The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Step-size Parameter Adaptation of Multi-channel Semi-blind ICA with Piecewise Linear Model for Barge-in-able Robot Audition

Ryu Takeda, Kazuhiro Nakadai, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno

Abstract— **This paper describes a step-size parameter adaptation technique of multi-channel semi-blind independent component analysis (MCSB-ICA) for a "barge-in-able" robot audition system. By "barge-in", we mean that the user can speak simultaneously when the robot is speaking. We focused on MCSB-ICA to achieve such an audition system because it can separate a user's and a robot's speech under reverberant environments. The problem with MCSB-ICA for robot audition is the slow speed of convergence in estimating a separation filter due to its step-size parameters. Many optimization methods cannot be adopted because their computational costs are proportional to the 2nd order of the reverberation time. Our method yields adaptive step-size parameters with MCSB-ICA at low computational costs. It is based on three techniques; 1) recursive expression of the separation process, 2) a piecewise linear model of the step-size of the separation filter, and 3) adaptive step-size parameters with a sub-ICA-filter. Experimental results show that our approach attains faster convergence speed and lower computational costs than those with a fixed step-size parameter.**

I. INTRODUCTION

A robot should recognize a user's speech from a mixture of sounds with the least prior information, because it has to work in unknown and / or dynamical environments. These may include the robot's own speech and a user's speech reverberations, because microphones are installed on its body, and not attached close to the user's mouth. Therefore, these should be suppressed to enhance the user's speech (Fig. 1). In human-robot and in human-computer interaction, the user often interrupts and begins speaking while the robot or the system is speaking. This situation is called "barge-in". Robot audition systems should be *"barge-in-able"* to enable smoother speech interactions.

To achieve such a barge-in-able system, we must deal with problems of echo cancellation (separation of the robot's/known speech) and blind dereverberation (separation of the user's speech reverberations) at the same time. We adopted multi-channel semi-blind independent component analysis (MCSB-ICA) [1], because: 1) it is theoretically robust against Gaussian noise, such as that from fans, 2) it can theoretically deal with separation of the known speech, user's speech, and other sound sources, including their reverberations. Other methods have not dealt with known-source signals [2], [3], [4], user's speech signals [5], or have not been able to deal with reverberation [6], [7].

R. Takeda, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno are with the Department of Intelligence Science and Technology. Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan. {rtakeda, tall, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

K. Nakadai is with Honda Research Institute Japan Co., Ltd., Wako, Saitama, 351-0114, Japan. nakadai@jp.honda-ri.com

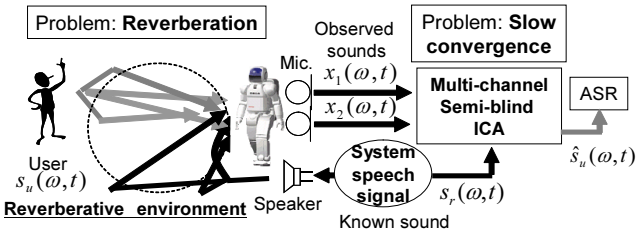

Fig. 1. Data flow and our problems

The requirements for MCSB-ICA to achieve robot audition are: a) fast convergence speed for estimating the separation filter of source signals, and b) low computational cost. ICA originally used many computational resources because it estimated the separation filter by an iterative learning, which was especially caused by a fixed step-size parameter. This parameter controlled the convergence speed, and poorly chosen step-sizes resulted in poor separation or slow convergence of the learning filter. This problem has usually been solved with a kind of the Newton methods, such as [8], [9]. However, it is difficult to apply these methods to MCSB-ICA because: 1) the MCSB-ICA model is different to standard ICA, 2) its evaluation function is highly nonlinear, and 3) the separation filter has large dimensions. These often results in increasing the computational cost.

We solved this step-size scheduling problem by adapting step-size parameters with a small sub-ICA filter. The direct implementation of this sub-ICA filter increased the computational cost. We avoided this problem by focusing on the recursive expression of the separated signals and piecewise linear step-size modeling. These approaches decreased the number of parameters to be estimated in the sub-ICA filter and reduced the computational cost.

This paper introduces three techniques we used to adapt the step-size parameters: 1) recursive expression of the separation process, 2) a piecewise linear model of the step-size of the separation filter, and 3) adaptive step-size parameters with a sub-ICA filter. Section 2 explains the MCSB-ICA and its step-size problem. Section 3 explains the three techniques in detail, and discusses our evaluations of our method in Section 4 and 5. The last section concludes the paper and discusses future work.

II. PROBLEMS WITH MULTI-CHANNEL SEMI-BLIND ICA

This section explains MCSB-ICA [1] and its problems with robot audition. Note that the MCSB-ICA model is de-

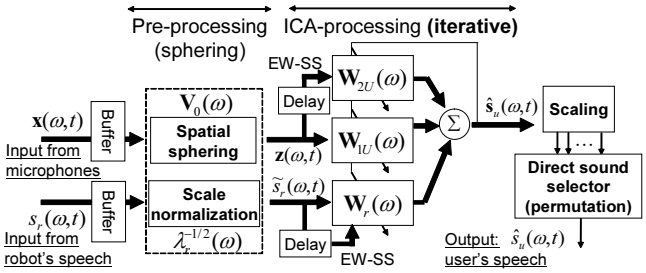

Fig. 2. Signal flow of MCSB-ICA

scribed here with a short-time Fourier transformation (STFT) representation [2] which is a form of multi-rate processing. We denote the spectrum after STFT as $s(\omega, t)$ at frequency ω and frame t. For the sake of simplicity, we have skipped denoting the frequency index, ω. Fig. 2 outlines MCSB-ICA, and we explain how the filter was estimated in this section.

A. Observation and Separation Model

We denote the observed spectra at microphones $M_1, \ldots, M_L$ as $x_1(t), \ldots, x_L(t)$ (L is the number of microphones), and its vector form as $\boldsymbol{x}(t) = [x_1(t), x_2(t), \ldots, x_L(t)]^T$. With the spectrum of user's utterance, $s_u(t)$, and known-source (robot's) spectrum, $s_r(t)$, the observed signals, $\boldsymbol{x}(t)$, can be described as the following finite impulse response (FIR) filter model:

$$\boldsymbol{x}(t) = \sum_{n=0}^{N} \boldsymbol{h}_u(n) s_u(t-n) + \sum_{m=0}^{M} \boldsymbol{h}_r(m) s_r(t-n), \quad (1)$$

where $\boldsymbol{h}_u(n)$ and $\boldsymbol{h}_r(m)$ correspond to the N- and M-dimensional FIR coefficient vectors of the user's and known-source spectra.

Before explaining the MCSB-ICA separation model, let us define the observed vector, $\boldsymbol{X}(t)$, and the known-source vector, $\boldsymbol{S}_r(t)$, as:

$$\boldsymbol{X}(t) = [\boldsymbol{x}(t), \boldsymbol{x}(t-1), \ldots, \boldsymbol{x}(t-N)]^T \text{and} \quad (2)$$
$$\boldsymbol{S}_r(t) = [s_r(t), s_r(t-1), \ldots, s_r(t-M)]^T, \quad (3)$$

The separation model for MCSB-ICA is set so that the direct sound frame of user's speech, $s_u(t)$, is independent of the delayed-observed and known sound spectra, $\boldsymbol{X}(t-d)$ and $\boldsymbol{S}_r(t)$. Here, $d(>0)$ is an initial-reflection interval parameter, and we consider the dependence between the direct and adjacent frame of $s_u(t)$. The separation model is written as:

$$\begin{pmatrix} \hat{\boldsymbol{s}}(t) \\ \boldsymbol{X}(t-d) \\ \boldsymbol{S}_r(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_{1u} & \boldsymbol{W}_{2u} & \boldsymbol{W}_r \\ \boldsymbol{0} & \boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_r \end{pmatrix} \begin{pmatrix} \boldsymbol{x}(t) \\ \boldsymbol{X}(t-d) \\ \boldsymbol{S}_r(t) \end{pmatrix}, \quad (4)$$

where $\hat{\boldsymbol{s}}(t)$ is an estimated signal vector with an L dimension, $\boldsymbol{W}_{1u}$ and $\boldsymbol{W}_{2u}$ correspond to $L \times L$ and $L \times L(N+1)$ separation matrices, and $\boldsymbol{W}_r$ is the $L \times (M+1)$ separation matrix. $\boldsymbol{I}_2$ and $\boldsymbol{I}_r$ correspond to optimally-sized unit matrices. Note that the estimated signal, $\hat{\boldsymbol{s}}(t)$, includes direct and some reflected signals of the user's speech.

B. Estimation of Filter Parameters

The filter parameter set, $\boldsymbol{W} = \{\boldsymbol{W}_{1u}, \boldsymbol{W}_{2u}, \boldsymbol{W}_r\}$, is estimated by minimizing the Kullback-Leibler divergence (KLD) between the joint Probability Density Function (PDF) and the products of the marginal PDF of $\boldsymbol{s}(t)$, $\boldsymbol{X}(t-d)$, and $\boldsymbol{S}_r(t)$. The explicit evaluation function is expressed as,

$$J(\boldsymbol{W}) = -\sum_{i=1}^{L} \mathrm{E}[\log p_i(s_i(t))] - \log|\det \boldsymbol{W}_{1u}| + H, \quad (5)$$

where $p_i(s_i)$ is a PDF of random variable s_i, $\mathrm{E}[\cdot]$ is a time-averaging operator, and H represent a joint entropy of $\{\boldsymbol{x}(t), \boldsymbol{X}(t-d), \boldsymbol{S}_r(t)\}$.

We obtain the following iterative update rules for $\boldsymbol{W}$ with a natural gradient method [10].

$$\boldsymbol{D} = \boldsymbol{\Lambda} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\hat{\boldsymbol{s}}^H(t)], \quad (6)$$
$$\boldsymbol{W}_{1u}^{[j+1]} = \boldsymbol{W}_{1u}^{[j]} + \mu \boldsymbol{D}\boldsymbol{W}_{1u}^{[j]}, \quad (7)$$
$$\boldsymbol{W}_{2u}^{[j+1]} = \boldsymbol{W}_{2u}^{[j]} + \mu\Big(\boldsymbol{D}\boldsymbol{W}_{2u}^{[j]} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\boldsymbol{X}^H(t-d)]\Big), \quad (8)$$
$$\boldsymbol{W}_{r}^{[j+1]} = \boldsymbol{W}_{r}^{[j]} + \mu\Big(\boldsymbol{D}\boldsymbol{W}_{r}^{[j]} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\boldsymbol{S}_r^H(t)]\Big), \quad (9)$$

where $\cdot^H$ denotes the conjugate transpose operation, and $\boldsymbol{\Lambda}$ is a non-holonomic constraint matrix, i.e., $\mathrm{diag}(\mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\hat{\boldsymbol{s}}^H(t)])$ [11]. The μ is a step-size parameter and $\boldsymbol{\phi}(\boldsymbol{x})$ is a non-linear function vector, $[\phi(x_1), \cdots, \phi(x_L)]^H$. $\phi(x)$, defined as

$$\phi(x) = -\frac{\mathrm{d}\log p(x)}{\mathrm{d}x}. \quad (10)$$

We assume that the source PDF is a noise-robust function $p(x) = \exp(-|x|/\sigma^2)/(2\sigma^2)$ with variance σ^2, and $\phi(x)$ becomes $x^*/(2\sigma^2|x|)$. Here, x^* denotes the conjugate of x. The two functions are defined in the continuous area, $|x| > \varepsilon$.

We use enforced spatial sphering as pre-processing, which is an approximation of sphering. We assume that the variance described above is almost 1 ($\sigma^2 \approx 1$) because sphering decorrelates the input signals and normalizes the variances [8]. The observed signal, $\boldsymbol{X}(t)$, and the known signal, $\boldsymbol{S}_r(t)$, are transformed as the following rules:

$$\boldsymbol{z}(t) = \boldsymbol{V}_u\boldsymbol{x}(t), \quad \boldsymbol{V}_u = \boldsymbol{E}_u\boldsymbol{\Lambda}_u^{-1/2}\boldsymbol{E}_u^H, \quad (11)$$
$$\tilde{s}_r(t) = \lambda_r^{-1/2} s_r(t), \quad (12)$$

where $\boldsymbol{E}_u$ and $\boldsymbol{\Lambda}_u$ are the eigenvector matrix and eigenvalue diagonal matrix of $\boldsymbol{R}_u = \mathrm{E}[\boldsymbol{x}(t)\boldsymbol{x}^H(t)]$. After sphering, $\boldsymbol{x}$ and s_r in Eqs. (4) – (9) are substituted into $\boldsymbol{z}$ and $\tilde{s}_r$.

C. Problem with MCSB-ICA: Step-size parameter

The step-size parameter in Eqs. (7)–(9) is essential because this controls the convergence speed and efficiency of separation. For example, large values result in fast adaptation with divergence and poor separation, while small values result in slow convergence and good separation.

Many step-size methods of control or fast methods of adaptation cannot be applied mainly because: of the i) high-nonlinearity and complexity of the evaluation function (Eq. (5)), the ii) the asymmetric structure of the separation model

(Eq. (4)), and the iii) the high-calculation cost proportional to the second order of the reverberation time. For example, Nakajima's method [9] requires the calculation of H in Eq. (5), and this is almost impossible. Instead of it, he set another evaluation function easy to calculate which converges the same separation filter theoretically. However, it is unclear whether our separation model is valid with the function. A fast-ICA algorithm [8] assumes perfect sphering and this increases the calculation cost, proportional to the third order of reverberation time.

We previously proposed a heuristic step-size method of scheduling combined with annealing and an exponentially-weighted step-size [12]. The step-size, μ_k, of the separation matrix at the j-th iteration and k-th delayed frame is defined by

$$\mu_k^{[j]} = \alpha\lambda^k/j + \beta, \tag{13}$$

where α, β, and λ are constant values. However, this approach needs some parameters to be set in advance. The optimum value of λ especially depends on the configurations and situation.

To accomplish fast MCSB-ICA processing for robot audition, we need to adapt the step-size parameters to the optimum one at a low computational cost.

III. Step-size Adaptation with Sub-ICA filter

This section explains the adaptation of step-size in MCSB-ICA with a sub-ICA filter.

A. Recursive Expression of Separation Process and Adaptive Step-size Parameter Problem

Let us introduce an expression for the recursive separation process to enable computational efficiency. Before explaining the recursive expression, we rewrite Eq. (4) into

$$\hat{\boldsymbol{s}}(t) = \sum_{n=0}^{N} \boldsymbol{W}_u(n)\boldsymbol{x}(t-n) + \sum_{m=0}^{M} \boldsymbol{W}_r(m)s_r(t-m), \tag{14}$$

where $\boldsymbol{W}_u(n)$ and $\boldsymbol{W}_r(m)$ are the $L \times L$ and $L \times 1$ separation matrix, respectively. Since we have omitted the range $(1 \le n < d)$ of $\boldsymbol{W}_u(n)$ for simplicity, we assume that $\boldsymbol{W}_u(n)$ equals 0 in that range. Note that $\boldsymbol{W}_{1u}$, $\boldsymbol{W}_{2u}$, and $\boldsymbol{W}_r$ correspond to $\boldsymbol{W}_u(0)$, $[\boldsymbol{W}_u(d), ..., \boldsymbol{W}_u(N)]$, and $[\boldsymbol{W}_r(0), ..., \boldsymbol{W}_r(M)]$. In addition, Eqs. (7)–(9) can be rewritten as, $\boldsymbol{W}_x^{[j+1]} = \boldsymbol{W}_x^{[j]} + \mu\Delta\boldsymbol{W}_x^{[j]}$, by using the incremental symbol, $\Delta\boldsymbol{W}_x$.

These notations and Eq. (14) yield the following recursive expression of the estimated signal $\hat{\boldsymbol{s}}(t)$,

$$\hat{\boldsymbol{s}}^{[j+1]}(t) = \sum_{n=0}^{N} \left(\boldsymbol{W}_u^{[j]}(n) + \mu\Delta\boldsymbol{W}_u^{[j]}(n)\right)\boldsymbol{x}(t-n) + \sum_{m=0}^{M} \left(\boldsymbol{W}_r^{[j]}(m) + \mu\Delta\boldsymbol{W}_r^{[j]}(m)\right)s_r(t-m) \tag{15}$$

$$= \hat{\boldsymbol{s}}^{[j]}(t) + \sum_{n=0}^{N} \mu\boldsymbol{y}_u^{[j]}(t-n) + \sum_{m=0}^{M} \mu\boldsymbol{y}_r^{[j]}(t-m), \tag{16}$$

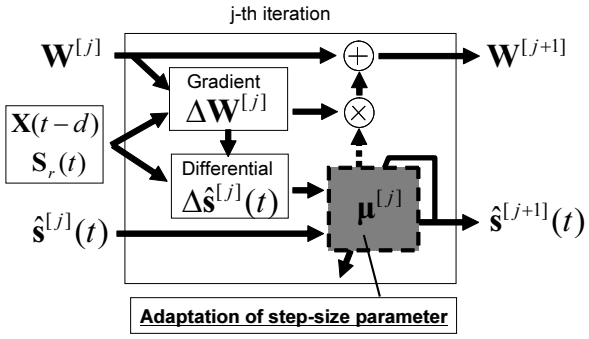

Fig. 3. Relation between parameters and signal flow with our method

where $\boldsymbol{y}_u^{[j]}(t-n) = \Delta\boldsymbol{W}_u^{[j]}(n)\boldsymbol{x}(t-n)$ and $\boldsymbol{y}_r^{[j]}(t-m) = \Delta\boldsymbol{W}_r^{[j]}(m)s_r(t-m)$. To generalize the problem of estimating the step-size, we modify the step-size, μ, to differ with each source $\hat{s}_i^{[j+1]}$, the delayed frames m and n, and each iteration, j. With the frame-variant $L \times L$ diagonal step-size matrices $\boldsymbol{\mu}_r^{[j]}(n) = \mathrm{diag}(\mu_{1,u}^{[j]}(n), ..., \mu_{L,u}^{[j]}(n))$ and $\boldsymbol{\mu}_u^{[j]}(m) = \mathrm{diag}(\mu_{1,r}^{[j]}(m), ..., \mu_{L,r}^{[j]}(m))$, Eq. (16) can be rewritten as:

$$\hat{\boldsymbol{s}}^{[j+1]}(t) = \hat{\boldsymbol{s}}^{[j]}(t) + \sum_{n=0}^{N} \boldsymbol{\mu}_u^{[j]}(n)\boldsymbol{y}_u(t-n) + \sum_{m=0}^{M} \boldsymbol{\mu}_r^{[j]}(m)\boldsymbol{y}_r(t-m). \tag{17}$$

The update rules for the separation matrices also change as to

$$\boldsymbol{W}_x^{[j+1]}(n) = \boldsymbol{W}_x^{[j+1]}(n) + \boldsymbol{\mu}_x^{[j]}(n)\Delta\boldsymbol{W}_x^{[j]}(n). \tag{18}$$

The optimum step-size parameters, $\boldsymbol{\mu}^{[j]} = \{\boldsymbol{\mu}_u^{[j]}, \boldsymbol{\mu}_r^{[j]}\}$, are estimated by minimizing the following evaluation function.

$$J(\boldsymbol{\mu}^{[j]}) = J(\boldsymbol{W}^{[j+1]}) - J(\boldsymbol{W}^{[j]}) \propto J(\boldsymbol{W}^{[j+1]}) \tag{19}$$

$$= -\sum_{i=1}^{L} \mathrm{E}[\log p_i(s_i^{[j+1]}(t))] - \log|\det \boldsymbol{W}_u^{[j+1]}(0)| \tag{20}$$

$$\approx -\sum_{i=1}^{L} \mathrm{E}[\log p_i(s_i^{[j+1]}(t))] \tag{21}$$

Here, we assumed that $\log|\det \boldsymbol{W}_u^{[j+1]}(0)|$ is very small compared with the 1st term in Eq. (20) because of the effect of spatial sphering [8]. This approximation enables us to minimize each $-\mathrm{E}[\log p_i(s_i^{[j+1]}(t))]$ independently instead of minimizing $J(\boldsymbol{\mu})$. However, as Eq. (21) has many parameters to be estimated, this results in a high computational cost. We have to reduce the number of parameters by setting some assumptions or models about $\boldsymbol{\mu}^{[j]}$. After this, we will focus on the i-th estimated source, $\hat{s}_i$.

B. Piecewise Linear Modeling of Step-size

Let us assume that the i-th element's step-size parameters for the observed term, $\mu_{i,u}^{[j]}(n)$, and the known source term, $\mu_{i,u}^{[j]}(m)$, are almost the same in terms of the number of

delayed frames, i.e., $\mu_i^{[j]}(n) = \mu_{i,u}^{[j]}(n) = \mu_{i,r}^{[j]}(n)$. By denoting the i-th element of $\boldsymbol{y}_u^{[j]}$ and $\boldsymbol{y}_u^{[j]}$ as $y_{i,u}^{[j]}$ and $y_{i,r}^{[j]}$, respectively, Eq.(17) can be reduced to

$$\hat{s}_i^{[j+1]}(t) = \hat{s}_i^{[j]} + \sum_{n=0}^{N} \mu_i^{[j]}(n)\Big(y_{i,u}^{[j]}(t-n) + y_{i,r}^{[j]}(t-n)\Big) \quad (22)$$

$$= \hat{s}_i^{[j]} + \sum_{n=0}^{N} \mu_i^{[j]}(n) y_i^{[j]}(t-n), \quad (23)$$

where $y_i^{[j]}(t-n) = y_{i,u}^{[j]}(t-n) + y_{i,r}^{[j]}(t-n)$.

We approximate $\mu_i(n)$ as a piecewise linear function divided into P ranges, and we define it as

$$\mu_i^{[j]}(n) = \begin{cases} \mu_{i,p}^{[j]} + \frac{n-B_p}{B_{p+1}-B_p}(\mu_{i,p+1}^{[j]} - \mu_{i,p}^{[j]}), \\ \qquad\qquad (B_p \le n < B_{p+1}) \\ 0, \qquad\qquad \text{otherwise} \end{cases} \quad (24)$$

where $\mu_{i,p}^{[j]}$ and $\mu_{i,p+1}^{[j]}$ are the edge points in the p-th range $[B_p \;\; B_{p+1})$ of n. We assume that the edge of final piecewise P is 0, that is, $\mu_{i,P}^{[j]} = 0$. In this model, the number of parameters to be estimated reduces into P. This model is illustrated in Fig. 4.

With this piecewise linear model, the 2nd term in Eq.(23) is rewritten as

$$\sum_{n=0}^{N} \mu_i^{[j]}(n) y_i^{[j]}(t-n) = \sum_{p=0}^{P-1} \sum_{n=B_p}^{B_{p+1}-1} \left(\mu_{i,p}^{[j]} + \frac{n-B_p}{B_{p+1}-B_p}(\mu_{i,p+1}^{[j]} - \mu_{i,p}^{[j]})\right) y_i^{[j]}(t-n). \quad (25)$$

By summarizing the terms for $\mu_{i,p}^{[j]}$ and expressing its coefficient as $Y_i(t,p)$, we modify the equation as

$$\hat{s}_i^{[j+1]}(t) = \hat{s}_i^{[j]} + \sum_{p=0}^{P-1} \mu_{i,p}^{[j]} Y_i(t,p), \quad (26)$$

or the matrix representation:

$$\begin{pmatrix} \hat{s}_i^{[j+1]}(t) \\ \boldsymbol{Y}_i(t) \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{\mu}_i^{[j]T} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \hat{s}_i^{[j]}(t) \\ \boldsymbol{Y}_i(t) \end{pmatrix}, \quad (27)$$

$$\boldsymbol{\mu}_i^{[j]} = [\mu_{i,0}^{[j]}, ..., \mu_{i,P-1}^{[j]}]^T, \quad (28)$$

$$\boldsymbol{Y}_i(t) = [Y_i(t,0), ..., Y_i(t,P-1)]^T. \quad (29)$$

Since we use the same evaluation function, J, of MCSB-ICA, the adaptive step-size problem equals the semi-blind ICA problem with the P dimensions. The calculation cost of this adaptation is not high because P is much smaller than that of Eq. (4). Note that a Weiner filter is estimated definitely as an optimal filter in terms of the second order statistics with a Gaussian approximation of $\phi(x)$. However, this filter causes a few performance degradation.

C. Adaptation of Step-size with Sub-ICA Filter

The update rules for $\boldsymbol{\mu}_i^{[j]}$ from the l-th iteration to $(l+1)$-th are

$$\boldsymbol{\mu}_i^{[j,l+1]} = \boldsymbol{\mu}_i^{[j,l]} - \gamma \mathrm{E}[\mathrm{Re}[\phi(\hat{s}_i^{[j+1]})\boldsymbol{Y}_i(t)]], \quad (30)$$

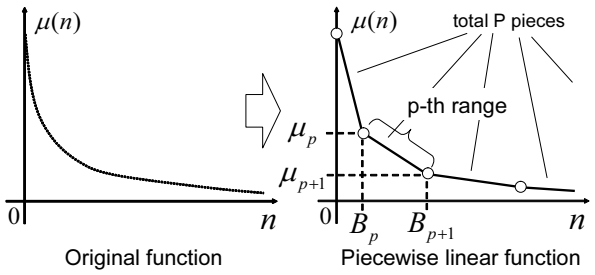

Fig. 4. Piecewise linear approximation of $\mu(n)$

where γ is a step-size parameter, and $\mathrm{Re}[x]$ denotes the real part of x. We need to set two parameters, γ and P in this adaptation.

We can use many techniques to accelerate the convergence speed, such as KL-transformation. Here, we use the sphering of $\boldsymbol{Y}_i(t)$ and it is transformed by the next rule:

$$\boldsymbol{P}_i(t) = \boldsymbol{V}_i \boldsymbol{Y}_i(t), \quad (31)$$

$$\boldsymbol{V}_i = \boldsymbol{E}_i \boldsymbol{\Lambda}_i^{-1/2} \boldsymbol{E}_i^T, \quad (32)$$

where $\boldsymbol{E}_i$ and $\boldsymbol{\Lambda}_i$ are the eigenvector matrix and eigenvalue diagonal matrix of $\boldsymbol{R}_i = \mathrm{E}[\mathrm{Re}[\boldsymbol{Y}_i(t)\boldsymbol{Y}_i^H(t)]]$. All $\boldsymbol{Y}_i(t)$ in Eqs. (27)–(30) are substituted into $\boldsymbol{P}_i(t)$. Here, the step-size in Eq. (18) is modified to $\boldsymbol{\mu}_i^{[j]}\boldsymbol{V}_i$.

In practice, we do not wait until $\boldsymbol{\mu}_i^{[j]}$ converges at the j-th iteration of $\boldsymbol{W}$ and we can stop the iteration of $\boldsymbol{\mu}_i^{[j]}$ at q times. This is because this adaptation is one part of estimating $\boldsymbol{W}$, and we can reuse $\boldsymbol{\mu}_i^{[j]}$ as the initial value of $\boldsymbol{\mu}_i^{[j+1]}$ at the $(j+1)$-th iteration. To schedule step-size parameter γ at the j-th iteration, we employ the annealing method as,

$$\gamma^{[j]} = \alpha_\gamma / j + \beta_\gamma, \quad (33)$$

where α_γ and β_γ are constant parameters.

D. Comparison of Calculation Costs

Our adaptive method needs $O(LP^2)$ at each iteration of $\boldsymbol{W}$. If it takes p_1 iterations for estimating $\boldsymbol{W}$ and q_1 times for $\boldsymbol{\mu}_i$, the total cost becomes $p_1(L^2(N+M) + q_1(LP^2))$. The processing time ratio (PTR) between our method and the standard method with p_2 iterations can be roughly evaluated as

$$\mathrm{PTR} \propto \frac{p_1(L^2(N+M) + q_1(LP^2))}{p_2(L^2(N+M))} \quad (34)$$

$$= \left(1 + \frac{q_1 P^2}{L(N+M)}\right)\frac{p_1}{p_2}. \quad (35)$$

These equations mean that our method is efficient with large N and M with many microphones, L. However, our method is not efficient with short filter lengths and small numbers of microphones. Whether our method works efficiently or not depends on the trade-off between the cost of adaptation and the scale of the system.

TABLE I

CONFIGURATION FOR DATA AND SEPARATION

Impulse response	16 kHz sampling
Reverberation time (RT_{20})	240 msec and 670 msec
Distance and direction	1.5 m and 0°, 45°, 90°, -45°, -90°
Number of microphones	Two (embedded in ASIMO's head)
STFT analysis	Hanning: 64 msec and shift: 24 msec
Input wave data	[-1.0 1.0] normalized

TABLE II

CONFIGURATION FOR SPEECH RECOGNITION

Test set	200 sentences
Training set	200 people (150 sentences each)
Acoustic model	PTM-Triphone: 3-state, HMM
Language model	Statistical, vocabulary size of 20 k
Speech analysis	Hanning: 32 msec and shift: 10 msec
Features	MFCC 25 dim. (12+Δ12+ΔPow)

IV. EXPERIMENTS

A. Experimental Settings

The impulse responses for speech data were recorded at 16 kHz in two different rooms,

Env. I): A normal room (RT_{20}=240 msec), and
Env. II): A hall-like room (RT_{20}=670 msec).

Here, RT_{20} means the reverberation time. The first room was 4.2 m $\times$ 7.0 m and the second was 7.55 m $\times$ 9.55 m. The speaker was 1.5 m apart from a microphone mounted on the head of Honda ASIMO, and the angles between the speaker and the front of ASIMO were five patterns of 0, 45, 90, -45, -90 degrees. We also recorded the impulse response from the robot's speech in each environment. All data (16 bits, PCM) were normalized to $[-1.0 \;\; 1.0]$.

We used 200 Japanese sentences for the user's and robot's speech, and they were convoluted in the corresponding recorded impulse responses. Julius[1] was used for HMM-based ASR with the statistical language model. Mel-frequency cepstral coefficients (MFCC) (12+Δ12+ΔPow) were obtained after STFT with a window size of 512 points and a shift size of 160 points for the speech features, and we then applied Cepstral Mean Normalization. A triphone-based acoustic model (three-state and four-mixture) was trained with 150 sentences of clean speech uttered by 200 male and female speakers (word-closed). The statistical language model consisted of 20,000 words, which were extracted from newspapers. The other experimental conditions are summarized in Tables I and II.

B. Evaluation Criteria

We carried out two experiments in each environments.

Exp. A): Dereverberation efficiency, and
Exp. B): Dereverberation and echo cancellation efficiency.

Word correctness (WC) and the number of iteration were evaluated. Note that the sounds only included the user's speech in Exp. A (**non-barge-in**), and they include the user's and robot's speech in Exp. B (**barge-in**). All data were used to estimate the matrices $\boldsymbol{W}_{1u}$, $\boldsymbol{W}_{2u}$ and $\boldsymbol{W}_r$ (batch-processing). Two microphones were used in these experiments. We also evaluated the PTR in both experiments.

C. Compared Methods and Separation Parameters

The same parameters for STFT were chosen, and the window size was $1,024$ points (64 msec) which is sub-optimal size [7], and its shift-size was 384 points (24 msec). The frame interval parameter d is 2, and the filter length, $N = M$, was 9 in Env. I and 23 in Env. II. The scaling and permutation problems are solved by the method mentioned in [1].

We compared three methods under all experimental conditions:

1) Our method: Ours,
2) Fixed one step-size: FIX, and
3) Annealing-based step-size: AN (Eq. (13)).

The parameters for our method were $\alpha_\gamma = 0.2$, and $\beta_\gamma = 5.0e^{-3}$. Dimension P of the piecewise linear function were 2, 1, and 0. We set $(B_0, B_1, B_2) = (0, 4, N)$ for $P = 2$, and $(B_0, B_1) = (0, N)$ for $P = 1$. Here, $P = 0$ means that we use the same adapted step-size for every $\mu(n)$. The maximum number of iteration of the sub-ICA filter was set to three. Three kinds of exponential parameters were used for the annealing-based method, $\lambda = 1.0$, 0.9, and 0.8. The other common parameters for this method were $\alpha = 6.0e^{-1}$ and $\beta = 5.0e^{-3}$. We tried three parameters for fixed step-sizes of, $\mu = 0.1$, 0.05, and 0.01.

V. RESULTS

Figures 5 and 6 present the position-averaged results for Exp. A, and Figs. 7 and 8 present those for Exp. B. The WC of clean speech is about 93%.

We can see that a large step-size (FIX2: $\mu = 1.0E^{-1}$) converges fast with the fixed step-size method, and performance is worse than that with a small step-size (FIX1: $\mu = 5.0E^{-2}$). However, FIX1 converges slowly and this means that a fixed step-size can not achieve the separation performance and convergence speed. We skip the result of $\mu = 1.0E^{-2}$ because its convergence speed is too slow.

The annealing methods outperformed the fixed step-size methods in all experiments. We showed only the result of $\lambda = 0.9$ (AN) because its averaged performance is better than that of others.

Our method, especially $P = 2$ and $P = 1$, outperformed all other methods in almost all situations. Since the result of $P = 0$ means the performance with one adapted step-size, we showed the importance both of the adaptation and the piece-wise linear model of the step-size.

Table III lists the PTR in Exp. A and Exp. B. As mentioned in Section III-D, PTR improved from 2.0 to 1.4 in the environment with a long reverberation and the known source signal. This means that our method should improve WC with half or two thirds the number of iterations of the other methods. This seemed to be achieved in Fig. 6–8 because the "Cross-point" in Figs. represents the number of iterations which our method needs to perform as well as other

[1]http://julius.sourceforge.jp/

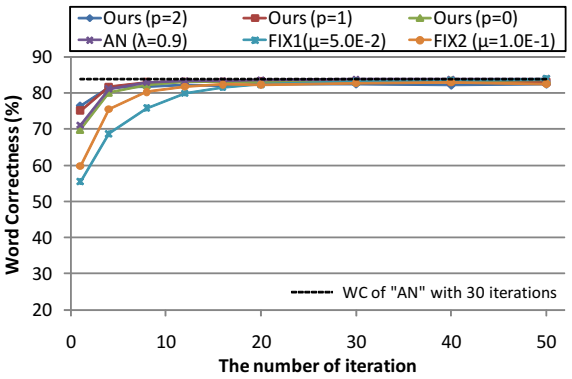

Fig. 5. Results of Exp.A in Env.I

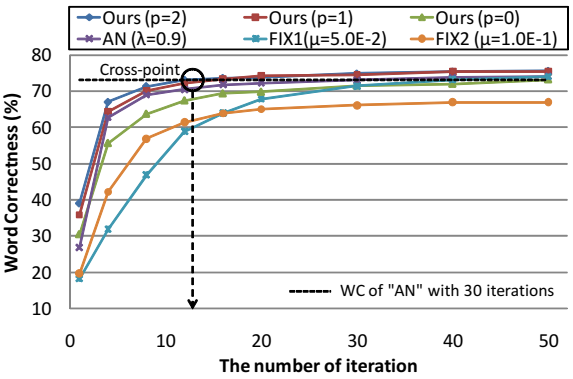

Fig. 6. Results of Exp.A in Env.II

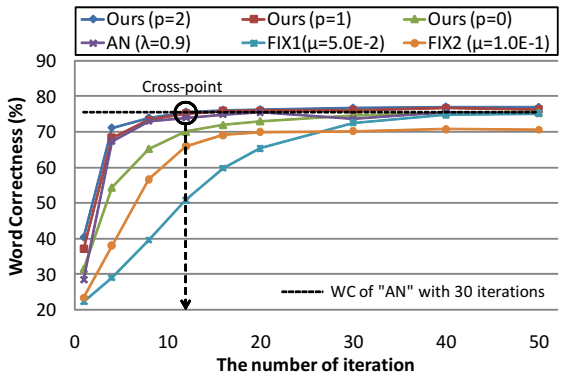

Fig. 7. Results of Exp.B in Env.I

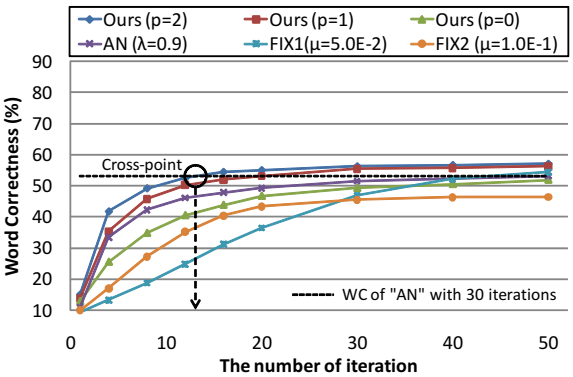

Fig. 8. Results of Exp.B in Env.II

TABLE III
PROCESSING TIME RATIO (PTR)

Configuration	PTR	inverse PTR
Exp.A, N=9	1.983	0.504
Exp.A, N=23	1.564	0.609
Exp.B, N=9	1.727	0.579
Exp.B, N=23	1.406	0.711

method with 30 iterations. In addition, the real-time factor ($\frac{\text{processing time}}{\text{data duration}}$) with 16 iterations in Exp. B and Env. II was less than 1.0 in our environment.

We concluded that our method accomplished the adaptation of step-sizes and the computational efficiency in almost all situations, excepting in the low reverberation situation without a known source.

VI. CONCLUSION AND FUTURE WORK

We developed a robot audition system that enabled barge-in for smooth speech interaction. To speed up the convergence of MCSB-ICA, we used three techniques: 1) recursive expression of the separation process, 2) a piecewise linear model of the step-size of the separation filter, and 3) adaptive step-size parameter with a sub-ICA filter. The experimental results demonstrated the effectiveness of our methods.

In the future, we intend to work on an incremental separation scheme of MCSB-ICA because this ICA includes delay time signals unlike conventional frequency-domain ICA. We also intend to evaluate MCSB-ICA as a blind source separation method. We also need to integrate MCSB-ICA with other methods to enable real-time processing for robot audition.

VII. ACKNOWLEDGMENTS

This research is partially supported by the Global COE Program and the Grant-in-Aid for Scientific Research (S).

REFERENCES

[1] R. Takeda *et al.*, "ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition," in *Proc. of ICASSP09*, 2009, pp. 3677–2633.

[2] T. Nakatani *et al.*, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *ICASSP08*. 2008, pp. 85–88, IEEE.

[3] R. Gomez *et al.*, "Distant-talking robust speech recognition using late reflection components of room impulse response," in *ICASSP08*. 2008, pp. 4581–4584, IEEE.

[4] T. Yoshioka *et al.*, "An integrated method for blind separation and dereverberation of convolutive audio mixtures," in *EUSIPCO08*, 2008.

[5] J.-M. Yang *et al.*, "A new adaptive filter algorithm for system identification using independent component analysis," in *ICASSP07*. 2007, pp. 1341–1344, IEEE.

[6] S. Miyabe *et al.*, "Barge-in- and noise-free spoken dialogue interface based on sound field control and semi-blind source separation," in *EUSIPCO07*, 2007, pp. 232–236.

[7] S. Araki *et al.*, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. on Speech & Audio Proc.*, vol. 11, pp. 109–116, 2003.

[8] A. Hyvarinen *et al.*, *Independent Component Analysis*, Wiley-Interscience, 2001.

[9] H. Nakajima *et al.*, "Adaptive step-size parameter control for real-world blind source separation," in *Proc. of ICASSP*, 2008, pp. 149–152.

[10] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[11] S. Choi *et al.*, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in *Int'l Workshop on ICA and BBS*, 1999, pp. 371–376.

[12] S. Makino *et al.*, "Exponentially weighted stepsize nlms adaptive filter based on the statistics of a room impulse response," *IEEE Trans. on Speech & Audio Proc.*, vol. 1, no. 1, pp. 101–108, 1993.