TALKER TRACKING DISPLAY ON AUTONOMOUS MOBILE ROBOT WITH A MOVING MICROPHONE ARRAY

Takanobu Nishiura[†,††]*, Masaya Nakamura*[‡]*, Akinobu Lee*[‡]*, Hiroshi Saruwatari*[‡]*, and Kiyohiro Shikano*[‡]

[†] Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama, 640-8510 Japan
[††] ATR Spoken Language Translation Research Labs.
[‡] Graduate School of Information Science, Nara Institute of Science and Technology
nishiura@sys.wakayama-u.ac.jp

ABSTRACT

A microphone array is the most effective technique for hands-free speech communication with a robot. However, it is known that the estimation accuracy of the sound source direction degrades under reverberant conditions. The estimation of the source direction is difficult under highly reverberant conditions even if the CSP (Cross-power Spectrum Phase analysis)-based method, which is one of the conventional methods for the estimation of source directions, is used. Therefore, it is necessary to use a method for source-direction estimation which is robust against reverberation when the microphone array is carried on an autonomous mobile robot.

In this paper, we propose a new reverberation-robust method for the estimation of sound-source directions based on the synchronous addition of CSP coefficients obtained by a circular microphone array. Also, we propose a sound-source localization method that improves the localization accuracy by using the moving microphone array on an autonomous mobile robot. To evaluate the proposed method, the source-localization accuracy of the moving microphone array was quantified by computer simulation. We also carried out an experiment in which a mobile robot approached a talker by using only information on the source localization. The results show that the proposed method can estimate the sound-source location accurately and that the robot can approach the talker even under the condition that the reverberation time is 0.85 sec.

1. INTRODUCTION

It is very important for an autonomous mobile robot to be able to capture distant talking speech and that the image of the talker with high quality. To achieve these aims, talker localization is needed, and a microphone array is an ideal candidate for that purpose. However, conventional talker localization methods in multiple sound source environments not only have difficulty localizing the multiple sound sources accurately, but also have difficulty localizing the target talker among known multiple sound source positions.

To cope with these problems, we proposed a new talker localization method consisting of two algorithms. One algorithm is for multiple sound source localization based on CSP (Cross-power Spectrum Phase analysis) method [1]. The other algorithm is for sound source identification among localized multiple sound sources towards talker localization[2, 3]. The performance of these algorithms with a fixed microphone array was confirmed in previous paper[3]. In this paper, we particularly focus on the talker localization performance of an autonomous mobile robot with a moving microphone array for talker tracking display and distant talking speech capture. We try to achieve a higher talker tracking display performance by using a moving microphone array. Our final goal is to acquire information on the acoustic environment by using autonomous mobile robots.

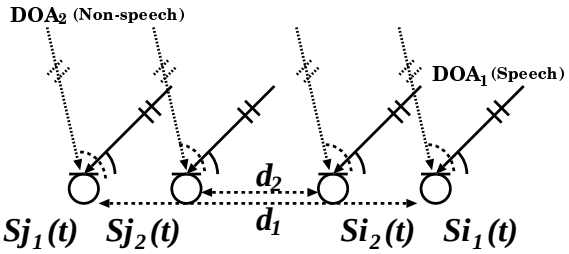


Figure 1: Capture of two sound signals with a fixed microphone array.

2. TALKER LOCALIZATION WITH A FIXED MICROPHONE ARRAY

As shown in Figure 1, we assume that the desired speech DOA_1 comes from the right and the undesired noise (non-speech) DOA_2 comes from the left. In this situation, talker localization is necessary for effectively capturing distant talking speech with a fixed microphone array.

Accordingly, we proposed a new talker localization algorithm[1], as shown in Figure 2. First, multiple sound DOAs (Directions Of Arrival) are estimated with the CSP coefficient addition method after multiple sound signals are captured. Then, the sound signals of the estimated DOAs are enhanced by steering the microphone array toward them. Finally, after identification between "speech" or "non-speech" using statistical speech and environmental sound models, the talker can be localized among the enhanced multiple sound signals.

2.1. DOA estimation with CSP coefficient addition method

DOA (Direction Of Arrival) must be estimated in order to automatically steer the microphone array. CSP (Cross-power Spec-

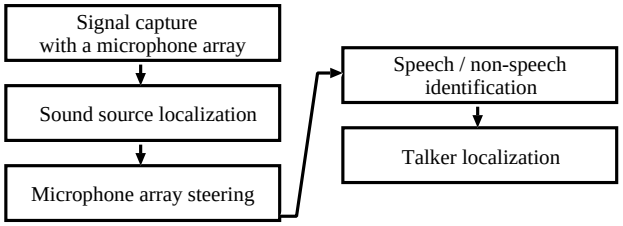


Figure 2: Talker localization algorithm overview.

trum Phase analysis) is a very popular method for estimating DOA. However, multiple DOA estimation with CSP is very difficult because of the cross-correlation of multiple sound signals. To overcome this problem, at ICASSP2000, we proposed a CSP coefficient addition method to estimate multiple DOAs[1]. In the environment of Figure 1, the CSP coefficients are derived from Equation (1).

$$\mathrm{CSP}_{i_n,j_n}(k) = \mathrm{IDFT}\left[\frac{\mathrm{DFT}\left[s_{i_n}(t)\right]\mathrm{DFT}\left[s_{j_n}(t)\right]^*}{\left|\mathrm{DFT}\left[s_{i_n}(t)\right]\right|\left|\mathrm{DFT}\left[s_{j_n}(t)\right]\right|}\right], \tag{1}$$

where t and k are the time index, $\mathrm{DFT}\,[\cdot]$ (or $\mathrm{IDFT}[\cdot]$) is the discrete Fourier transform (or the inverse discrete Fourier transform), and the symbol * is the complex conjugate. Then, CSP coefficients are added, as shown in Equation (2).

$$\begin{aligned}\mathrm{CSP}_{i,j}(\theta) &= \sum_{n=1}^{N}\mathrm{CSP}_{i_n,j_n}(\theta),\\ subject\quad to\quad \theta &= \cos^{-1}\left(\frac{c\cdot k/F_s}{d_n}\right),\end{aligned} \tag{2}$$

where N is the number of additions, d_n is the distance between two adjacent transducers, c is the sound propagation speed, and F_s is the sampling frequency. The DOAs can be accurately estimated by finding the maximum values of the added CSP coefficients by Equation (3).

$$\mathrm{DOA_n} = \underset{\theta}{\mathrm{argmax}}(\mathrm{CSP}_{ij}(\theta)). \tag{3}$$

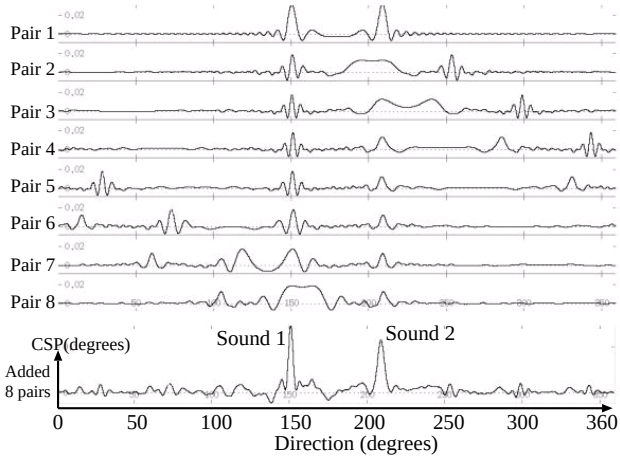
Figure 3 shows an example of the CSP coefficient addition result with a circular microphone array. As a result, we can confirm that true DOAs can be accurately estimated by using CSP coefficient addition method, even in a highly reverberant environment.

2.2. Microphone array steering for speech enhancement

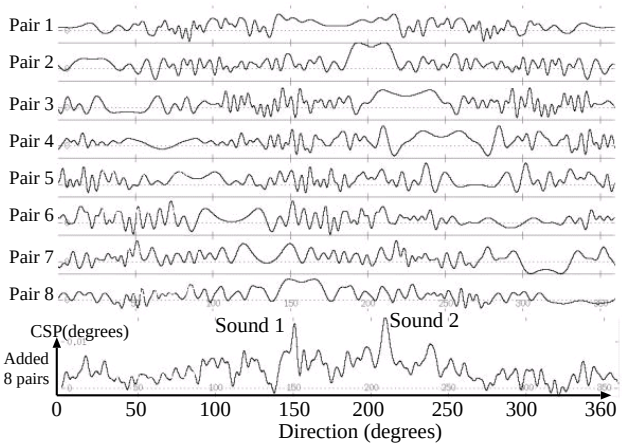
Microphone array steering is necessary to capture distant signals effectively. In this paper, a delay-and-sum beamformer [5] is used to steer the microphone array. Multiple sound signals of estimated DOAs are enhanced by the microphone array steering because the delay-and-sum beamformer can form directivity to the estimated DOAs.

2.3. Speech/non-speech identification based on GMMs

Multiple sound signals are captured effectively and enhanced by microphone array steering. Therefore, the talker can be localized by identifying the enhanced multiple sound signals. Until now, a



(A) Anechoic environment.



(B) High reverberant environment.

Figure 3: CSP coefficient addition result.

speech model alone was usually used for speech/non-speech segmentation [4] or identification. However, a single speech model has problems in that it not only requires a threshold to identify between "speech" and "non-speech", but also degrades the identification performance in noisy reverberant environments. To overcome these problems, at EUROSPEECH2001, we proposed a speech/non-speech identification algorithm that uses statistical speech and environmental sound GMMs (Gaussian Mixture Models)[2]. The multiple sound signals enhanced by using microphone array steering are identified by Equation (4).

$$\hat{\lambda} = \underset{\lambda}{\mathrm{argmax}}\, P(S(w)|\lambda_s, \lambda_n), \tag{4}$$

where $S(w)$ is the enhanced signal with microphone array steering (frequency domain), λ_s represents the statistical speech model, and λ_n represents the statistical environmental sound model. The enhanced signals are identified as "speech" or "non-speech" by es-

timating the maximum likelihood in Equation (4). This algorithm allows the talker to be localized from among estimated DOAs.

2.3.1. Speech and environmental sound database

Numerous sound sources are necessary to design the speech and environmental sound GMMs. Therefore, we use the ATR speech Database (ATR-DB) [6] to design the speech model and the RWCP (Real World Computing Partnership) sound scene database (RWCP-DB) [7, 8] which includes various environmental sounds to design the non-speech model. The RWCP-DB also includes numerous impulse responses measured in various acoustical environments. These impulse responses are used to conduct evaluation experiments in various acoustical environments.

3. TALKER LOCALIZATION WITH A MOVING MICROPHONE ARRAY

Figure 4 shows an example of talker localization with a moving microphone array on an autonomous mobile robot. The autonomous mobile robot can localize the talker position by finding the crossing points based on estimated DOAs with the CSP coefficient addition method. However, a multiple microphone array is necessary to find the crossing points. To cope with this problem, we try to localize the talker position and track the talker with a moving microphone array as shown in Figure 4. The algorithm for this is shown in Figure 5.

Speech detection is first carried out based on likelihood with speech model[4]. If speech can be detected, multiple sound sources are localized with the CSP coefficient addition method after multiple sound signals have been captured with the microphone array. Then, these localized sound signals are enhanced by steering the microphone array toward them. Then, after identification between "speech" or "non-speech" using statistical speech and environmental sound models based on GMMs, the talker can be localized among the enhanced multiple sound signals. Finally, the autonomous mobile robot can localize the talker position effectively by moving the microphone array based on the estimated talker direction. Also, if speech cannot be detected for a long time, the autonomous mobile robot with the microphone array will pause. In Figure 5, W is the threshold for detecting the non-speech terms.

With above algorithms, we try to localize and track the talker by using an autonomous mobile robot with a moving microphone array for talker tracking display.

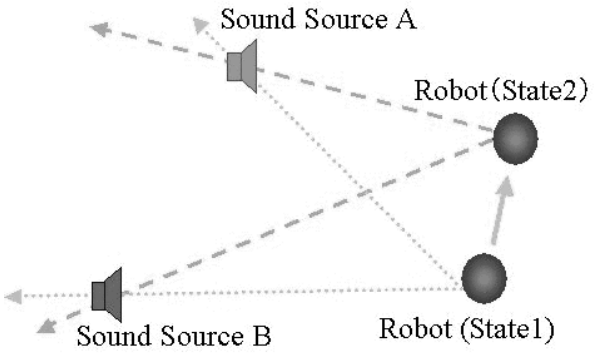


Figure 4: Moving microphone array on autonomous mobile robot.

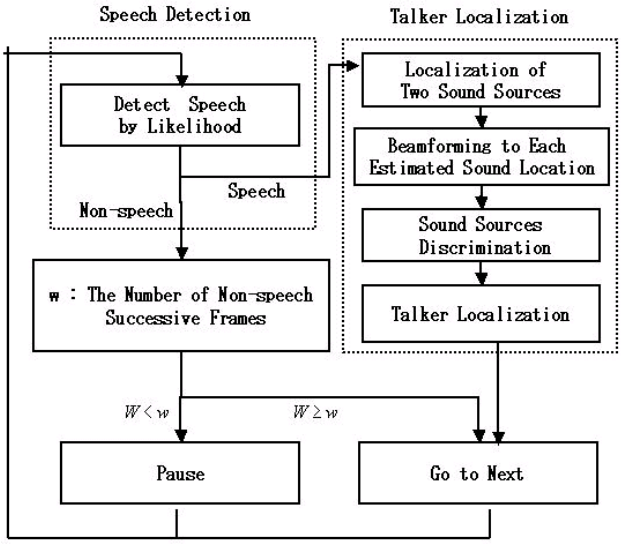


Figure 5: Talker tracking algorithm for moving microphone array on autonomous mobile robot.

Table 1: Experimental conditions for the moving microphone array

Microphone array	Circular type, 16 transducers, and 60 cm diameters
Moving speed	20 cm/sec
Reverberation time $T_{[60]}$	0.55 and 0.85 sec.
Sampling frequency	16 kHz
SNR	0, 5,10, and 15 dB

4. EVALUATION EXPERIMENTS

Evaluation experiments are conducted by computer simulation. The experimental environment is shown in Figure 6 and the experimental condition is shown in Table 1 and Table 2. Reverberation time in the room ($T_{[60]}$) was 0.55 sec and 0.85 sec. We simulated an autonomous mobile robot with a circular type microphone array which has 16 transducers and 60 cm diameters. Two sound sources, speech and non-speech, exist in the room. The SNR was 0, 5, 10, and 15 dB. In this condition, we evaluated the talker tracking display performance in an experiment in which the mobile robot approached a talker using only information on the source localization.

5. EXPERIMENTAL RESULTS

Figure 7 shows the talker tracking results for the SNR = 10 dB, 2 sound sources, and $T_{[60]}$ = 0.85 sec environment. As shown in Figure 7, we were able to confirm that the autonomous mobile robot localizes the talker position and approaches him/her effectively under highly reverberant conditions by using proposed algorithm with the moving microphone array. We were also able to confirm the same tendency results for the $T_{[60]}$ = 0.55 sec environment. Therefore, we confirmed that the autonomous mobile robot

Table 2: Experimental conditions for speech / non-speech identification

Frame length	32 msec. (Hamming window)
Frame interval	8 msec.
Feature vector	MFCC (16 orders, 4 mixtures), ΔMFCC (16 orders, 4 mixtures), Δpower (1 order, 2 mixtures)
Number of models	Speech: 1 model Non-speech: 1 model
Speech DB	ATR speech DB SetA [6]
Speech model training	200 words × 5 subjects (2 females and 3 males)
Non-speech DB	RWCP-DB [7, 8]
Non-speech model training	92 sounds × 4 sets
Test data (Open)	
Speech:	216 words × 1 subjects (1 male)
Non-speech:	White Gaussian noise

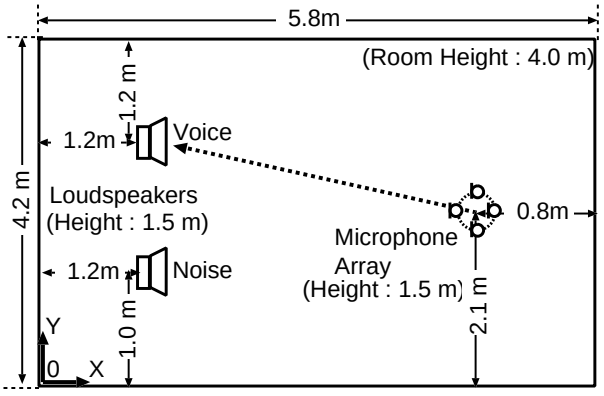


Figure 6: Experimental environment.

with the moving microphone array could display talker tracking results based on the proposed talker localization algorithm.

6. CONCLUSION

In this paper, we particularly focused on the talker localization performance of an autonomous mobile robot with a moving microphone array for talker tracking display and distant talking speech capture. From evaluation experiment results, we confirmed that the autonomous mobile robot can localize the talker position and approaches him/her effectively by using the moving microphone array, even under highly reverberant conditions. Therefore, we also confirmed that the autonomous mobile robot with the moving microphone array can display talker tracking results based on the proposed talker localization algorithm. In future work, we will evaluate the performance in real acoustic environments. Our final goal is to acquire information on the acoustic environment by using autonomous mobile robots.

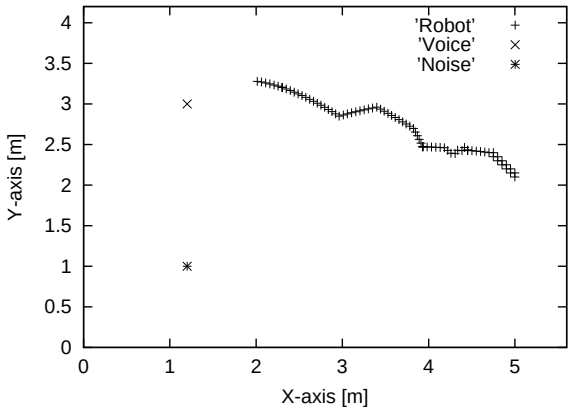


Figure 7: Talker tracking results (SNR = 10 dB, $T_{[60]}$ = 0.85 sec.)

7. ACKNOWLEDGMENT

This research was partially supported by the "Auditory Brain" project of JST CREST and The Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid No. 14780288.

8. REFERENCES

[1] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of Multiple Sound Sources Based on a CSP Analysis with a Microphone Array," Proc. ICASSP2000, pp. 1053–1056, Jun. 2000.

[2] T. Nishiura, S. Nakamura, and K. Shikano, "Statistical Sound Source Identification in Real Acoustic Environment for Robust Speech Recognition Using a Microphone Array," Proc. EUROSPEECH2001, pp. 2611–2614, Sep. 2001.

[3] T. Nishiura, S. Nakamura, and K. Shikano, "Talker Localization in a Real Acoustic Environment Based on DOA Estimation and Statistical Sound Source Identification," Proc. ICASSP2002, pp. 893–896, May 2002.

[4] R. Singh, M.L. Seltzer, B. Raj, and R.M. Stern, "Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination," Proc. ICASSP2001, pp. 273–276, May 2001.

[5] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," J. Acoust. Soc. Am., Vol. 78, No. 5, pp. 1508–1518, Nov. 1985.

[6] K. Takeda, Y. Sagisaka, and S. Katagiri, "Acoustic-Phonetic Labels in a Japanese Speech Database," Proc. European Conference on Speech Technology, Vol. 2, pp. 13–16, Oct. 1987.

[7] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, "Data Collection in Real Acoustical Environments for Sound Scene Understanding and Hands-Free Speech Recognition," Proc. Eurospeech99, pp. 2255–2258, Sep. 1999.

[8] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada,"Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition," Proc. LREC2000, pp. 965–968, May. 2000.