# DEVELOPMENT OF ZONAL BEAMFORMER AND ITS APPLICATION TO ROBOT AUDITION

*Nobuaki Tanaka*[1], *Tetsuji Ogawa*[2], *Kenzo Akagiri*[1], *and Tetsunori Kobayashi*[1]

[1]Dept. of Computer Science, Waseda University
[2]Waseda Institute for Advanced Study

## ABSTRACT

We have proposed a zonal beamformer (ZBF), which enhances the sound source located in a zonal space, and applied the ZBF to noise reduction systems for robot audition. A conversational partner of a robot does not always remain stationary with respect to the robot. In order to cope with such a situation, we have proposed a fan-like beamformer (FBF), which enhances the sound source located in a fan-like space in front of the robot under the assumption that the partner is in front of the robot. However, the FBF may degrade the noise reduction performance when directional noise sources are located behind the target source because the FBF widens the space as the distance from the robot increases. The ZBF can better improve the performance of eliminating the directional noise coming from behind the target source than the FBF because the ZBF has a considerably sharper directivity than the FBF.

## 1. INTRODUCTION

Robots that converse with people in real environments need to extract and recognize the speech utterances of their conversational partners. In addition, autonomous mobile robots need miniature microphone systems and signal processing devices because such robots have restrictions with respect to the weights and sizes of the devices that are mounted on them. Because of the miniaturization of these devices, robot audition systems require noise reduction systems that achieve high performances using low-computational-cost algorithms. Moreover, the target sound source (i.e., conversational partner) is not always at the same position.

Noise reduction methods using microphone arrays[1, 2] have been applied to the preprocessing of noisy speech recognition systems. Most of the beamforming techniques need a considerable number of microphones and large microphone spacings. Adaptive beamforming based on an independent component analysis (ICA) can carry out sound source separation by sequentially estimating the direction-of-arrivals (DOAs) of the sound source[3]. However, this method complicates the algorithm of real-time sound source separation for moving sources because this method induces unavoidable delays for the convergence of adaptive filters after the sound source positions are estimated. Therefore, these beamforming techniques may not be suitable for robot audition systems.

In contrast, we assume that a conversational partner of a robot is in front of the robot. Under this assumption, we proposed a fan-like beamformer (FBF), which enhances the sound sources located in the fan-like space in front of the robot, using compact microphone arrays[4, 5]. Since this method enhances not a sound source but a space, it can cope with the moving of the conversational partner without the delays that occur in adaptive beamforming. However, in this method, the enhanced space becomes larger as the distance between a sound source and the robot increases. In this case, directional noise sources located behind the target source degrade the performance of noise reduction. In order to solve this problem, we propose a zonal beamformer (ZBF), which enhances the sound sources located in the zonal space in front of the robot. The ZBF can improve the performance of eliminating the directional noise coming from behind the target sound source.

The rest of this paper is organized as follows. The microphone systems used are described in Sect. 2. The noise reduction methods are described in Sect. 3. In Sect. 4, details of the experimental investigation of the proposed method in terms of speech recognition performance, noise reduction performance, and speech quality are provided. Finally, in Sect. 5, the concluding remarks are presented.

## 2. MICROPHONE SYSTEM

We used compact and light-weight microphone arrays that are suitable for autonomous mobile robots.

### 2.1 MEMS microphones

We used four-line or six-line analog micro electro mechanical systems (MEMS) microphones, which are constructed on the basis of a semiconductor integrated technology and are significantly compact and light in weight. We used SPM0208HD5 made by Knowles Co., Ltd. The width, depth, and height of the microphone are 4.72 mm, 3.76 mm, and 1.25 mm, respectively. We prepared 1.5-cm$^2$ substrates, each of which consisted of a MEMS microphone and peripheral circuits, including a pre-amplifier. These substrates were mounted on the robot head.

### 2.2 Microphone arrangement

We placed microphone arrays on the top of the robot head, as shown in Fig. 1. We used the microphone arrays shown in Figs. 1(a) and 1(b) for developing the FBF and ZBF, respectively. These microphone arrangements were aimed at suppressing the influences of the reflections and diffractions induced by the robot head and body. Microphone channels were labeled as shown in Fig. 1. The front, right, and left direction of the robot were defined as zero, positive, and negative degrees, respectively. In this study, we assumed that the target speech utterances arrived from the front of the robot.

## 3. DIRECTIONAL NOISE REDUCTION

In this section, we review directional noise reduction method using FBF, which is our previous work[5, 6]. Next, we describe ZBF, which is an extension of FBF, in detail. In the present paper, $x_i(t)$ denotes a signal received by the microphones Chi$^{(F)}$ and Chi$^{(Z)}$ at a discrete time $t$, and $X(\omega,k)$ denotes an STFT coefficient of $x_i$, where $k$ and $\omega$ denote a discrete frame and a discrete frequency, respectively.

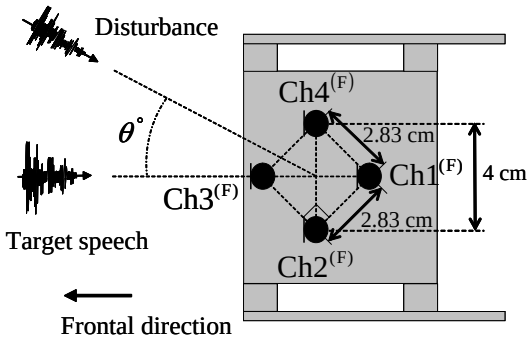
### 3.1 Fan-like beamformer (FBF)

We developed FBFs using the microphone array shown in Fig. 1(a). In this system, null beamformers and subtractive beamformers were developed, and then time-frequency masking was carried out using the outputs of these beamformers. $C_1(\omega,k)$ and $C_2(\omega,k)$ denote the spectral components of the outputs of the null beamformers that were developed by delay addition followed by subtraction using $X_1(\omega,k)$ and $X_3(\omega,k)$. $C_1(\omega,k)$ and $C_2(\omega,k)$ were computed as follows:

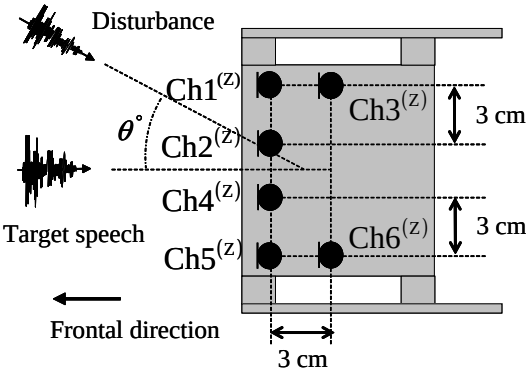
$$C_1(\omega,k) = X_3(\omega,k)\cdot\exp(-j\omega\tau_d) - X_1(\omega,k) \quad (1)$$

$$C_2(\omega,k) = X_1(\omega,k)\cdot\exp(-j\omega\tau_d) - X_3(\omega,k) \quad (2)$$

 ISSN 2076-1465 




(a) Microphone array for FBF.




(b) Microphone array for ZBF.

Figure 1: Microphone arrangements.

where $\tau_d$ denotes a delay corresponding to the spacing of the microphones placed in a diagonal position. The directivity patterns of $C_1$ and $C_2$ are shown in Fig. 2(a). In this case, $C_1$ and $C_2$ indicate the directivity with a null in the direction of 0° and that with a null in the direction of 180°, respectively.

$S_1(\omega,k)$ denotes a spectral component of the output of the subtractive beamformer developed using $X_1(\omega,k)$ and $X_3(\omega,k)$, and $S_2(\omega,k)$ denotes a spectral component of the output of the subtractive beamformer developed using $X_2(\omega,k)$ and $X_4(\omega,k)$. $S_1(\omega,k)$ and $S_2(\omega,k)$ were computed as follows:

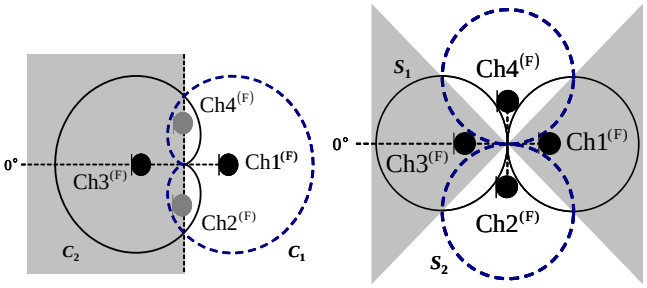
$$S_1(\omega,k) = X_1(\omega,k) - X_3(\omega,k) \quad (3)$$
$$S_2(\omega,k) = X_4(\omega,k) - X_2(\omega,k) \quad (4)$$

The directivity patterns of $S_1$ and $S_2$ are shown in Fig. 2(b). $S_1$ indicates the directivity that has maximum gains in the directions of 0° and 180°, and nulls in the directions of 90° and -90°. $S_2$ indicates the directivity that has maximum gains in the directions of 90° and -90°, and nulls in the directions of 0° and 180°. Note that the directivity patterns shown in Fig. 2 hold for the frequency range approximately from 300 to 2500 Hz.

In the case of the FBF, the spectral component of the target sound source, $\hat{S}_{\mathrm{FBF}}$, was estimated by the following time-frequency masking:

$$\hat{S}_{\mathrm{FBF}}(\omega,k) = \begin{cases} S_1(\omega,k), \\ \quad \text{if } |S_1(\omega,k)| > |S_2(\omega,k)| \\ \quad \text{and } |C_1(\omega,k)| < |C_2(\omega,k)| \\ \beta \cdot S_1(\omega,k), \quad \text{otherwise} \end{cases} \quad (5)$$

where $\beta$ denotes a flooring coefficient. This time-frequency masking suppressed the directional noise coming from the side of the robot by selecting the time-frequency components in which $S_1(\omega,k)$ was larger than $S_2(\omega,k)$, and then it suppressed the noise coming from behind the robot by selecting the components in which $C_2(\omega,k)$ was larger than $C_1(\omega,k)$. As a result, the FBF extracted the spectral components of the signal coming from the sound source located in the fan-like space in front of the robot (i.e., the common shaded area of Figs. 2(a) and 2(b)) as the target source.




(a) Null beamformers. (b) Subtractive beamformers.

Figure 2: Directivity patterns of beamformers for developing FBF.

### 3.2 Zonal beamformer (ZBF)

ZBF, which is a new contribution in the present paper, uses the microphone array shown in Fig. 1(b). The microphone arrangement of ZBF is intended to develop zonal spacial filter that can eliminate the directional noise coming from behind the target sound source. In this system, eight null beamformers and two subtractive beamformers were developed, then time-frequency masking was carried out using the outputs of these beamformers. Eight null beamformers were computed as follows:

$$C_1(\omega,k) = X_2(\omega,k) - X_1(\omega,k) \cdot \exp(-j\omega\tau_n) \quad (6)$$
$$C_2(\omega,k) = X_1(\omega,k) - X_2(\omega,k) \cdot \exp(-j\omega\tau_n) \quad (7)$$
$$C_3(\omega,k) = X_5(\omega,k) - X_4(\omega,k) \cdot \exp(-j\omega\tau_n) \quad (8)$$
$$C_4(\omega,k) = X_4(\omega,k) - X_5(\omega,k) \cdot \exp(-j\omega\tau_n) \quad (9)$$
$$C_5(\omega,k) = X_1(\omega,k) - X_3(\omega,k) \cdot \exp(-j\omega\tau_n) \quad (10)$$
$$C_6(\omega,k) = X_3(\omega,k) - X_1(\omega,k) \cdot \exp(-j\omega\tau_n) \quad (11)$$
$$C_7(\omega,k) = X_5(\omega,k) - X_6(\omega,k) \cdot \exp(-j\omega\tau_n) \quad (12)$$
$$C_8(\omega,k) = X_6(\omega,k) - X_5(\omega,k) \cdot \exp(-j\omega\tau_n) \quad (13)$$

where $\tau_n$ denotes a delay corresponding to the spacing of the neighbor microphones. Figure 3 shows the directivity patterns of these beamformers.

$S_1(\omega,k)$ and $S_2(\omega,k)$ denote spectral components of the outputs of the subtractive beamformers computed as follows:

$$S_1(\omega,k) = X_1(\omega,k) - X_3(\omega,k) \quad (14)$$
$$S_2(\omega,k) = X_5(\omega,k) - X_6(\omega,k) \quad (15)$$

The directivity patterns of $S_1$ and $S_2$ are shown in Fig. 4. Note that the directivity patterns shown in Figs. 3 and 4 hold for the frequency range approximately from 300 to 3500 Hz.

The ZBF estimated the spectral component of the target source, $\hat{S}_{\mathrm{ZBF}}$, by the following time-frequency masking:

$$\hat{S}_{\mathrm{ZBF}}(\omega,k) = \begin{cases} (S_1(\omega,k) + S_2(\omega,k))/2, \\ \quad \text{if } \alpha \cdot |C_1(\omega,k)| > |C_2(\omega,k)|, \\ \quad \alpha \cdot |C_4(\omega,k)| > |C_3(\omega,k)|, \\ \quad |C_5(\omega,k)| > |C_6(\omega,k)|, \\ \quad \text{and } |C_7(\omega,k)| > |C_8(\omega,k)| \\ \beta \cdot (S_1(\omega,k) + S_2(\omega,k))/2, \\ \quad \text{otherwise} \end{cases} \quad (16)$$

where $\alpha$ denotes a positive constant for adjusting the width of the target zonal space and $\beta$ denotes a flooring coefficient. Figure 5

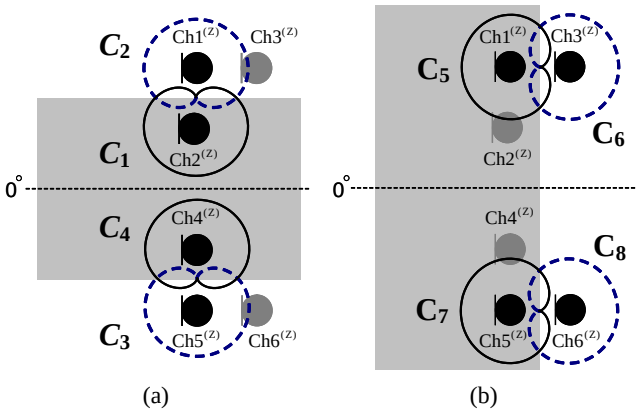



Figure 3: Directivity patterns of null beamformers for developing ZBF.

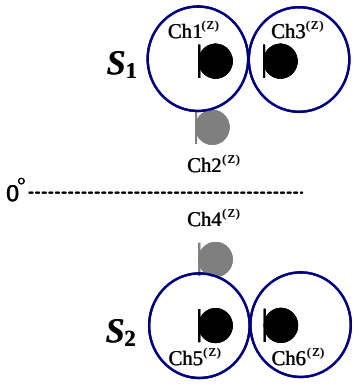



Figure 4: Directivity patterns of subtractive beamformers for developing ZBF.

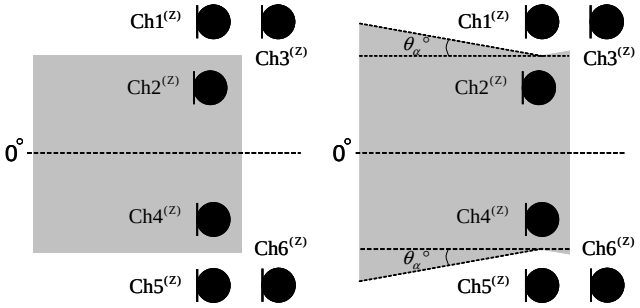



Figure 5: Directivity patterns of ZBF.

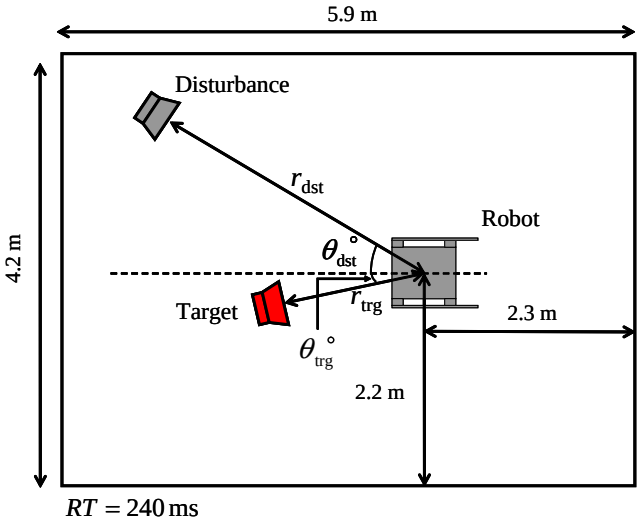



Figure 6: Recording environment.

shows the space enhanced by the original ZBF with $\alpha = 1.0$ and the space enhanced by the wider ZBF with $\alpha > 1.0$. Time-frequency masking with $\alpha = 1.0$ (original ZBF) enhanced the sound sources located in the shaded zonal space shown in Fig. 3(a) by selecting the time-frequency components in which $C_1(\omega,k)$ was larger than $C_2(\omega,k)$ and $C_4(\omega,k)$ was larger than $C_3(\omega,k)$, and then it enhanced the sound sources located in the shaded space shown in Fig. 3(b) by selecting the time-frequency components in which $C_5(\omega,k)$ was larger than $C_6(\omega,k)$ and $C_7(\omega,k)$ was larger than $C_8(\omega,k)$. As a result, the ZBF extracts the spectral components of the signal coming from the sound source located in the zonal space in front of the robot (i.e., the common shaded area of Figs. 3(a) and 3(b)) as the target source.

Since the original ZBF with $\alpha = 1.0$ formed a small width of zonal directivity, this beamformer could deteriorate the performance of estimating the target spectral components when the target source moved from a position directly in front of the robot. In order to solve this problem, we modified the original ZBF so that it could enhance the sound sources at wider angles (i.e., Fig. 5(b)) than the original ZBF (i.e., Fig. 5(a)) by using $\alpha > 1.0$.

## 4. NOISE REDUCTION EXPERIMENT

We carried out an experimental comparison of noise reduction performances between the FBF and ZBF under the condition that the target source was not always located at a position directly in front of the robot and the directional noise was simultaneously observed. In the present study, the FBF and ZBF were evaluated in terms of the automatic speech recognition performance on the basis of the word accuracy, the noise reduction performance on the basis of the noise reduction rate (NRR)[7], and the speech quality on the basis of the log-spectral distance (LSD)[8].

### 4.1 Experimental condition

#### *4.1.1 Speech materials*

The target speech utterances consisted of 100 sentences, which were spoken by 23 male speakers; these sentences were taken from a continuous speech database that contained sentences from Japanese newspaper articles[10]. In the case of directional noise (i.e., disturbance speech utterances), 100 sentences were selected from the same database; however, these sentences were different from the target speech utterances. In this case, each disturbance speech utterance was of approximately the same duration as the corresponding target speech utterance.

#### *4.1.2 Speech recording*

Figure 6 shows the recording environment. A microphone array was placed on the head of the conversation robot "ROBISUKE"[9]. We placed a target source at one of the five frontal positions $(\theta_{\rm trg}, r_{\rm trg}) = (0^\circ, 1.0$ m), $(\pm 5^\circ, 1.0$ m), or $(\pm 10^\circ, 1.0$ m), and we placed a disturbance source at one of the eight positions $(\theta_{\rm dst}, r_{\rm dst}) = (\pm 20^\circ, 2.5$ m), $(\pm 30^\circ, 2.5$ m), $(\pm 40^\circ, 2.5$ m), or $(\pm 50^\circ, 2.5$ m). The heights of the target and disturbance sources were 1.0 m. The target and disturbance speech utterances were recorded separately. 100 utterances were played back through a loudspeaker placed at each position and then they were observed at the microphones mounted on the robot head. In total, 500 utterances and 800 ut-

Table 1: Setup for noise reduction.

| | |
|---|---|
| sampling frequency | 16 kHz |
| frame length | 64 ms (w/ 64 ms zero padding) |
| frame shift | 16 ms |
| analysis window | Hamming window |
| analysis range | 300–5500 Hz |

Table 2: Setup for speech recognition.

| | |
|---|---|
| sampling frequency | 16 kHz |
| frame length | 25 ms |
| frame shift | 10 ms |
| analysis window | Hamming window |
| feature parameters | 12 MFCCs, 12 ΔMFCCs, and a Δlog energy |

terances were recorded for the target and disturbance speech utterances, respectively. A disturbance speech utterance was electronically superposed on the corresponding target speech utterance at an SNR of 0 dB.

#### *4.1.3 Noise reduction and speech recognition*

Experimental conditions for noise reduction and acoustic feature extraction are listed in Tables 1 and 2, respectively. For noise reduction, we used zero padding in the time domain to increase the frequency resolution. We used the analysis range from 300 to 5500Hz in order to reduce the influence of spacial aliasing to the beamformers. At high-frequency bands in that analysis range (e.g., above 3500Hz), the directivity patterns shown in Figs. 2, 3 and 4 do not hold. Despite this, preliminary experiments indicated that the frequency range from 300 to 5500 Hz was necessary for the purpose of speech recognition. The flooring coefficient used in time-frequency masking (i.e., $\beta$ in Eqs. 5 and 16) was 0.01. Acoustic models were trained with 20414 sentences spoken by 133 male speakers, taken from the ASJ database[10], which consisted of Japanese newspaper article sentences (ASJ-JNAS) and phonetically balanced sentences (ASJ-PB) recorded with close-talking microphones. We used tied-state triphones with 2000 states. The distribution function in each state of the models was represented by a 16-mixture Gaussian distribution with diagonal covariances. We used word trigram language models that were constructed using a lexicon with a vocabulary size of 20 K.

#### *4.1.4 Evaluation criterion*

In this experiment, we evaluated the performances of the proposed method in terms of the word accuracy, NRR, and LSD, which are frequently used in assessments of noise reduction systems and automatic speech recognition systems, for separated speech utterances.

The word accuracy was calculated as follows:

$$\text{WA}\ \ (\%) = \frac{N-D-S-I}{N} \times 100 \tag{17}$$

where $N$, $D$, $S$, and $I$ denote the number of words included in the correct word sequences, deletion errors, substitution errors, and insertion errors, respectively.

The NRR was computed as follows:

$$\text{NRR (dB)} \quad = \quad \text{SNR}^{(\text{O})} - \text{SNR}^{(\text{I})} \tag{18}$$

where $\text{SNR}^{(\text{O})}$ and $\text{SNR}^{(\text{I})}$ denote the output SNR, which is computed using the separated signal of the target sound and that of the disturbance sound, and the input SNR, which is computed using the observed signal of the target sound and that of the disturbance sound, respectively.

The LSD was computed as follows:

$$\text{LSD (dB)} = \frac{10}{K}\sum_{k=0}^{K-1}\left[\frac{1}{W}\sum_{\omega=0}^{W-1}\left(\log_{10}\frac{|\hat{S}_{\text{BF}}(\omega,k)|^2}{|S_{\text{ref}}(\omega,k)|^2}\right)^2\right]^{\frac{1}{2}} \tag{19}$$

where $\hat{S}_{\text{BF}}(\omega,k)$ denotes the spectral component of the sound separated using the FBF or ZBF, and $S_{\text{ref}}(\omega,k)$ denotes the spectral component of the reference sound, which was recorded when only the target sound was observed. $K$ and $W$ denote the number of discrete frames and that of discrete frequencies, respectively.

### 4.2 Experimental result

Figures 7, 8, and 9 show the word accuracies, NRRs, and LSDs of the FBF, original ZBF ($\alpha = 1.0$ in Eq. 16, $\theta_\alpha = 0^\circ$ in Fig. 5), and 10°-wider ZBF ($\alpha = 1.42$ in Eq. 16, $\theta_\alpha = 10^\circ$ in Fig. 5) as a function of the DOAs of a disturbance source (i.e., $\theta_{\text{dst}} = \pm 20^\circ$, $\pm 30^\circ$, $\pm 40^\circ$, and $\pm 50^\circ$), respectively. Each bar represents the average performance for five DOAs of a target source (i.e., $\theta_{\text{trg}} = 0^\circ$, $\pm 5^\circ$, and $\pm 10^\circ$). The reason why the performances shown in these figures were asymmetrical to the DOAs of the disturbance source was that the phase characteristics of the microphones used were slightly different.

In the evaluation in terms of the speech recognition performance, the word accuracy was below 0% (e.g., -3.0%), on an average for the DOAs of the target and disturbance source, without any noise reduction. The word accuracy did not improve (e.g., 1.7%) even with the application of the conventional delay-and-sum (DS) beamformer. In these cases, speech recognition was not effective because considerable insertion errors were induced by the directional noise. In contrast, the word accuracies achieved when the FBF, original ZBF, and 10°-wider ZBF were carried out were 27.0%, 39.1%, and 52.3%, on an average for the DOAs of the target and disturbance source, respectively. The rest of this subsection deals with the results of the FBF, original ZBF, and 10°-wider ZBF in detail. Figure 7 shows that the ZBF reduced word errors of the FBF. The FBF could reduce the directional noise from the disturbance source located at the DOA of $|\theta_{\text{dst}}| > 45^\circ$ at least in principle. In contrast, the ZBF could reduce the directional noise even when the disturbance source was located at the DOA of $|\theta_{\text{dst}}| < 45^\circ$. In this experiment, the target source was not always located at a position directly in front of the robot (i.e., $\theta_{\text{trg}} = 0^\circ$). In this case, the directivity of the original ZBF might be too sharp to cope with the movement of the target source from the position directly in front of the robot. In fact, the 10°-wider ZBF gave better word accuracies than the original ZBF.

From the evaluation in terms of noise reduction (Fig. 8), the ZBF gave better NRRs than the FBF, irrespective of the DOAs of the disturbance source. In particular, the original ZBF gave the best NRRs, irrespective of the DOAs of the disturbance source. In this case, the noise reduction performance of the original ZBF was higher than the 10°-wider ZBF because the original ZBF had sharper directivity than the 10°-wider ZBF.

Instead, the original ZBF found it difficult to avoid more distortions included in the separated sounds as compared to the 10°-wider ZBF. This result was observed from the evaluation in terms of the speech quality based on the LSD (Fig. 9), i.e., the 10°-wider ZBF achieved lower LSDs than the original ZBF. In addition, 10°-wider ZBF achieved better speech quality (i.e., lower LSDs) than the FBF, irrespective of the DOAs of the disturbance source.

We mainly used the noise reduction systems in order to improve the performance of the speech recognition systems for robot audition. Therefore, from the above results, we can conclude that the 10°-wider ZBF is suitable for robot audition systems.

## 5. CONCLUSION

We developed a zonal beamformer, which enhances only the sound source located in the zonal space in front of the microphone array,

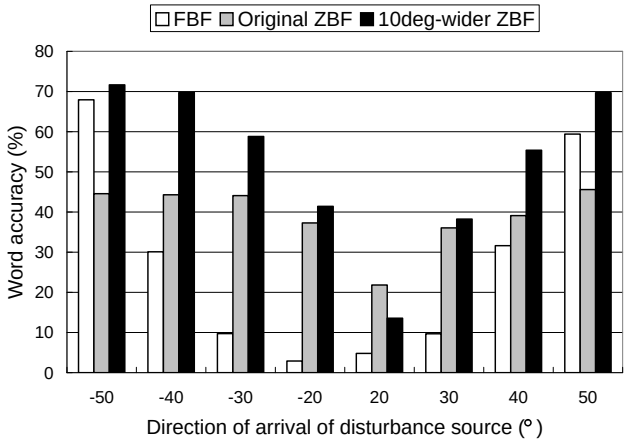



Figure 7: Word accuracy as a function of DOAs of a disturbance source. Each bar represents the average for five DOAs of a target source.

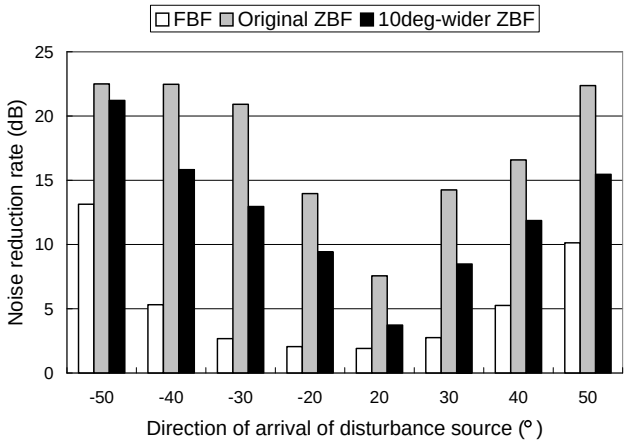



Figure 8: NRR as a function of DOAs of a disturbance source. Each bar represents the average of five DOAs of a target source.

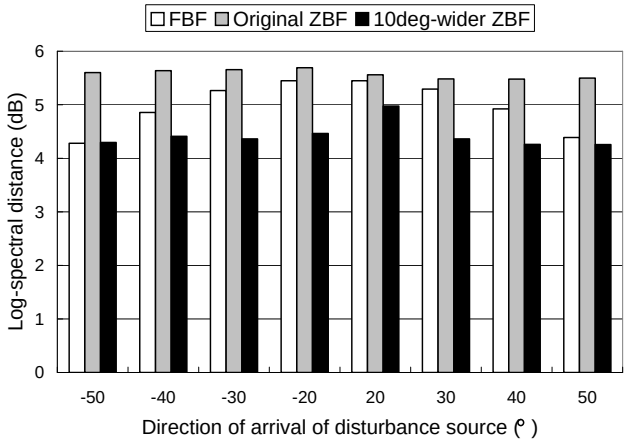



Figure 9: LSD as a function of DOAs of a disturbance source. Each bar represents the average of five DOAs of a target source.

and applied this beamformer to the preprocessing of speech recognition systems for a spoken dialog robot. Experimental results showed that this beamformer could better improve the performances of the reduction in the directional noise coming from behind the target source than a fan-like beamformer.

## REFERENCES


[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, pp. 4–24, 1988.

[2] M. S. Brandstein and D. B. Ward, *Microphone arrays: signal processing techniques and applications*, Springer-Verlag, Berlin, 2001.

[3] S. Makino *et al.*, *Blind speech separation*, Springer, 2007.

[4] S. Takada *et al.*, "Sound source separation using null-beamforming and spectral subtraction for mobile devices," *Proc. WASPAA2007*, Oct. 2007.

[5] T. Ogawa *et al.*, "Ears of the robot: noise reduction using four-line ultra-micro omni-directional microphones mounted on a robot head," *Proc. EUSIPCO2008*, Aug. 2008.

[6] K. Hosoya *et al.*, "Robot auditory system using head-mounted square microphone array," *Proc. IROS2009*, pp. 2736–2741, Oct. 2009.

[7] H. Saruwatari *et al.*, "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Applied Signal Process.*, vol. 2003, no. 11, pp. 1135–1146, Nov. 2003.

[8] J. Li *et al.*, "A two-microphone noise reduction method in highly non-stationary multiple-noise-source environments," *IEICE Trans. Fundamentals*, vol. 91E-A, no. 6, pp. 1337–1346, June 2008.

[9] Y. Matsuyama *et al.*, "Designing communication activation system in group communication," *Proc. Humanoids2008*, pp. 629–634, Dec. 2008.

[10] K. Itou *et al.*, "The design of the newspaper based Japanese large vocabulary continuous speech recognition corpus," *Proc. ICSLP1998*, pp. 3261–3264, Nov. 1998.