SOUND SOURCE SEPARATION OF MOVING SPEAKERS FOR ROBOT AUDITION

Kazuhiro Nakadai[†,‡], *Hirofumi Nakajima*[†], *Yuji Hasegawa*[†], *and Hiroshi Tsujino*[†]

† Honda Research Institute Japan Co., Ltd
8-1 Honcho, Wako-shi, Saitama, 351-0188 Japan
‡ Tokyo Institute of Technology

ABSTRACT

This paper addresses sound source separation and speech recognition for moving sound sources. Real-world applications such as robots should cope with both moving and stationary sound sources. However, most studies assume only stationary sound sources. We introduce two key techniques to cope with moving sources, that is, *Adaptive Step-size control (AS)* and *Optima Controlled Recursive Average (OCRA)* to improve blind source separation. We implemented a real-time robot audition system with these techniques for our humanoid robot ASIMO with an 8ch microphone array by using HARK which is our open-source software for robot audition. The performance of the system will be shown through sound source separation for moving sources and automatic speech recognition of separated speeches.

Index Terms— robot audition, sound source separation, speech recognition, adaptive step-size control, optima controlled recursive average, moving sound sources

1. INTRODUCTION

"Robot Audition" is important for natural human-robot interaction [1]. A robot should recognize users' voices without using any additional equipment such as a headset or microphone attached to the users. Actually, various robot audition systems have been reported [2, 3, 4]. However, in the real world, environmental factors such as source positions, background noises and room reverberations change dynamically. Since most parameters in their robot audition systems were tuned to a static environment, it is still difficult to recognize a user's voice in a dynamically-changing environment. To solve this problem, the parameters should be automatically adapted to the environment, because optimal values of the parameters are dynamically changing. We reported an adaptive step-size (AS) control method[5] for general Blind Source Separation (BSS) methods to achieve automatic adaptation of optimal parameters, and showed its effectiveness in a simulated environment. This paper proposes an optima controlled recursive average (OCRA) method as another effective method to cope with dynamically-changing environments. In addition, we applied these two methods to a real-time frequency-domain BSS method, online Geometric Source Separation (online GSS) [6] We, then, show its effectiveness through sound source separation experiments for moving sources by using a real robot called Honda ASIMO which was modified to have an 8ch microphone array in its head.

2. FORMULATION OF ONLINE GEOMETRIC SOURCE SEPARATION AND ITS ISSUES

Online GSS [6] is promising as one of the adaptive FD-BSS algorithms for robot audition, because it requires a smaller calculation cost than the other BSS algorithms. We, thus, focused on online GSS. Suppose that there are *M* sources and *N* ($\geq$ *M*) microphones. A spectrum vector of *M* sources at frequency ω, $\mathbf{s}(\omega)$, is denoted as $[s_1(\omega)\ s_2(\omega)\ \ldots\ s_M(\omega)]^T$, and a spectrum vector of signals captured by the *N* microphones at frequency ω, $\mathbf{x}(\omega)$, is denoted as $[x_1(\omega)\ x_2(\omega)\ \ldots\ x_N(\omega)]^T$. The source separation is then formulated as

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega), \tag{1}$$

where $\mathbf{W}(\omega)$ is called a *separation matrix*. The separation with the general FD-BSS is defined as finding $\mathbf{W}(\omega)$ which satisfies the condition that output signal $\mathbf{y}(\omega)$ is the same as $\mathbf{s}(\omega)$. In order to estimate $\mathbf{W}(\omega)$, GSS introduces two cost functions, that is, separation sharpness (J_{SS}) and geometric constraints (J_{GC}) defined by

$$\begin{aligned} J_{SS}(\mathbf{W}) &= \|E[\mathbf{E}_{SS}]\|^2 & (2)\\ \mathbf{E}_{SS} &= \mathbf{y}\mathbf{y}^H - \mathrm{diag}[\mathbf{y}\mathbf{y}^H], \\ J_{GC}(\mathbf{W}) &= \|\mathbf{E}_{GC}\|^2 & (3)\\ \mathbf{E}_{GC} &= \mathrm{diag}[\mathbf{W}\mathbf{D} - \mathbf{I}], \end{aligned}$$

where $\|\cdot\|^2$ indicates the Frobenius norm, $\mathrm{diag}[\cdot]$ is the diagonal operator, $E[\cdot]$ represents the expectation operator and H represents the conjugate transpose operator. $\mathbf{D}$ means a transfer function matrix based on a direct sound path between a sound source and each microphone. The total cost function $J(\mathbf{W})$ is represented as

$$J(\mathbf{W}) = \alpha_S J_{SS}(\mathbf{W}) + J_{GC}(\mathbf{W}), \tag{4}$$

where α_S means the weight parameter that controls the weight between the separation cost and the cost of the geometric constraint. This parameter is usually set to $\|\mathbf{x}^H\mathbf{x}\|^{-2}$ according to [6].

In addition, the online GSS updates $\mathbf{W}$ by minimizing $J(\mathbf{W})$ so that the optimal separation matrix $\mathbf{W}_{opt}$ can be obtained by using

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu\mathbf{J}'(\mathbf{W_t}). \tag{5}$$

where $\mathbf{W}_t$ denotes $\mathbf{W}$ at the current time step t, $\mathbf{J}'(\mathbf{W})$ is defined as an update direction of $\mathbf{W}$, and μ means a step-size parameter. $\mathbf{J}'(\mathbf{W})$ is derived from its complex gradient [8].

$$\begin{aligned} \mathbf{J}'(\mathbf{W}) &= \alpha_S\mathbf{J}'_{SS}(\mathbf{W}) + \mathbf{J}'_{GC}(\mathbf{W}) & (6)\\ \mathbf{J}'_{SS}(\mathbf{W}) &= 2\mathbf{E}_{SS}\mathbf{W}\mathbf{x}\mathbf{x}^H \\ \mathbf{J}'_{GC}(\mathbf{W}) &= \mathbf{E}_{GC}\mathbf{D}^H. \end{aligned}$$

However, the online GSS has two issues in a dynamically-changing environment. One is that the step-size parameter μ and the weight parameter α_S are fixed values decided heuristically or empirically, although they should be frequency-dependent and time-variant values according to environmental changes. The other is the

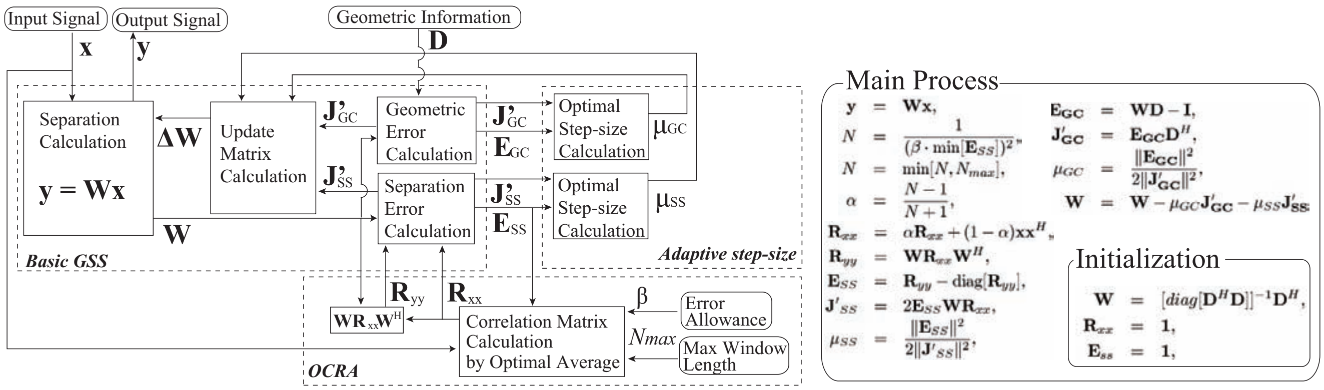


Fig. 1. Diagram of GSS with the proposed methods

calculation error of $\mathbf{J}'_{SS}(\mathbf{W})$ in Eq. (6). The online GSS omitted the expectation operation and just used an instantaneous product for incremental processing. This disturbs the convergence of $\mathbf{W}$, and $\mathbf{W}$ perturbs around the optimal value $\mathbf{W}_{opt}$. For the first issue, we already reported an *Adaptive Step-size (AS)* method [5] which controls both μ and α_S optimally, and showed its effectiveness in simulated environments. For the second issue, we propose a new method called *Optima Controlled Recursive Average (OCRA)* which makes the convergence of the separation matrix smoother, and improves the separation performance of stationary states.

2.1. Adaptive step-size control

AS is well-studied in the field of echo cancellation [7]. We extended AS to support multi-channel input and complex number signals for FD-BSS by using the multi-dimensional version of Newton's method and linear approximation formula for a complex gradient matrix[5]. By using our AS, Eqs. (5) and (6) are redefined as

$$\begin{aligned}\mathbf{W}_{t+1} &= \mathbf{W}_t - \mu_{SS}\mathbf{J}'_{SS}(\mathbf{W_t}) - \mu_{GC}\mathbf{J}'_{GC}(\mathbf{W_t}), \quad (7)\\ \mu_{SS} &= \|\mathbf{E}_{SS}\|^2/(2\|2\mathbf{E}\mathbf{W}_t\mathbf{x}\mathbf{x}^H\|^2),\\ \mu_{GC} &= \|\mathbf{E}_{GC}\|^2/(2\|2\mathbf{E}_{GC}\mathbf{D}^H\|^2).\end{aligned}$$

where μ_{SS} and μ_{GCq} correspond to $\mu\alpha_S$ and μ, respectively. They become large values 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. Thus, step-size and weight paramters are controlled optimally at the same time.

3. OPTIMA CONTROLLED RECURSIVE AVERAGE

OCRA estimates a precise correlation matrix by using an adaptively controlled window. In online systems, correlation matrix $\mathbf{R}_{xx}$ at time frame t is estimated as $\hat{\mathbf{R}}_{xx}(t)$ from a partial signal of $\mathbf{x}(t)$ by using a time window $w(\cdot)$.

$$\begin{aligned}\hat{\mathbf{R}}_{xx}(t) &= w(t) * [\mathbf{x}(t)\mathbf{x}^H(t)]\\ &= \sum_{\tau=0}^{\infty} w(\tau)[\mathbf{x}(t-\tau)\mathbf{x}^H(t-\tau)]. \quad (8)\end{aligned}$$

Since rectangular and exponential windows are commonly used, we discuss their estimation errors, $w_{Rct}(\tau)$ and $w_{Exp}(\tau)$ defined as

$$w_{Rct}(\tau) = \begin{cases} 1/N & 0 \le \tau < N \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

$$w_{Exp}(\tau) = \begin{cases} (1-\alpha)\alpha^{-\tau} & 0 \le \tau \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where N is a rectangular window of length, and α is the decay parameter, which decides the equivalent window length. We use the root mean squared error $\bar{e}_{ij}$ defined as

$$\bar{e}_{ij} = \sqrt{E[|w(t) * [x_i(t)x_j^*(t)] - E[x_i(t)x_j^*(t)]|^2]}. \quad (11)$$

By assuming that $x_i(t)$ is a linear combination of Gaussian complex variables, errors for the rectangular window $\bar{e}_{ij,Rct}$ and for the exponential window, $\bar{e}_{ij,Exp}$ are calculated as

$$\begin{aligned}\bar{e}_{ij,Rct} &= 1/\sqrt{N}, \quad (12)\\ \bar{e}_{ij,Exp} &= \sqrt{(1-\alpha)/(1+\alpha)}. \quad (13)\end{aligned}$$

These equations imply that for reducing estimation errors, a long window length is required. For better estimations, the window length must be long, however, it makes the adaptation speed slow. Therefore, it is necessary to set the optimal length for required precision. Because the required precision is proportional to the separation sharpness, we propose an optimal control method for window length defined by

$$N(t) = (\beta \cdot \min[\mathbf{E_{SS}}(t)])^{-2}, \quad (14)$$

where $N(t)$ is the window length for the rectangular window, β is an allowable error parameter, and $\min[\mathbf{A}]$ represents the minimum element's value in matrix $\mathbf{A}$. To avoid an extraordinary long window, we introduced the maximum value of $N(t)$, N_{max} shown in Fig. 1, and it is set to 1,000. For β, we empirically used 0.99. The decay parameter α for the exponential window which is equivalent to the rectangular length $N(t)$ is defined as

$$\alpha(t) = (N(t)-1)/(N(t)+1). \quad (15)$$

Finally, the correlation matrix is recursively estimated by using OCRA with the exponential window defined by

$$\hat{\mathbf{R}}_{xx}(t) = \alpha\hat{\mathbf{R}}_{xx}(t-1) + (1-\alpha)\mathbf{x}\mathbf{x}^H. \quad (16)$$

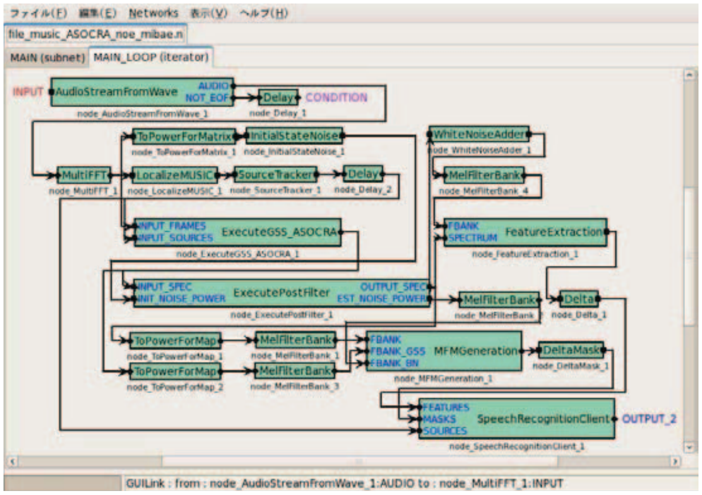


Fig. 2. HARK-based real-time robot audition system using the proposed GSS module (shown as “ExecuteGSS_ASOCRA” in the center)

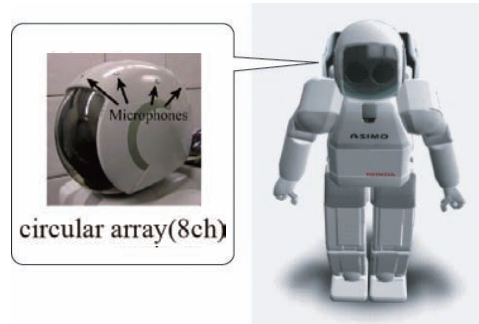


Fig. 3. An 8 ch microphone array embedded in ASIMO’s head. The microphones are circularly layouted on the head. Each microphone is omni-directional.

4. IMPLEMENTATION

4.1. GSS improved by AS and OCRA

Fig. 1 shows a diagram of GSS introducing our proposed method, that is, AS and OCRA, and its initialization and main processes. The step-size and weight parameters are adaptively controlled as μ_{SS} and μ_{GC} with AS. Our GSS uses correlation matrices $\mathbf{R}_{xx}$ and $\mathbf{R}_{yy}$ estimated by OCRA instead of using the corresponding instantaneous products.

4.2. Real-time robot audition system

We implemented GSS depicted in Fig. 1 as a new module of HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) which is our open source software for robot audition[1] [9]. HARK consists of a complete set of modules for robot audition[2] as component blocks on FlowDesigner[10][3], which works on Linux in real time. Many multi-channel sound cards are supported to build a real-time robot audition system easily. For preprocessing, sound source localization, tracking and separation are available. These preprocessing modules are able to be integrated with automatic speech recognition (ASR) based on the missing feature theory (MFT). For MFT, modules such as acoustic feature extraction for ASR, automatic missing feature mask generation, and ASR interface are prepared. Missing-feature-theory based ASR (MFT-ASR) is provided as a patch for Julius/Julian[11] which are Japanese/English open source speech recognition systems. Only MFT-ASR is implemented as a non-FlowDesigner module in HARK, but it connects with FlowDesigner by using modules from the ASR interface. Users are able to flexibly build robot audition systems by using the GUI interface. Fig. 2 shows our robot audition system using a newly-developed module for GSS with AS and OCRA.

5. EVALUATION

We evaluated our proposed GSS in terms of two points:

Ex.1 performance of our proposed GSS,

Ex.2 performance of our robot audition system with the proposed GSS.

In **Ex.1**, sound source separation of a mixture of two white noise sources was performed. One sound source was located in front of a robot. The other was a kind of moving sound source, that is, its direction alternately switched between 90° and $90^\circ + \Delta d$. Δd was one of 5°, 30°, 45°, 60°, or 90°. The timing of the direction switches was set to be one of 0.125, 0.25, 0.5, 1, 2, or 4 seconds. The distance between robot and each sound source was 1 m. The input sound was synthesized by using measured impulse responses between a sound source and a microphone array which is embedded in Honda ASIMO shown in Fig. 3. The directions for sound sources were given to GSS. The performance of online GSS (GSS), GSS with AS (GSSAS), and GSS with AS and OCRA (GSSASO) were compared. For a metric for separation, we used the mean of correlation coefficients (CC) defined in time-frequency domain as

$$CC\ \text{[dB]} = 10\log_{10} E_\omega[CC_\omega(\omega)], \tag{17}$$

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

where $E_\omega[\cdot]$ and $E_t[\cdot]$ are the average powers in frequency and time, respectively. $y_i(\omega, t)$ shows 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 sources are separated completely.

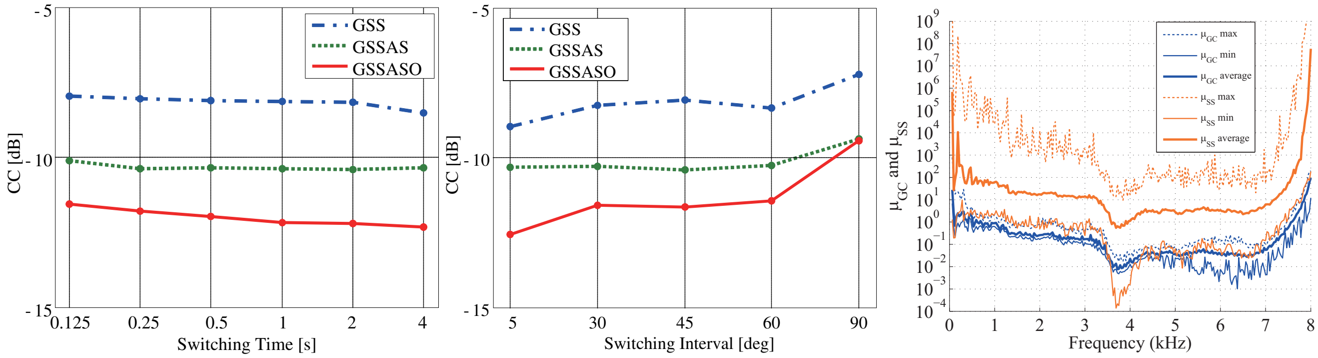
In **Ex.2**, isolated word recognition for a stationary speech source, a moving speech source and a mixture of stationary and moving speech sources was performed. The stationary speaker stands at 60° left to ASIMO in a 4.0 m × 7.0 m room with 0.3–0.4 s of RT_{20}. The moving speaker moved around ASIMO from 0° to -90°. We asked two persons to utter 236 isolated words included in ASIMO’s word database, that is, real speech data. In this case, GSS worked as a module in our robot audition system, and speaker directions estimated by a MUSIC-based sound localization module were sent to the GSS module. An acoustic model for MFT-ASR was trained with the Japanese Newspaper Article Sentences (JNAS) corpus. For a metric, we used a word correct rate.

In both experiments, the sampling rate was 16 kHz. The window length and shift length for GSS were 32 ms and 16 ms, respectively. The Hanning window was used as the window function.

[1] “HARK” has a meaning of “listen” in old English. Available at http://winnie.kuis.kyoto-u.ac.jp/HARK/.

[2] The latest HARK-0.1.7 includes 30 modules.

[3] http://flowdesigner.sourceforge.net/



a) CC vs. switching time ($\Delta d = 30°$)

b) CC vs. switching direction (switching time = 1 s)

3) ranges of μ_{GC} and μ_{SS} for GSSASO

Fig. 4. Result of **Ex.1** – Performance of GSS

Table 1. Result of **Ex.2** – word correct rate of isolated word recognition (%)

		GSS	GSSAS	GSSASO
single	stationary	95.8	95.8	95.8
	moving	90.7	90.5	90.5
double (simultaneous)	stationary	58.3	72.3	73.1
	moving	60.2	72.9	74.4

Fig. 4 shows results of **Ex.1**. Fig. 4a) shows the change of CC for switching speed. Obviously, GSSAS and GSSASO outperform online GSS, and GSSASO has the best performance. The switching speed does not affect the performance of these three methods. Fig. 4b) shows the change of CC for switching intervals. In this case, also GSSASO has the best performance. However, as the interval is larger, the performance drops. Fig. 4c) shows the maximum, minimum and average values of μ_{GC} and μ_{SS} for GSSASO in **Ex.1**. The average values show that the optimal value varies according to frequency. The average lines for μ_{GC} and μ_{SS} have similar frequency characteristics. This reflects the frequency characteristics of the impulse response used for data synthesis. From minimum and maximum values, we found that μ_{GC} changed from 10^{-1} to 10^{1} compared with the average value, and μ_{SS} changed from 10^{-2} to 10^{2}. This shows that a fixed step-size parameter is of less use for these kinds of dynamic changing situations.

Tab. 1 shows the speech recognition results in **Ex.2**. For both stationary and moving single speech sources, the three methods have the same performance in speech recognition. This shows that even online GSS has the capability to deal with a moving source in less noisy cases, because W converged fast enough. However, in noisy cases where simultaneous speeches occur, online GSS is of less use. GSSAS and GSSASO, thus, have better performance. Recognition of the separated speech for the stationary source also improved. We guess that this is caused by the leakage from the moving source, that is, a dynamically-changing noise. GSSAS and GSSASO were able to deal with such a noise, while online GSS was not. The effect of OCRA was small in **Ex.2**. This means that the improvement in source separation by OCRA affects ASR performance less. However, it may be more effective in another situation. We will, thus, conduct further experiments such as robot's moving cases.

6. CONCLUSION

We proposed optima controlled recursive average which is applicable for general blind source separation to cope with dynamically-changing environments. It is applied to GSS with previously-proposed adaptive step-size control which also improves the separation performance in dynamically-changing environments. In addition, we developed a real-time robot audition system with the proposed GSS. Our experiments for the GSS module and the robot audition system showed that the combination of these two methods improved sound source separation and speech recognition. Our future work includes more detailed evaluation such as sentences, crossing speakers and situations with moving robots.

7. REFERENCES

[1] K. Nakadai *et al*, "Active audition for humanoid," in *17th National Conf. on Artificial Intelligence (AAAI2000)*. AAAI, 2000, pp. 832–839.

[2] K. Nakadai *et al*, "Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots," *Speech Communication*, vol. 44, pp. 97–112, 2004.

[3] S. Yamamoto *et al.*, "Design and Implementation of A Robot Audition System for Automatic Speech Recognition of Simultaneous Speech," IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-2007), IEEE, 2007.

[4] I. Hara *et al*, "Robust speech interface based on audio and video information fusion for humanoid hrp-2," *IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2004, pp. 2404–2410.

[5] H. Nakajima *et al*, "Adaptive step-size parameter control for real-world blind source separation," *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, vol. AE-P2, No. 2375.

[6] J. Valin *et al*, "Enhanced robot audition based on microphone array source separation with post-filter," *IEEE Int'l Conf. on Intelligent Robots and Systems (IROS2004)*. IEEE/RSJ, 2004, pp. 2123–2128.

[7] 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.

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

[9] K. Nakadai *et al.*, "An Open Source Software System For Robot Audition HARK and Its Evaluation," *IEEE Humanoids 2008*. , IEEE, 2008, pp. 561-566.

[10] C. Côté *et al.*, "Reusability tools for programming mobile robots," *IEEE Int'l Conf. on Intelligent Robots and Systems (IROS2004)*. IEEE, 2004, pp. 1820–1825.

[11] T. Kawahara and A. Lee, "Free software toolkit for Japanese large vocabulary continuous speech recognition," *ICSLP-2000*, vol. 4, 2000, pp. 476–479.