ADAPTIVE STEP-SIZE PARAMETER CONTROL FOR REAL-WORLD BLIND SOURCE SEPARATION

Hirofumi Nakajima, Kazuhiro Nakadai, Yuji Hasegawa and Hiroshi Tsujino

Honda Research Institute Japan Co., Ltd
Honcho 8-1, Wako-shi, Saitama, 351-0188 Japan

ABSTRACT

This paper describes a method to adaptively control a step-size parameter which is used for updating a separation matrix to extract a target sound source accurately in blind source separation (BSS). The design of the step-size parameter is essential when we apply BSS to real-world applications such as robot audition systems, because the surrounding environment dynamically changes in the real world. It is common to use a fixed step-size parameter that is obtained empirically. However, due to environmental changes and noises, the performance of BSS with the fixed step-size parameter deteriorates and the separation matrix sometimes diverges. We propose a general method that allows adaptive step-size control. The proposed method is an extension of Newton's method utilizing a complex gradient theory and is applicable to any BSS algorithm. Actually, we applied it to six types of BSS algorithms for an 8 ch microphone array embedded in Honda ASIMO. Experimental results show that the proposed method improves the performance of these six BSS algorithms through experiments of separation and recognition for two simultaneous speeches.

Index Terms— robot audition, blind source separation, adaptive step-size, Newton's method

1. INTRODUCTION

For natural human-robot interaction, a robot should have auditory functions [1]. In the real-world environment where the robot is expected to work properly, the robot should cope with dynamically-changing noise sources including its own motor noises and speech interference like barge-in. *Sound Source Separation (SSS)* is, thus, essential for the robot. *Blind Source Separation (BSS)* is often used as an SSS algorithm [2, 3], because it shows high performance without using any transfer function between a microphone and a sound source. However, most BSS algorithms have difficulties in separation speed and accuracy in a dynamically-changing environment, because they use a fixed step-size parameter which is manually tuned to a specific stationary environment. Therefore, we propose a general framework to allow an adaptive step-size parameter based on Newton's method to improve BSS performance in the real world.

2. ADAPTIVE STEP-SIZE PARAMETER CONTROL

2.1. General BSS Formulation

Fig. 1 shows the general system model for SSS. Suppose that there are M sources and N ($\leq M$) microphones. A spectrum vector of M sources at frequency ω, $\mathbf{s}(\omega)$, is denoted as $[s_1(\omega)s_2(\omega)...s_M(\omega)]^T$, and a spectrum vector of signals captured by the N microphones at frequency ω, $\mathbf{x}(\omega)$, is denoted as $[x_1(\omega)x_2(\omega)...x_N(\omega)]^T$. $\mathbf{x}(\omega)$ is, then, calculated as

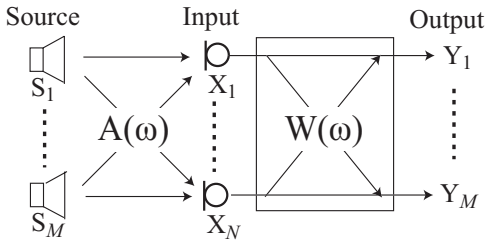

Fig. 1. System Model for Blind Source Separation

$$\mathbf{x}(\omega) = \mathbf{H}(\omega)\mathbf{s}(\omega), \tag{1}$$

where $\mathbf{H}(\omega)$ is a transfer function (TF) matrix. Each component H_{ji} of the TF matrix represents the TF from the i-th source to the j-th microphone. SSS is then formulated as

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

where $\mathbf{W}(\omega)$ is called a *separation matrix*. SSS is defined as a problem to find $\mathbf{W}(\omega)$ which satisfies the condition that output signal $\mathbf{y}(\omega)$ is the same as $\mathbf{s}(\omega)$. If $\mathbf{H}(\omega)$ is obtained precisely, $\mathbf{W}(\omega)$ is easily estimated by calculating the pseudo inverse $\mathbf{H}^{+}(\omega)$. However, it is difficult to obtain $\mathbf{H}(\omega)$ precisely.

BSS solves this problem because it is able to separate sound sources even when $\mathbf{H}(\omega)$ is unknown or only a part of $\mathbf{H}(\omega)$ such as direct sound components is given. BSS is formulated by obtaining an optimal separation matrix $\mathbf{W}_{opt}$ without using any prior information such as $\mathbf{H}(\omega)$. $\mathbf{W}_{opt}$ is estimated by minimizing a cost function $J(\mathbf{y})$ which denotes the mixture degree of $\mathbf{y}$.

$$\mathbf{W}_{opt} = \underset{\mathbf{W}}{argmin}[J(\mathbf{y})] = \underset{\mathbf{W}}{argmin}[J(\mathbf{W}\mathbf{x})]. \tag{3}$$

To obtain $\mathbf{W}_{opt}$, BSS updates $\mathbf{W}$ to minimize $J(\mathbf{y})$ by using

$$\mathbf{W}_{t+1} \quad = \quad \mathbf{W}_t - \mu\mathbf{J}'(\mathbf{W_t}). \tag{4}$$

where $\mathbf{W}_t$ denotes $\mathbf{W}$ at the current time step t, $\mathbf{J}'(\mathbf{W})$ is defined as the update direction of $\mathbf{W}$, and μ means a step-size parameter. Most BSS algorithms use a fixed frequency-independent value as the step-size parameter. However, the fixed step-size has several problems as mentioned in Sec. 1.

2.2. General Formulation of Adaptive Step-Size Parameter Control for BSS

This section describes the formulation of an adaptive step-size parameter control method which is generally applicable to BSS. The use of an adaptive step-size parameter is well-studied in the field

of echo cancellation [4]. However, most adaptive step-size methods for echo cancellation like normalized LMS assume a single channel input and signal processing only with real numbers. To apply such an adaptive step-size method to BSS, we extended it to support multi-channel input and complex number signals. To realize this, we introduced the multi-dimensional Newton's method and linear approximation formula for a complex gradient matrix. According to the complex gradient theory [5], $J(\mathbf{W})$ around $J(\mathbf{W}_t)$ is approximated as

$$J(\mathbf{W}) \approx J(\mathbf{W}_t) + 2\mathrm{MA}(\nabla_{w*}J(\mathbf{W}), \mathbf{W} - \mathbf{W}_t), \quad (5)$$

where $\mathrm{MA}(\mathbf{A}, \mathbf{B}) = Re[\sum_{i,j} a^*_{i,j} b_{i,j}]$, which represents the real-part sum of all products of the matrices $\mathbf{A}^*$ and $\mathbf{B}$, and ∇_{w*} is the complex gradient operator [5]. μ becomes the optimal value μ_{opt} when $J(\mathbf{W}) = 0$. Thus, from Eqs. (4) and (5), μ_{opt} is defined as

$$\mu_{\mathrm{opt}} = \frac{J(\mathbf{W}_t)}{2\mathrm{MA}(\nabla_{w*}J(\mathbf{W}_t), \mathbf{J}'(\mathbf{W}_t))} \quad (6)$$

Eq. (6) shows the general formulation of the adaptive method. It is easily applicable to any kind of BSS by replacing $J(\mathbf{W})$ with that for the target BSS algorithm. If $\mathbf{J}'(\mathbf{W}) = \nabla_{w*}J(\mathbf{W})$, μ_{opt} is simplified as

$$\mu_{opt} = \frac{J(\mathbf{W}_t)}{2\|\mathbf{J}'(\mathbf{W}_t)\|^2}, \quad (7)$$

where $\|\cdot\|^2$ means the Frobenius norm.

Using our adaptive method, the step-size becomes large when a separation error is high, for example, due to source position changes. It will be low when the error is small due to the convergence of the separation matrix.

3. APPLICATION TO BSS ALGORITHMS

We applied our proposed adaptive step-size parameter control to six types of BSS algorithms, *Decorrelation based Source Separation (DSS)*, *Independent Component Analysis (ICA)*, *Geometric-constrained Source Separation (GSS)*, *Geometric-constrained ICA (GICA)*, *High-order DSS (HDSS)*, and *Geometric-constrained HDSS (GHDSS)*. The basic formulation of these BSS algorithms is defined in Eqs. (2) – (4). The differences between them are the definitions of $J(\mathbf{W})$ and $\mathbf{J}'(\mathbf{W})$. Therefore, μ_{opt} defined in Eq. (7) changes when our adaptive step-size parameter control method is applied. The six BSS algorithms with adaptive step-size parameters are described in the following sections.

3.1. Decorrelation-based Source Separation (DSS)

The cost function of DSS is defined by

$$\begin{aligned} J_{DSS}(\mathbf{W}) &= \|E[\mathbf{E}]\|^2 \quad (8)\\ \mathbf{E} &= \mathbf{y}\mathbf{y}^H - \mathrm{diag}[\mathbf{y}\mathbf{y}^H], \end{aligned}$$

where $E[\cdot]$ represents an expectation operator. The update direction $\mathbf{J}'(\mathbf{W})$ is calculated by

$$\mathbf{J'}_{DSS}(\mathbf{W}) = 2\mathbf{E}\mathbf{W}\mathbf{x}\mathbf{x}^H \quad (9)$$

which is obtained by taking $\nabla_{w*}J(\mathbf{W})$ and removing $E[\cdot]$. The optimal step-size is

$$\mu_{opt_{DSS}} = \frac{\|\mathbf{E}\|^2}{2\|2\mathbf{E}\mathbf{W}_t\mathbf{x}\mathbf{x}^H\|^2} \quad (10)$$

3.2. Independent Component Analysis (ICA)

We selected a conventional ICA algorithm based on Kullback-Liebler divergence [2] and natural gradient method [6] for applying our proposed method. In this ICA, $J(\mathbf{W})$ and $\mathbf{J}'(\mathbf{W})$ are given by

$$\begin{aligned} J_{ICA}(\mathbf{W}) &= \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{q(\mathbf{y})} d\mathbf{y}, \quad (11)\\ \mathbf{J'}_{ICA}(\mathbf{W}) &= \mathbf{E}_\phi \mathbf{W}, \quad (12)\\ \mathbf{E}_\phi &= \phi(\mathbf{y})\mathbf{y}^H - \mathrm{diag}[\phi(\mathbf{y})\mathbf{y}^H], \end{aligned}$$

where $p(\mathbf{y})$ is the joint Probability Density Function (PDF) of $\mathbf{y}$.$q(\mathbf{y})$ is the product of the marginal PDF, i.e., $\prod_k p(y_k)$. $\phi(\mathbf{y})$ means a nonlinear function defined as

$$\begin{aligned} \phi(\mathbf{y}) &= [\phi(y_1), \phi(y_2), \cdots, \phi(y_N)]^T \quad (13)\\ \phi(y_i) &= -\frac{\partial}{\partial y_i} \log p(y_i). \end{aligned}$$

There are a variety of definitions for $\phi(y_i)$. In this paper, we selected a hyperbolic-tangent-based function [7] defined by

$$\phi(y_i) = \tanh(\eta|y_i|)e^{j\cdot\theta(y_i)}, \quad (14)$$

where η means the scaling parameter.

Since it is almost impossible to calculate J_{ICA}, we used $\|\mathbf{E}_\phi\|^2$ instead of the J_{ICA}. The optimal step-size is defined by

$$\begin{aligned} \mu_{opt_{ICA}} &= \frac{\|\mathbf{E}_\phi\|^2}{2\mathrm{MA}(\mathbf{E}_\phi\mathbf{W}_t, 2\mathbf{E}\tilde{\phi}(\mathbf{y})\mathbf{x}^H)} \quad (15)\\ \tilde{\phi}(\mathbf{y}) &= [\tilde{\phi}(y_1), \tilde{\phi}(y_2), ..., \tilde{\phi}(y_N)]^T\\ \tilde{\phi}(y_i) &= \phi(y_i) + y_i \frac{\partial \phi(y_i)}{\partial y_i}. \end{aligned}$$

3.3. Geometric-constrained Source Separation (GSS)

GSS relaxes limitations in ICA such as permutation and scaling problems by introducing "geometric constraints" obtained from the locations of microphones and sound sources. Therefore, it is suitable for real-world applications such as robot audition systems [8]. $J(\mathbf{W})$ for GSS consists of two cost functions – $J_{DSS}(\mathbf{W})$ in Eq. (8) and $J_{GC}(\mathbf{W})$ which corresponds to geometric constraints.

$$J_{GSS}(\mathbf{W}) = J_{DSS}(\mathbf{W}) + \lambda J_{GC}(\mathbf{W}) \quad (16)$$

where λ means a weight factor.

When a cost function based on delay-and-sum beamforming (C1 in [9]) is selected as $J_{GC}(\mathbf{W})$, it is denoted as

$$\begin{aligned} J_{GC}(\mathbf{W}) &= \|\mathbf{E}_{GC}\|^2 \quad (17)\\ \mathbf{E}_{GC} &= diag[\mathbf{W}\mathbf{D} - \mathbf{I}] \end{aligned}$$

where $\mathbf{D}$ means a transfer function matrix based on a direct sound path between a sound source and each microphone. $\mathbf{J}'(\mathbf{W})$ is given by

$$\begin{aligned} \mathbf{J'}_{GSS}(\mathbf{W}) &= \mathbf{J'}_{DSS}(\mathbf{W}) + \lambda \mathbf{J'}_{GC}(\mathbf{W}) \quad (18)\\ \mathbf{J'}_{GC}(\mathbf{W}) &= \mathbf{E}_{GC}\mathbf{D}^H. \end{aligned}$$

The update equation of the separation matrix for GSS is defined by

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu_{DSS}\mathbf{J'}_{DSS}(\mathbf{W_t}) - \mu_{GC}\mathbf{J'}_{GC}(\mathbf{W_t}). \quad (19)$$

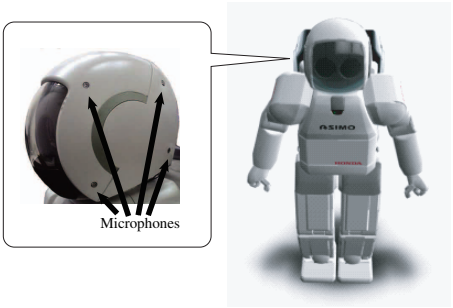

Fig. 2. ASIMO with 8 microphones

In the case where a fixed step-size parameter is used, μ_{GC} is defined by

$$\mu_{GC} = \lambda \cdot \mu_{DSS}. \tag{20}$$

In our adaptive step-size method, both μ_{DSS} and μ_{GC} are optimized. The optimal step-size for μ_{DSS} is defined as Eq. (10), and that for μ_{GC} is calculated as

$$\mu_{opt_{GC}} = \frac{\|\mathbf{E}_{GC}\|^2}{2\|2\mathbf{E}_{GC}\mathbf{D}^H\|^2} \tag{21}$$

3.4. Geometric-constrained ICA (GICA)

GICA is an ICA algorithm with geometric constraints. Thus, it is formulated by replacing $J_{DSS}(\mathbf{W})$ with $J_{ICA}(\mathbf{W})$ in Eq. (16).

$$J_{GICA}(\mathbf{W}) \quad = \quad J_{ICA}(\mathbf{W}) + \lambda J_{GC}(\mathbf{W}) \tag{22}$$

Therefore, the optimal step-size parameters for GICA are obtained from Eqs. (15) and (21). Although GICA was also reported in [10], it requires accurate geometric information to achieve good performance. Since our GICA formulation allows constraint errors to some extent, it is more suitable for real-world applications.

3.5. High-order DSS (HDSS)

$J(\mathbf{W})$ and $\mathbf{J}'(\mathbf{W})$ for HDSS are defined by

$$J_{HDSS}(\mathbf{W}) \quad = \quad \|E[\mathbf{E}_\phi]\|^2 \tag{23}$$

$$\mathbf{J}'_{HDSS}(\mathbf{W}) \quad = \quad 2\mathbf{E}_\phi \tilde{\phi}(\mathbf{y})\mathbf{x}^H \tag{24}$$

Thus, the optimal step-size for HDSS is defined by

$$\mu_{opt_{\mathrm{HDSS}}} = \frac{\|\mathbf{E}_\phi\|^2}{2\|2\mathbf{E}_\phi \tilde{\phi}(\mathbf{y})\mathbf{x}^H\|^2}. \tag{25}$$

3.6. Geometric-constrained High-order DSS (GHDSS)

GHDSS is a Geometric-constrained version of HDSS. Therefore, its cost function, $J_{GHDSS}(\mathbf{W})$ is defined by

$$J_{GHDSS}(\mathbf{W}) \quad = \quad J_{HDSS}(\mathbf{W}) + \lambda J_{GC}(\mathbf{W}). \tag{26}$$

The optimal step-size parameters for GHDSS are obtained from Eqs. (25) and (21).

4. EVALUATION

We evaluated the adaptive step-size control method through the performance of the above six BSS algorithms with/without adaptive step-size control.

We used an 8 ch microphone array embedded in Honda ASIMO shown in Fig. 2. The positions of the microphones are bilaterally symmetric. First, by using this microphone array, we measured background noise including ASIMO's own motor noises and impulse responses using a loudspeaker (GENELEC 1029A) in a room. The size of the room was 4.0 m × 7.0 m × 3.0 m, and the reverberation time (RT_{20}) was 0.3–0.4 s. The input data was, then, synthesized as a mixture of two Japanese-speech sources originating from the front direction (S_1) and 90° to the right (S_2) of ASIMO by using the measured impulse responses and background noise.

Both sources are assumed to be 1.5 m away from the robot and to have the same power. The background noise level was 10-20 dB lower than each speech source. The setting of the six BSS algorithms is described in Table 1. For BSS with a fixed step-size parameter, three kinds of μ values, i.e., 0.1, 0.01, 0.001, were used. The weight factors λ in Eqs. (16), (22) and (26) are set to $\|\mathbf{y}\mathbf{y}^H\|^{-2}$ according to [8]. Besides the six algorithms, we also evaluated two other conditions to know the baseline performance. One case was with one microphone input selected, and another case was with a simple delay-and-sum beamformer applied. Basically, BSS algorithms in the frequency domain have two problems – scaling and permutation. In this work, the permutation problems are solved by maintaining $\|\mathbf{W}\| = 1$ at every time frame [11]. The scaling problems are avoided by reordering row vectors in $\mathbf{W}$ according to geometric information on sound source directions estimated at the first time frame. Three metrics - signal-to-noise ratio (SNR), mean of correlation coefficient (CC) and word correct rate (WCR) by using automatic speech recognition (ASR) - were used for evaluation. SNR and CC were measured for 10 s speech input in all algorithms, and WCR was measured only for speech separated by GSS, because it has the best performance in SNR when our proposed method was applied.

SNR is defined by

$$SNR = 10\log_{10}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{|y|^2}{|\hat{n}|^2}\right], \tag{27}$$

where y means a separated signal (output) and $\hat{n}$ is the noise signal included in y. The n is calculated by using $\hat{n} = y - \hat{s}$, where $\hat{s}$ represents a separated signal for the signal generated by the convolution of S_i and the measured impulse response.

CC is defined in time-frequency domain as

$$CC \ [\mathrm{dB}] \quad = \quad 10\log_{10} E_\omega[CC_\omega(\omega)], \tag{28}$$

$$CC_\omega(\omega) \quad = \quad \frac{|E_t[|y_1^*(\omega,t)y_2(\omega,t)|]}{\sqrt{E_t[|y_1(\omega,t)|^2]}\cdot\sqrt{E_t[|y_2(\omega,t)|^2]}}$$

where $E_\omega[\cdot]$ and $E_t[\cdot]$ mean the average powers in frequency and time respectively. $y_i(\omega,t)$ means the i-th output signal at time t and frequency ω. Because CC represents the correlation between the two sound sources, it is expected to be $-\infty$ dB when the two speeches are separated completely.

To measure WCR, we used Japanese automatic speech recognizer, Julian[12], which supports network grammar as a language model. Isolated word recognition for an ATR phonetically-balanced Japanese word dataset which includes 216 words per speaker was performed using a clean acoustic model.

Table 1. BSS Setting

sampling frequency	16 kHz
window function	Hanning
window length	512 (32 ms)
shift length	256 (16 ms)
scaling parameter η	1

4.1. Results

Figs. 3 – 5 show SNR, CC and WCR of the separated speech, respectively.

In Fig. 3, our proposed method (AS) shows optimal SNR improvement in four BSS algorithms – GSS, DSS, HDSS, and GHDSS. This means that adaptive step-size control is effective to improve BSS. However, in the ICA and GICA algorithms, the performance of the proposed method was lower than we expected. We found that noises at low frequency were emphasized in the separated speech in ICA and GICA. This decreased SNR, but sound source separation worked well at the frequency bands which include speech signals.

In Fig. 4, AS shows the best performance in all BSS algorithms. This means that our proposed method is also effective in ICA and GICA in terms of decorrelation, that is, separation.

For WCR, AS shows the best performance in Fig. 5. Sound source separation is often used to improve ASR performance as preprocessing. Thus, this means that AS is effective for real-world applications using ASR such as robot audition systems.

5. CONCLUSION

We proposed an adaptive step-size control method to improve sound source separation in the real-world environment. It is an extension of Newton's method and is applicable to any kind of blind source separation algorithm. We implemented six types of BSS algorithms with adaptive step-size control. Through the experiments of sound source separation for two simultaneous speeches, we proved the effectiveness and the general applicability of the proposed method. Because the evaluation was performed in a simulated environment using speech data synthesized by using measured impulse responses, evaluation in dynamically-changeable acoustic environments remains as future research. Construction of a real-time robot audition system by introducing our proposed method is another future challenge.

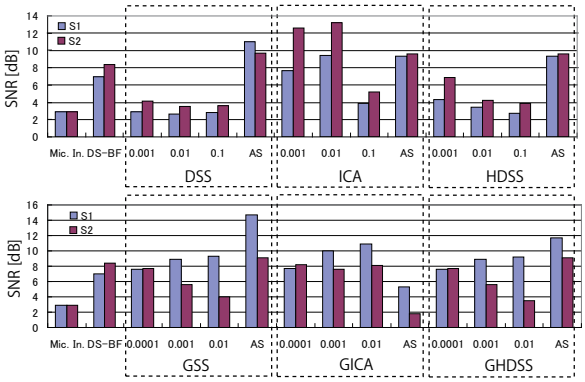

Fig. 3. Improvement in SNR for two simultaneous speeches

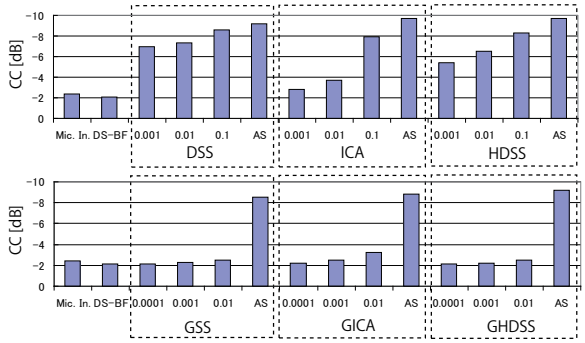

Fig. 4. Correlation Coefficient (CC)

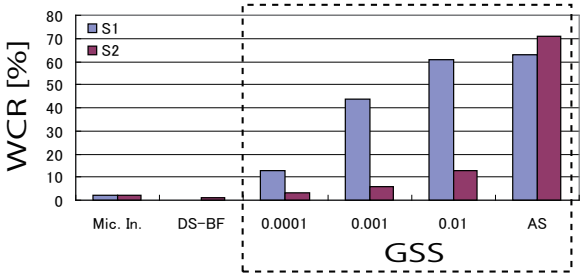

Fig. 5. Improvement in WCR of separated speech

6. REFERENCES

[1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *17th National Conf. on Artificial Intelligence (AAAI2000)*. AAAI, 2000, pp. 832–839.

[2] S. Ikeda and N. Murata, "A method of ica in time-frequency domain," *Workshop Indep. Compom. Anal. Signal.*, pp. 365–370, 1999.

[3] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ica and beamforming," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[4] S. Yamamoto and S. Kitayama, "An adaptive echo canceller with variable step gain method," *Trans. of the IECE of Japan*, vol. E 65, no. 1, pp. 1–8, 1982.

[5] D.H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc.*, vol. 130, no. 1, pp. 251–276, 1983.

[6] S. Amari, "Natural gradient works effeciently in learning," *Neural Compt.*, vol. 10, pp. 251–276, 1998.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2002, pp. 1001–1004.

[8] J. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *2004 International Conference on Intelligent Robots and Systems (IROS2004)*. IEEE/RSJ, 2004, pp. 2123–2128.

[9] L. Parra and C. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[10] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Trans. on Speech and Audio Processing*, vol. 15, no. 2, pp. 715–726, 2007.

[11] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[12] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," in *7th European Conf. on Speech Communication and Technology*, 2001, vol. 3, pp. 1691–1694.