DOA ESTIMATION METHOD BASED ON SPARSENESS OF SPEECH SOURCES FOR HUMAN SYMBIOTIC ROBOTS

Masahito Togami, Akio Amano, Takashi Sumiyoshi, and Yasunari Obuchi

Central Research Laboratory, Hitachi Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan

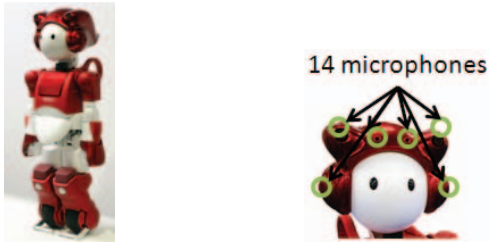
ABSTRACT

In this paper, direction of arrival (DOA) estimation methods (both azimuth and elevation) based on sparseness of human speech, "modified delay-and-sum beamformer based on sparseness (MDSBF)" and "stepwise phase difference restoration (SPIRE)", are introduced for human symbiotic robots. MDSBF can achieve good DOA estimation, whose computational cost is proportional to resolution of azimuth and elevation space. DOA estimation result of SPIRE is less accurate than that of MDSBF, but computational cost is independent of resolution. To achieve more accurate DOA estimation result than SPIRE with small computational cost, we propose a novel DOA estimation method which is combination of MDSBF and SPIRE. In the proposed method, MDSBF with rough resolution is performed prior to SPIRE execution, and SPIRE precisely estimates DOA of sources after MDSBF. Experimental results show that sparseness based methods are superior to conventional methods. The proposed combination method achieved more accurate DOA estimation result than SPIRE with smaller computational cost than MDSBF.

Index Terms— DOA estimation, microphone array, sparseness, human symbiotic robot

1. INTRODUCTION

Hitachi has been developed human symbiotic robots. EMIEW [1] is the first one which was demonstrated at Aichi Expo in Japan. The second robot of Hitachi, EMIEW2, is smaller than EMIEW so that it can move quickly with sufficient safety. The appearance of EMIEW2 is shown in Fig. 1 (a).



(a) Appearance of EMIEW2 (b) Alignment of microphone array

Fig. 1. Appearance of EMIEW2 and alignment of microphone array

The configuration of EMIEW and EMIEW2 is summarized in Table. 1. Estimation of direction of arrival (DOA) of human speech is necessary for human symbiotic robots. To turn around to the speaker direction or eye-to-eye communication are the popular applications of DOA estimation. DOA estimation is also necessary as pre-processor of the noise reduction system. Noise statistics estimation based on DOA estimation for the noise reduction system is proposed by some of the authors [2]. DOA estimation techniques have been widely studied. Many conventional techniques estimate only azimuth. However, for small robots such as EMIEW2, DOA estimation of both azimuth and elevation is necessary.

Table 1. Configuration of EMIEW and EMIEW2

	EMIEW	EMIEW2
height	1.3 m	0.8 m
weight	70 kg	13 kg
maximum speed	6 km/h	6 km/h
microphone array	8 elements mounted on shoulders and ears	14 elements mounted on head

Modified delay-and-sum beamformer based on sparseness (MDSBF) is developed by some of the authors [2] for communication robots. A speech source is known to be composed of a few frequency components at each time [3]. MDSBF can localize sparse sources such as human speech with high-accuracy. However, computational cost of MDSBF is proportional to resolution of azimuth and elevation space. Computational cost is an important factor for auditory system of robots. Preciseness of MDSBF estimation is upper-limited by computational resource on robots. Another method based on sparseness, SPIRE (Stepwise Phase dIfference REstoration), is also proposed by some of the authors [4][5]. Computational cost of SPIRE is independent of resolution. In this paper, MDSBF and SPIRE are evaluated using a microphone array mounted on the head of EMIEW2 (Fig. 1 (b)) . These microphones is utilized for DOA estimation and automatic speech recognition. To achieve correct recognition results, 14 microphones are mounted. Multichannel noise signals were recorded using EMIEW2 at a exhibition hall for low SNR evaluation. The experimental results show that MDSBF and SPIRE can estimate DOA of speech sources more correctly than the conventional methods, but SPIRE is less accurate than that of MDSBF. To achieve more accurate DOA estimation result than SPIRE, we propose a combination method of MDSBF and SPIRE. MDSBF with rough resolution is performed prior to SPIRE execution. SPIRE estimates DOA of sources around MDSBF estimation result. Experimental results show that the proposed method can achieve more accurate DOA estimation result than SPIRE with smaller computational cost than MDSBF.

2. PROBLEM STATEMENT

M is the number of the microphones. The received signals at the m-th microphone signal $x_m(t)$ is converted into time-frequency

domain signal as $x_m(f,\tau)$. f is the frequency index, and τ is the frame index of short term Fourier transform. $\boldsymbol{x}(f,\tau) = [x_1(f,\tau),\ldots,x_M(f,\tau)]$ can be modeled as follows:

$$\boldsymbol{x}(f,\tau) = \sum_{i=1}^{N_s} s_i(f,\tau)\boldsymbol{a}_i(f) + \boldsymbol{N}(f,\tau), \quad (1)$$

where N_s is the number of the sources $s_i(f,\tau)$ is the original source signal of the i-th source, $\boldsymbol{a}_i(f)$ is the steering vector of the i-th source, which is defined by DOA of the speech source, and $\boldsymbol{N}(f,\tau)$ is noise signal. Assuming that the speech source is sufficiently far from the microphones, the steering vector $\boldsymbol{a}_i(f)$ is defined as follows:

$$\boldsymbol{a}(f) = [\exp(j2\pi f T_1(\theta_i,\phi_i)),\ldots,\exp(j2\pi f T_M(\theta_i,\phi_i))], \quad (2)$$

where θ_i is the azimuth of the i-th source direction, ϕ_i is the elevation of the i-th source direction, $T_m(\theta_i,\phi_i)$ are the time difference between the source position (θ_i,ϕ_i) and the microphone position m.

3. DOA ESTIMATION BASED ON SPARSENESS ASSUMPTION

3.1. Modified delay-and-sum beamformer based on sparseness (MDSBF)

Assuming that the sound sources are human speech and multiple sources do not overlap each other at the same time frequency point, the input signal $\boldsymbol{x}(f,\tau)$ is approximated as follows:

$$\boldsymbol{x}(f,\tau) = s_{active}(f,\tau)\boldsymbol{a}_{active}(f) + \boldsymbol{N}(f,\tau), \quad (3)$$

where $active$ is the index of the source which is only one source at frequency f and frame τ. The volume difference of the input signal at each microphone is assumed to be normalized. Assuming that the distribution of $\boldsymbol{N}(f,\tau)$ is white Gaussian, maximum likelihood estimation of DOA of the active source (θ_m,ϕ_{ml}), is obtained as

$$\begin{aligned}
(\theta_{ml},\phi_{ml}) &= \underset{\theta,\phi}{\operatorname{argmax}} \min_{s_{\theta,\phi}(f,\tau)} P(\boldsymbol{x}(f,\tau)|s_{\theta,\phi}(f,\tau)\boldsymbol{a}_{\theta,\phi}(f)),\\
&= \underset{\theta,\phi}{\operatorname{argmax}} \min_{s_{\theta,\phi}(f,\tau)} P(\boldsymbol{N}(f,\tau)|s_{\theta,\phi}(f,\tau)\boldsymbol{a}_{\theta,\phi}(f)),\\
&= \underset{\theta,\phi}{\operatorname{argmin}} \min_{s_{\theta,\phi}(f,\tau)} |\boldsymbol{x}(f,\tau) - s_{\theta,\phi}(f,\tau)\boldsymbol{a}_{\theta,\phi}(f)|^2,\\
&= \underset{\theta,\phi}{\operatorname{argmin}} |\boldsymbol{x}(f,\tau) - \boldsymbol{a}_{\theta,\phi}(f)^*\boldsymbol{x}(f,\tau)\boldsymbol{a}_{\theta,\phi}(f)|^2,\\
&= \underset{\theta,\phi}{\operatorname{argmax}} |\boldsymbol{a}_{\theta,\phi}(f)^*\boldsymbol{x}(f,\tau)|^2.
\end{aligned}$$

Maximum Likelihood DOA estimation (θ_{ml},ϕ_{ml}) is obtained at each time-frequency point. By peak-searching for a histogram made from (θ_{ml},ϕ_{ml}) at all time-frequency points, DOA of the multiple sources can be obtained. The above DOA estimation algorithm is called modified delay-and-sum beamformer based on sparseness (MDSBF) [2]. In MDSBF, searching for whole azimuth and elevation space is necessary to estimate (θ_{ml},ϕ_{ml}) at each time-frequency point. In the viewpoint of computational cost, memory size of the steering vector $\boldsymbol{a}_{\theta,\phi}$ and processing time are proportional to the resolution of azimuth and elevation space. The closed-form solution (θ_{ml},ϕ_{ml}) without searching for whole azimuth and elevation space is required.

3.2. Stepwise phase difference restoration (SPIRE)

$G(f,\tau) = |\boldsymbol{a}_{\theta,\phi}(f)^*\boldsymbol{x}(f,\tau)|^2$ is expanded by the definition of the steering vector as follows:

$$\begin{aligned}
G(f,\tau) &= |\sum_{m=1}^{M} \exp(-j2\pi f T_m(\theta,\phi))x_m(f,\tau)|^2\\
&= \sum_{m_1=1}^{M}\sum_{m_2=1}^{M} \exp(-j2\pi f(T_{m_1}(\theta,\phi) - T_{m_2}(\theta,\phi))\frac{x_{m_1}(f,\tau)}{x_{m_2}(f,\tau)}.
\end{aligned}$$

The phase difference between the m_1-th microphone and the m_2-th microphone is defined as $\sigma_{m_1,m_2} = \frac{1}{j}\log\frac{x_{m_1}(f,\tau)}{x_{m_2}(f,\tau)}$, $j=\sqrt{-1}$, and τ_{m_1,m_2} is defined as $\tau_{m_1,m_2}(\theta,\phi) = T_{m_1}(\theta,\phi) - T_{m_2}(\theta,\phi)$, then $G(f,\tau)$ is expanded as follows:

$$\begin{aligned}
G(f,\tau) &= \sum_{m_1=1}^{M}\sum_{m_2=1}^{M} \exp(-j2\pi f\tau_{m_1,m_2}(\theta,\phi) + j\sigma_{m_1,m_2}),\\
&= M + \sum_{m_1=1}^{M}\sum_{m_2=m_1+1}^{M} 2\cos(-2\pi f\tau_{m_1,m_2}(\theta,\phi) + \sigma_{m_1,m_2}),
\end{aligned}$$

The first term of $G(f,\tau)$ is constant, and this term can be neglected to maximize $G(f,\tau)$, and $\bar{G}(f,\tau) = G(f,\tau) - M$. $\cos(x)$ can be approximated by the Taylor expansion around $x = 2\pi n$ as $\cos(x) \approx 1 + \frac{-1}{2}(x-2\pi n)^2$, and n is the arbitrary integer. p is defined as the index of the microphone pair (m_1,m_2). Therefore, $\bar{G}(f,\tau)$ can be approximated as $\bar{G}(f,\tau) \approx \sum_{p=1}^{\frac{M(M-1)}{2}} \left(2 - \frac{1}{2}(-2\pi f\tau_p(\theta,\phi) + \sigma_p + 2\pi n_p)^2\right)$. (θ_{ml},ϕ_{ml}) is obtained as

$$(\theta_{ml},\phi_{ml}) \approx \underset{(\theta,\phi)}{\operatorname{argmin}} \sum_{p=1}^{\frac{M(M-1)}{2}} (-2\pi f\tau_p(\theta,\phi) + \sigma_p + 2\pi n_p)^2. \quad (4)$$

When arbitrary integer n_p is given, the closed-form solution can be obtained. n_p is deeply related to the spatial aliasing problem. The definition range of σ_p is restricted from $-\pi$ to π. $\sigma_p + 2\pi n_p$ is regarded as the true phase difference. SPIRE estimates n_p in ascending order of distance between a microphone pair [5]. DOA estimation result of SPIRE can be obtained without searching for whole azimuth and elevation space, and allocated memory for the steering vectors is not necessary. In the viewpoint of computational cost, SPIRE is superior to MDSBF. However, SPIRE is approximation of MDSBF, and DOA estimation result of SPIRE is less accurate than that of MDSBF. To achieve more accurate DOA estimation result than SPIRE with small computational cost, combination method of MDSBF and SPIRE is proposed.

3.3. Combination method of MDSBF and SPIRE

The proposed method performs MDSBF with rough resolution prior to SPIRE execution. SPIRE searches for DOA within the limited region, which is shown in Fig. 3.3. The microphone pairs whose distances are less than $d_{mdsbf} = \frac{c}{2f_{max}\max(\sin\frac{\Delta\theta}{2}+\alpha_\theta, \sin\frac{\Delta\phi}{2}+\alpha_\phi)}$ are free from the spatial aliasing problem in the limited region. These pairs are described as the initial pairs. Let $\boldsymbol{p}_i = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ be the position vector of the i-th microphone. Here, b_1 and b_2 are the microphone indexes of the b-th microphone pair: $\boldsymbol{d}_b = \boldsymbol{p}_{b_1} - \boldsymbol{p}_{b_2}$. Let the distance matrix $\boldsymbol{D}$ be $[\boldsymbol{d}_1,\cdots,\boldsymbol{d}_K]^T$ (T: transpose of a matrix),

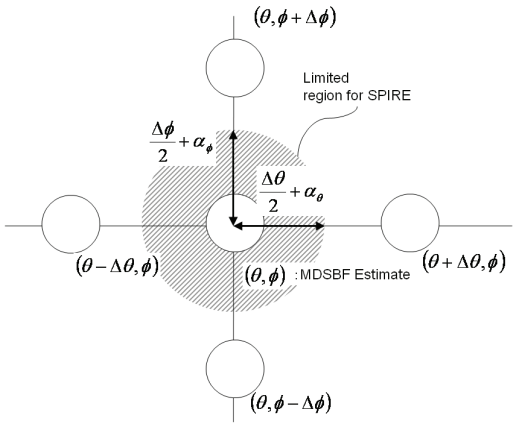


Fig. 2. Limited region by MDSBF

and $\boldsymbol{D}^+$ be the Moore and Penrose generalized inverse matrix of $\boldsymbol{D}$, where K is the number of microphone pairs. $\boldsymbol{r}_{initial}$ is composed of the phase differences of the initial pairs. The arbitrary integer of the initial pairs ,$\boldsymbol{n}_{initial}$, is estimated so as to fill the following equation:

$$\boldsymbol{r}_{mdsbf}-\pi\mathbf{1} \leq_{each} \boldsymbol{r}_{initial}+2\pi\boldsymbol{n}_{initial} \leq_{each} \boldsymbol{r}_{mdsbf}+\pi\mathbf{1}, \quad (5)$$

where $\boldsymbol{q}_{mdsbf}$ is DOA estimation result by MDSBF, and defined as $\begin{bmatrix} \cos\theta\cos\phi \\ \sin\theta\cos\phi \\ \sin\phi \end{bmatrix}$, and $\boldsymbol{r}_{mdsbf}$ is estimation of the phase differences of the initial pairs and is defined as $\boldsymbol{D}_{initial}\boldsymbol{q}_{mdsbf}$, $\boldsymbol{x} \leq_{each} \boldsymbol{y}$ means that each element of $\boldsymbol{y}$ is larger than or equivalent to each element of $\boldsymbol{x}$, and the vector $\mathbf{1}$ is a vector whose elements have the value 1. The remained microphone pairs are sorted in ascending order of the distance between each microphone pair and are divided into P clusters. The vector $\boldsymbol{n}_p$ composed of the arbitrary integers of the p-th cluster are estimated in the stepwise manner from the first cluster to the P-th cluster. $\boldsymbol{n}_p$ is estimated so as to fill the following equation:

$$\hat{\boldsymbol{r}}_p-\pi\mathbf{1} \leq_{each} \boldsymbol{r}_p+2\pi\boldsymbol{n}_p \leq_{each} \hat{\boldsymbol{r}}_p+\pi\mathbf{1}, \quad (6)$$

where $\hat{\boldsymbol{r}}_1 = \boldsymbol{D}_1(\boldsymbol{r}_{initial}+2\pi\boldsymbol{n}_{initial})$, $\hat{\boldsymbol{r}}_p = \boldsymbol{D}_p\boldsymbol{D}^+_{0\_p-1}(\boldsymbol{r}_{0\_p}+2\pi\boldsymbol{n}_{0\_p})$ when $p \geq 2$, $\boldsymbol{D}_{0\_p-1}$ is the distance matrix composed of all microphone pairs from the initial pairs to the $p-1$-th cluster. The stepwise process is approximation for maximization of Eq. 4. Finally, (θ_{ml},ϕ_{ml}) in Eq. 4 can be obtained by $\hat{\boldsymbol{r}}_P$.

4. EXPERIMENT

The sound sources were human speech convolved by the impulse responses obtained using the microphone array mounted on EMIEW2 in the reverberant environment ($RT_{60} = 300$ ms), and were mixed with multichannel noise signals recorded using EMIEW2 in a noisy exhibition hall (SNR=0 dB). The spectrogram of the noise source component is shown in Fig. 3. The sampling rate was set to be 8 kHz. Frame size was 512, and frame shift was 256. The average power of each speech source and noise source component were adjusted to be the same value, The proposed combination method of MDSBF and SPIRE is described as MDSBF+SPIRE. SPIRE, MDSBF and MDSBF+SPIRE were compared with MUSIC for wideband signals [6], and SRP-PHAT [7]. The number of the partitions of the resulting histogram was set to be 180 for azimuth and 90 for elevation. In

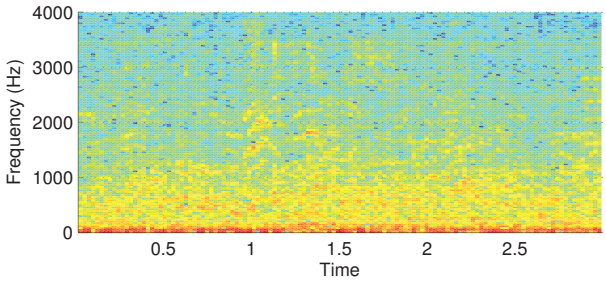


Fig. 3. Spectrogram of noise source component

MDSBF+SPIRE, $\frac{\Delta\theta}{2}$ and $\frac{\Delta\phi}{2}$ were set to be 20 degree. "MDSBF with rough resolution" is described as DOA estimation by MDSBF with 20 degree resolution. α_θ and α_ϕ were set to be 2 degree. There were two sound sources at each experiment. In Fig. 4, the ratio of correctly estimated DOA results for each method is shown. DOA estimation is performed at every 0.5 second. The right answers are

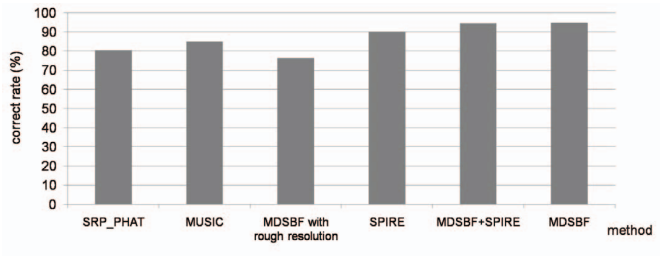


Fig. 4. Correct rate of DOA estimation results: Admissible error is 10 degrees.

made by MUSIC using the impulse response of each source. The distance between two sources and EMIEW2 was 1 m. DOA difference between two sources is varying (180, 120, 60, 30 degree). The sparseness based methods are superior to MUSIC or SRP-PHAT. SPIRE is slightly inferior to MDSBF. MDSBF+SPIRE is superior to SPIRE, and MDSBF+SPIRE is almost as correct as MDSBF. In Fig. 5, the resulting histograms are shown. The wavelength is about 3 s. In "close case", two sources are close to each other. In this case, the resulting histograms of MUSIC or SRP-PHAT has only one peak. On the other hand, the resulting histograms of MDSBF, SPIRE, and MDSBF+SPIRE has two peaks. When the sources are close to each other, the sparseness based methods are shown to be superior to MUSIC or SRP-PHAT. The peaks of MDSBF+SPIRE is slightly sharper than that of SPIRE. The proposed combination method (MDSBF+SPIRE) is shown to be effective. Even if average SNR is low, speech sources are expected to be concentrated on a few time-frequency points (*sparse TF points*), in which SNR is higher than average SNR, and the peaks of the histogram are considered to be composed of DOA estimation results at sparse TF points. On other time-frequency points, SNR is expected to be lower than average SNR. In these time-frequency points, DOA estimation results are not considered to concentrate in a particular direction but spread on whole azimuth and elevation plane. To evaluate accuracy of DOA estimation results at sparse TF points, the following measure C is

(a) Two sources are separate ("separate case"): $(\theta, \phi) = (-60, 30)$ or $(60, 30)$.

(b) Two sources are close to each other ("close case"): $(\theta, \phi) = (0, 0)$ or $(0, 30)$.

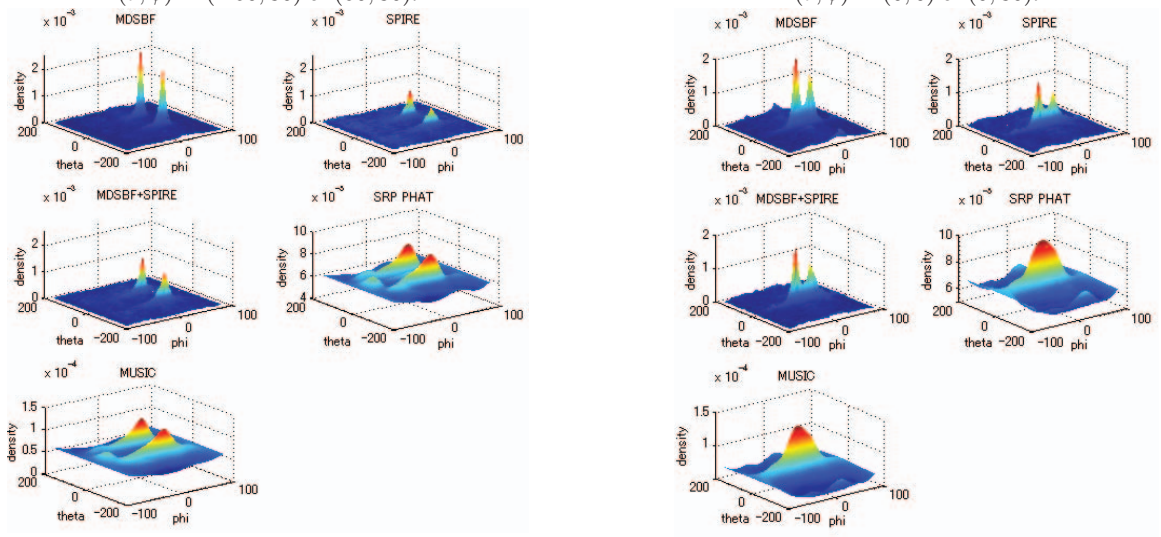


Fig. 5. Resulting DOA histograms of two speech sources: wavelength is about 3 s.

used:

$$C = \frac{\sum_{\tau,f} \delta(p_{main} \geq 10dB)\delta(|\hat{\theta} - \theta| < Th_\theta)\delta(|\hat{\phi} - \phi| < Th_\phi)}{\sum_{\tau,f} \delta(p_{main} \geq 10dB)}, \quad (7)$$

where $(\hat{\theta}, \hat{\phi})$ is DOA estimation result, (θ, ϕ) is the correct DOA of the source with the maximum power at (τ, f), $\delta(x)$ is 1 when x is true, otherwise $\delta(x) = 0$, $p_{main} = 20 \log_{10} \max_i \frac{|\boldsymbol{y}_i(f,\tau)|}{|\boldsymbol{x}(f,\tau)| - |\boldsymbol{y}_i(f,\tau)|}$, and $\boldsymbol{y}_i(f, \tau)$ is the multichannel signal composed of the i-th source component at each microphone. In Table. 2, the results of C for "separate case" and "close case" are shown. Furthermore, processing time which is estimated on a personal computer (Core2 Quad 2.66 GHz, Windows XP) and allocated memory size are also shown. MDSBF is superior to SPIRE, but computational cost of MDSBF is shown to be much larger than SPIRE. MDSBF+SPIRE was shown to be more correct than SPIRE, and computational cost of MDSBF+SPIRE is considerably smaller than MDSBF.

Table 2. DOA estimation accuracy at each time frequency point

	SPIRE	MDSBF	MDSBF+SPIRE
C ("separate case")	0.36	0.51	0.46
C ("close case")	0.36	0.41	0.38
processing time	0.02	1.0	0.02
Allocated memory size for DOA estimation	1 MB	500 MB	5 MB

5. CONCLUSION

In this paper, we focused on DOA estimation of human speech sources using the microphone array mounted on the head of the robots. For human symbiotic robots, DOA estimation of human speech sources are important to communicate with human. We introduced two methods based on sparseness assumption of speech sources, MDSBF and SPIRE. Furthermore, a combination method of MDSBF and SPIRE was proposed to obtain more accurate DOA estimation result with smaller computational cost than MDSBF. MDSBF with rough resolution is performed prior to SPIRE execution. SPIRE estimates DOA of sources around MDSBF estimation result. Experimental results with EMIEW2 showed that sparseness based methods can estimate sources direction more accurately than conventional methods. Additionally, the proposed combination method (MDSBF+SPIRE) achieved more accurate DOA estimation result than SPIRE and computational cost was remarkably smaller than MDSBF.

6. REFERENCES

[1] Y. Hosoda, et al., "Basic design of human-symbiotic robot EMIEW," In *Proc. of IROS2006*, pp.5079-5084, 2006.

[2] M. Togami, Y. Obuchi, and A. Amano, "Automatic Speech Recognition of Human-Symbiotic Robot EMIEW," in "Human-Robot Interaction", pp. 395-404, I-tech Education and Publishing, 2007.

[3] M. Aoki, et al., "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," " *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149-157, 2001.

[4] M. Togami, T. Sumiyoshi, and A. Amano, "Stepwise phase difference restoration method for sound source localization using multiple microphone pairs," *Proc. ICASSP2007*, vol. I, pp. 117-120, 2007.

[5] M. Togami and Y. Obuchi, "Stepwise Phase Difference Restoration Method for DOA Estimation of Multiple Sources", *IEICE Trans. on Fundamentals*, vol. E91-A, no. 11, 2008 (to appear).

[6] F. Asano, et al., "Fusion of audio and video information for detecting speech events," *Proc. IF2003*, pp. 386-393, 2003

[7] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, chapter 8, pp. 157-180, Springer, Berlin, Germany, 2001.