ can be obtained as

$$E[||\boldsymbol{y}_k(\tau)||_2^2] = \alpha^2\frac{N_\omega\Gamma(1+\frac{2}{\beta}(N_\omega+1))}{(N_\omega+1)\Gamma(1+\frac{2}{\beta}N_\omega)}, \tag{33}$$

where $E[\cdot]$ denotes the expectation operation. Applying the well-known Starling's approximation to Gamma functions such as $\Gamma(1+z) \simeq \sqrt{2\pi z}(z/e)^z$ to eq. (33), we have

$$E[||\boldsymbol{y}_k(\tau)||_2^2] \simeq \alpha^2\left(\frac{N_\omega+1}{N_\omega}\right)^{\frac{2}{\beta}N_\omega-\frac{1}{2}}\left(\frac{2(N_\omega+1)}{\beta e}\right)^{\frac{2}{\beta}}, \tag{34}$$

which is larger than 10^{33} for $N_\omega = 1025$, $\beta = 0.2$, and $\gamma = 1$. Thus, such a huge magnitude can cause numerical instability or divergence even in floating point calculation.

To avoiding this, the scale normalization

$$W(\omega) \leftarrow W(\omega)/\sqrt{\frac{1}{N_\omega N_\tau K}\sum_k\sum_\tau ||\boldsymbol{y}_k(\tau)||_2^2} \tag{35}$$

is here introduced after each iteration of AuxIVA such that $E[|y_k(\omega,\tau)|^2] = 1$ is satisfied.

B. Clipping

Obviously, the weighting function $r^{\beta-2}$ diverges when $r = 0$. Even though the observation is not silent, an estimated source can be accidently close to silent at a frame during iterations. In this case, the weighting function becomes a very huge value at the frame, which leads numerical divergence or over-fitting of V_k to only the frame. To avoid this, the clipping of the weighting function is also introduced here as

$$\phi(r) = \min\{\phi_0, r^{\beta-2}\}, \tag{36}$$

where ϕ_0 denotes a clipping value.

VI. Experimental Evaluations

The separation performances of AuxIVA with different βs in eq. (36) were compared with experiments by using synthesized convolutive mixtures of speech. The impulse responses from nine directions recorded in two variable reverberation rooms (E2A and E2B) from RWCP Sound Scene Database in Real Acoustical Environments [10] were used. Note that the reverberation time of E2B is very long (1.3 s). Also, we selected nine speech utterances from the ATR Japanese speech database (Set B), assigned them to each of the nine directions, convoluted each of them after downsampling to 16 kHz, and mixed them. We prepared two mixtures ($K = 2$) and three mixtures ($K = 3$) of all combinations (${}_9C_2 = 36$ and ${}_9C_3 = 84$, respectively) of them. Other experimental conditions are summarized in Table I.

The AuxIVA update with the weighting function of eq. (36) including the normalization of eq. (35) was applied to all mixtures, and $\beta = 0, 0.2, 0.4, 0.6, 0.8$ and 1 were compared. Note that $\beta = 1$ and $\beta = 0$ corresponded to the time-invariant Laplacian source model [4], [6] and the time-varying Gaussian source model [8], respectively. The initial value of the demixing matrix was given by the identity matrix for simplicity. A clipping value, $\phi_0 = 1000$, was experimentally determined. Although AuxIVA almost converged at 10 or 20 iterations in most cases, we applied 50 iterations to evaluate the best separation performance in this experiment. No divergence happened over all trials for any βs. The estimated sources were

TABLE I
EXPERIMENTAL CONDITIONS

room type	E2A	E2B
reverberation time	0.3s	1.3s
microphone spacing	2.83cm	
source-microphone distance	2m	
source direction	10° to 170° by 20°	
frame length	4096	8192
frame shift	2048	4096
window function	hamming	
signal length	10s	
sampling frequency	16kHz	

calculated by applying the estimated demixing matrix with the projection back operation [9]. The performance was evaluated by the average of the SDR over all trials calculated by using the BSS toolbox [11].

The resultant SDRs for different βs at each of the four conditions (the number of sources, 2 or 3, by room type, E2A or E2B) are shown in Fig. 1. The bars and the error bars indicate the averaged SDR and the quantile range over all trials, respectively. The best choice of β slightly depends on the conditions, but $\beta = 0.2$ or $\beta = 0.4$ showed almost the best performance at all conditions.

VII. Conclusions

In this paper, AuxIVA with the power of vector norm type weighting function $\phi(r) = r^{\beta-2}$ was presented, where $0 \leq \beta \leq 2$. The normalization and the clipping operation were discussed for numerical stability. The experimental results suggest that $\beta = 0.2 \sim 0.4$ should be a good choice for speech separation.

References

[1] A. Hiroe, "Solution of Permutation Problem in Frequency Domain ICA Using Multivariate Probability Density Functions," *Proc. ICA*, pp. 601–608, 2006.

[2] T. Kim, T. Eltoft, and T.-W. Lee, "Independent Vector Analysis: An Extension of ICA to Multivariate Components," *Proc. ICA*, pp. 165–172, 2006.

[3] P. Smaragdis, "Blind Separation of Convolved Mixtures in the Frequency Domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[4] N. Ono, "Stable and Fast Update Rules for Independent Vector Analysis Based on Auxiliary Function Technique," *Proc. WASPAA*, pp. 189-192, Oct. 2011.

[5] N. Ono and S. Miyabe, "Auxiliary-function-based Independent Component Analysis for Super-Gaussian Sources," *Proc. LVA/ICA*, pp.165-172, 2010.

[6] N. Ono, "Fast Stereo Independent Vector Analysis and its Implementation on Mobile Phone," *Proc. IWAENC*, Sept. 2012.

[7] T. Itahashi and K. Matsuoka, "Stability of Independent Vector Analysis," Signal Processing, vol. 92, no. 8, pp. 1809-1820, 2012.

[8] T. Ono, N. Ono, and S. Sagayama, "User-guided Independent Vector Analysis with Source Activity Tuning," Proc. ICASSP, pp. 2417–2420, Mar. 2012.

[9] N. Murata, S. Ikeda, and A. Ziehe, "An Approach to Blind Source Separation Based on Temporal Structure of Speech Signals," *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

[10] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition," *Proc. LREC*, pp. 965-968, 2000.

[11] E. Vincent, C. Fevotte, and R. Gribonval, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

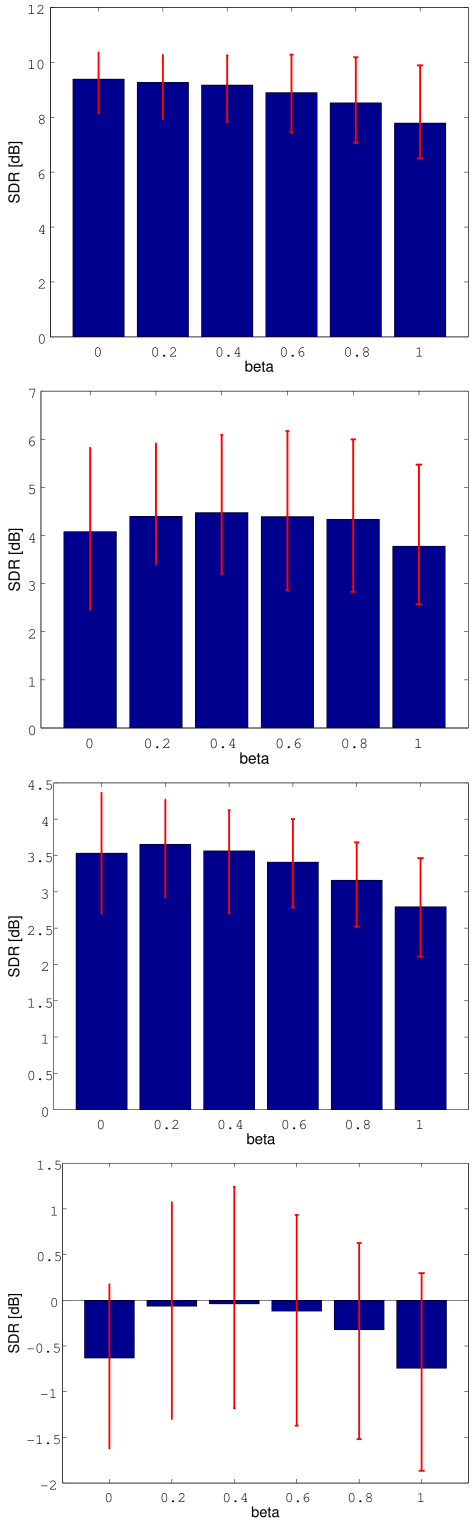

Fig. 1. The resultant SDR for different βs. From top to bottom, the results in the conditions of two mixtures in room E2A, two mixtures in room E2B, three mixtures in room E2A, and three mixtures in room E2B are shown.