Acoust. Sci. & Tech. **34**, 2 (2013)

PAPER

Multiple-nulls-steering beamformer based on both talker and noise direction-of-arrival estimation

Masato Nakayama[1,2,*], Takanobu Nishiura[1], Yoichi Yamashita[1] and Noboru Nakasako[2]

[1]*College of Information Science and Engineering, Ritsumeikan University, 1–1–1 Noji Higashi, Kusatsu, 525–8577 Japan*
[2]*Faculty of Biology-Oriented Science and Technology, Kinki University, 930 Nishi-mitani, Kinokawa, Wakayama, 649–6493 Japan*

(*Received 24 May 2012, Accepted for publication 7 August 2012*)

Abstract: Beamforming with a microphone-array is an ideal candidate for distant-talking speech recognition. An adaptive beamformer can achieve beamforming with a small microphone-array, but it had difficulty extracting distant-moving speech and reducing moving noises, because it must rapidly train long multiple-channel adaptive filters by using observed noises with a microphone-array. However, if positions of both talkers and noises can be estimated, adaptive filters may not need to be trained in real noisy environments. Therefore, we propose a multiple-nulls-steering beamformer based on both talker and noise localization that does not require adaptive training with observed noises. Finally, we confirmed the validity and effectiveness of the proposed method through computer simulations and evaluation experiments in real noisy environments.

Keywords: Beamforming, Microphone-array, Distant-talking speech recognition, Talkers and noises localization, Multiple-nulls-steering beamformer

PACS number: 43.60.Fg, 43.60.Jn [doi:10.1250/ast.34.80]

1. INTRODUCTION

Teleconference systems and control systems with speech require a hands-free speech interface. There are two main types of conventional hands-free speech interfaces. One has a wearable microphone, such as a headset. The wearable microphone is able to observe clean speech at high-signal-to-noise ratio (high-SNR) in noisy environments. However, the wearable microphone forces users to bear the inconvenience of wearing it. On the other hand, an interface with a remote microphone is proposed as another hands-free speech interface. The remote microphone does not make users conscious of it. However, the remote microphone has a problem in that reverberations and ambient noises degrade recorded speech. To overcome this problem, the high-SNR technology using a microphone-array is focused on.

Beamforming with a microphone-array controls directivity by using the designed steering filters and can extract a clean speech from distant-talking speech. Delay-and-sum array [1] and adaptive array [2–5] have been proposed as typical beamformers. A delay-and-sum array can form a directivity to desired sound sources without adaptive training, but a large microphone-array with many transducers is required in order to achieve beamforming with a high directivity. Meanwhile, an adaptive array can form nulls to noise sound sources by using a small microphone-array, but this method has difficulty adapting to the motion of observation points, talkers, and noises. This is because an adaptive array requires adaptive training to a multi-channel steering filter for null beamforming [2]. Also, a subtractive array without adaptive training has been proposed, but this method cannot perform sufficiently in real environments because it forms a narrow null [4].

On the other hand, a cross spectral phase analysis (CSP) method [6] and an acoustic distance measurement (ADM) method based on interference [7,8] have been proposed as talker localization technology. The CSP method estimates the direction of arrival (DOA) of a sound source by using time delay between a sound source and a microphone-array. Therefore, the CSP method can localize the talker by localizing the sound source to talker speech, but this method has difficulty determining whether a sound source is talker speech or noise. Also, the CSP

*e-mail: mnaka@fc.ritsumei.ac.jp

method cannot estimate a talker position in non-speech segments. The ADM method searches for an object in a short range by transmitting and receiving a sound from transducers just like an active radar. Therefore, the ADM method can localize a talker by localizing an object to talkers, but talker speech becomes noise in this method. By combining the CSP method and the ADM method, both talkers and noises can be localized with high accuracy.

When both talker and noise DOAs can be estimated with high accuracy, it is not necessary to perform adaptive training to environmental noises because we can control desired directions and null directions by using results of DOA estimation. Therefore, in this paper, we propose a multiple-nulls-steering beamformer based on both talker and noise DOA estimation. In the proposed method, talker and noise DOA are estimated by combining the CSP and the ADM methods. In addition, nulls-steering filters are pre-designed by Griffith-Jim array [2] with estimated talker and noise DOA. The proposed method reduces noise by minimizing output power from multiple-nulls-steering filters. However, in minimizing output power without constraint, an output signal has a large distortion, because output signal includes speech and noise. Thus, the proposed method minimizes output power with constraint on the basis of average speech spectrum [5]. Finally, we confirm the validity and effectiveness of the proposed method through computer simulations and evaluation experiments in real environments.

2. CONVENTIONAL GRIFFITH-JIM ARRAY [2]

Griffith-Jim array is an adaptive array with the following constraint conditions:

$$F(\omega) = 1, \tag{1}$$

where $F(\omega)$ is frequency characteristics between a desired sound source and an array output. This constraint condition shows that a distortion of desired signal is not tolerated. Griffith-Jim array requires a talker DOA and environmental noises in order to perform adaptive training. This is because this method minimizes noise power of array output under the condition that an input signal from a talker DOA is kept in a distortion-free condition. However, the error between an environment of adaptive training and an environment of beamforming arises as a distortion of array output. In addition, this method has difficulty adapting to the motion of observation points, talkers, and noises because it requires real-time design of multi-channel adaptive filters.

3. PROPOSED METHOD

A conventional adaptive array was able to perform well even in the case of a small microphone-array but had difficulty adapting to the motion of observation points, talkers, and noises. In this paper, therefore, we propose a multiple-nulls-steering beamformer, which is robust to the motion of observation points, talkers, and noises, based on both talkers and noises DOA estimation. In this section, we describe the algorithm of the proposed method. Figure 1 shows a flowchart of the proposed method.

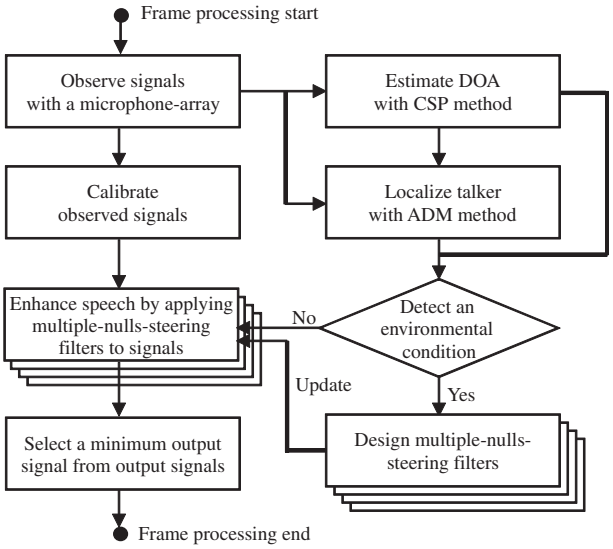


Fig. 1 Flowchart of the proposed method.

As shown in Fig. 1, in the first step of the proposed method, spatial information of real environments is obtained. DOA estimation is achieved by a CSP method, and the talker is localized by both CSP and ADM methods. By combining the CSP and ADM methods, we can determine whether the estimated DOAs are talker speech or noise. DOA is estimated and the talker is localized frame by frame. These processing are shown in Fig. 1 as "Estimate DOA with CSP method" and "Localize talker with ADM method."

From results of the estimated DOAs at each frame, the proposed method detects whether environmental conditions vary or not. If environmental conditions are varying or initializing, then multiple-null-steering-filters are designed by computer simulation on the basis of both talker and noise DOAs and those mobility predictions. Also, computer simulation for the multiple-null-steering-filters are performed on the basis of an algorithm of a Griffith-Jim array. The design method for the multiple-null-steering-filters is described in Sect. 3.1. These processes are shown in Fig. 1 as "Detect an environmental condition" and "Design multiple-null-steering-filters."

The proposed method requires calibration to be performed to measure the system because it utilizes multiple-nulls-steering filters pre-designed by computer simulation. Thus, variation between each microphone of a microphone-array is calibrated as follows:

$$\hat{S}_i(\omega) = \frac{\sigma_{g_1}}{\sigma_{g_i}} S_i(\omega), \tag{2}$$

where ω [Hz] is frequency, $\hat{S}_i(\omega)$ is i-th channel spectrum of observed signal after calibration ($i = 1, \ldots, M$), $S_i(\omega)$ is i-th channel spectrum of observed signal before calibration, and σ_{g_i} is i-th channel gain calculated from direct wave of impulse response. This processes is shown in Fig. 1 as "Calibrate observed signals." The direct wave of impulse response is calculated with the maximum peak detection.

By applying pre-designed multiple-nulls-steering filters to calibrated array signals, the multiple output signals are obtained as follows:

$$Y_j(\omega) = \sum_{i=1}^{M} \hat{S}_i(\omega) H_{ij}(\omega), \tag{3}$$

where $Y_j(\omega)$ is an output signal from j-th nulls-steering filter ($j = 1, \ldots, N$), and $H_{ij}(\omega)$ is i-th channel filter in j-th nulls-steering filter. To select suitable nulls-steering filter, the proposed method selects minimum output signal from N output signals. However, in minimizing output power without constraint, an output signal has a large distortion, because output signal includes speech and noise. Thus, the proposed method has an algorithm for minimizing output power with constraint on the basis of average speech spectrum. Thus, the proposed method minimizes output power with the weight function on the basis of average speech spectrum [5] as follows:

$$Y(\omega) = \frac{1}{W(\omega)} \min_{j=1}^{N} W(\omega) Y_j(\omega), \tag{4}$$

where $W(\omega)$ is weight function based on average speech spectrum. The design method of weight function is described in Sect. 3.2.

The proposed method is performed by frame processing. Thus, the computational cost of the proposed method seems very expensive because DOA is estimated and the talker is localized at each frame. However, this computational cost is not expensive because CSP and ADM methods can be achieved by a computational cost double that of Fourier transform. Also, the conventional beamformer also requires both DOA estimation and talker localization. In addition, multiple-nulls-steering filters can be applied with low computational cost because the observed array signals are transformed into frequency domain.

3.1. Design of Nulls-Steering Filters

In this section, we describe the design method for nulls-steering filters. In the proposed method, the nulls-steering filters are designed by the algorithm of Griffith-Jim array. Griffith-Jim array requires the desired DOA and the environmental noise that is not including a desired speech in order to design nulls-steering filters. The desired DOA is estimated by CSP and ADM methods. Also, the environmental noise $\boldsymbol{N}(\omega)$ is created by using the noise DOA, which is estimated by CSP method, as follows:

$$\begin{aligned}\boldsymbol{N}(\omega) = {} & \frac{1}{\sigma_d} \sum_k \boldsymbol{D}_k(\omega) \boldsymbol{X}_k(\omega) \\ & + \frac{A}{\sigma_a} \sum_{k=0}^{(\theta_s-\beta)/\alpha} \boldsymbol{D}_{\alpha \cdot k}(\omega) \boldsymbol{Y}_{\alpha \cdot k}(\omega) \\ & + \frac{A}{\sigma_a} \sum_{k=(\theta_s+\beta)/\alpha}^{180/\alpha} \boldsymbol{D}_{\alpha \cdot k}(\omega) \boldsymbol{Y}_{\alpha \cdot k}(\omega),\end{aligned} \tag{5}$$

where $\boldsymbol{N}(\omega)$ is the environmental noise, θ_s [degrees] is the desired DOA, k [degrees] is the noise DOA, $\boldsymbol{D}_k(\omega)$ is the spectrum of delay filter corresponding to the noise DOA, $\boldsymbol{X}_k(\omega), \boldsymbol{Y}_k(\omega)$ are spectra of white Gaussian noises that do not correlate with each other, σ_d is a standard deviation of the directional noise, σ_a is a standard deviation of the omni-directional ambient noise, A is gain coefficient of the omni-directional ambient noise, α [degrees] is step size of angles in designing the omni-directional ambient noise, and β [degrees] is range of angles in designing the omni-directional ambient noise. Equation (5) shows that the omni-directional ambient noise does not exist in the range between $\theta_s - \beta$ and $\theta_s + \beta$ degrees.

In the case of multiple talkers, the nulls-steering filters can be designed by changing θ_s in Eq. (5) as the desired DOA of each talker. To simulate all noisy environments, the proposed method requires a large computational cost. However, assuming that a number of the simultaneously arriving noises at each frame is one or few, we can narrow down the noisy environments to design nulls-steering filters. Also, we can design multiple nulls-steering filters by using fast servers in advance. In addition, by sending and receiving data online, we can design nulls-steering filters by using external servers because nulls-steering filters have small data size.

However, the proposed method requires larger memory than the conventional method. The proposed method requires 1-MByte memory under the conditions in which the number of speech sources is 1, the number of noises is 1, the desired DOA moves from 0 to 90 degrees, the noise DOA moves from 90 to 180 degrees, step size of angles is 5 degrees, the number of microphones is 4 channels, the filter length is 100 tap, and the float is 8 Bytes.

3.2. Design of Weight Function

An average speech spectrum and a weight function are calculated by using a speech database with manual labeling as follows:

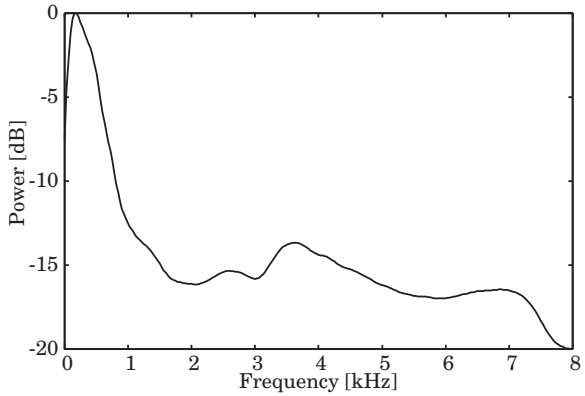


Fig. 2 Average speech spectrum.

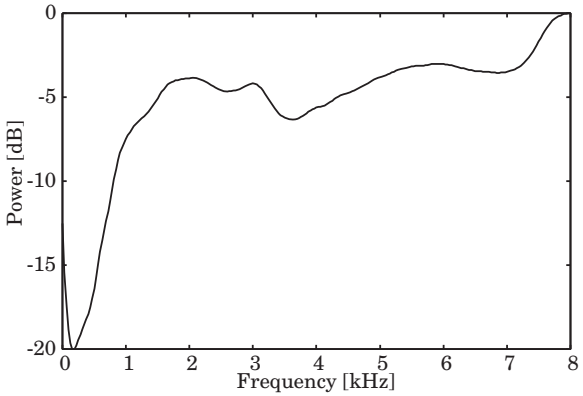


Fig. 3 Weight function.

$$|ASP(\omega)| = \sum_{l=1}^{L} \frac{1}{N_l} \sum_{n=1}^{N_l} |SP_l(\omega; n)| + g_{\mathrm{a}}, \tag{6}$$

$$W(\omega) = \frac{1}{|ASP(\omega)|}, \tag{7}$$

where L is amount of speech (words), N_l is number of l-th speech frames, $SP_l(\omega; n)$ is spectrum of speech signal, $ASP(\omega)$ is average speech spectrum, $W(\omega)$ is weight function, and g_{a} is a gain adjustment parameter. A gain adjustment parameter g_{a} adjusts an average speech spectrum to a specified range on a log-spectrum. An average speech spectrum is normalized by this adjustment.

In this paper, we calculated the average speech spectrum and weight function by using JNAS database [9] with manual phoneme labeling. Figures 2 and 3 show an average speech spectrum and weight function, respectively. The average speech spectrum and weight function are calculated under the conditions shown in Table 1. Also, a gain adjustment parameter was adaptively calculated to adjust difference between maximum and minimum spectrum into 20 dB. The weight function can reduce a distortion of the output desired speech.

Table 1 Design conditions for average speech spectrum.

Speech database	JNAS-DB [9]
Phoneme addition	30000 frames
Frame length	32 ms (Humming window)
Frame interval	8 ms
Sampling frequency	16 kHz
Quantization	16 bits

Table 2 Simulation conditions.

Parameter settings	
Sampling frequency	16 kHz
Quantization	16 bits
Frame length	32 ms (Hanning window)
Frame interval	16 ms
Filter design conditions	
Adaptive training	NLMS (Step size: 0.1)
Filter tap	101
Talker DOA	$\theta_{\mathrm{s}} = 140$ degrees
Noise source	White Gaussian noise
Noise DOA	35, 70, 90, 105 degrees
Ambient noise	$A = 0.3$, $\alpha = 5$, $\beta = 15$
Test data (open)	
Desired speech source	Phoneme balanced 216 words in ATR-DB [10] (4 males and 4 females)
Noise source	White Gaussian noise
SNR	0, 5, 10, 15, 20 dB
Acoustic model (HMM)	IPA speaker-independent clean monophone model [11]

4. COMPUTER SIMULATIONS

To confirm the validity of proposed method, we performed computer simulations to examine the proposed method at various SNRs in noisy environments.

4.1. Simulation Conditions

Table 1 shows design conditions of an average speech spectrum and a weight function, Table 2 shows simulation conditions, and Fig. 4 shows a simulation environment. In this paper, we assume that the desired and noise DOAs were known. To evaluate the proposed method, computer simulations and evaluation experiments using moving noises or moving talkers may be required. However, moving noises and moving talkers have difficulty in keeping repeatability. Thus, in this computer simulation, noise sources have four positions, and a noise source is switched from noise source 1 to noise source 4 at 0.2 s intervals as shown in Fig. 4. In other words, we simulate the conditions in which one noise source moves and one desired speech source does not. For the conventional method, we evaluated Griffith-Jim array, which is referred to as the conventional GJ, under the conditions of using only noisy speech without using the noise segments for adaptive filter. Adaptive training of the conventional GJ is

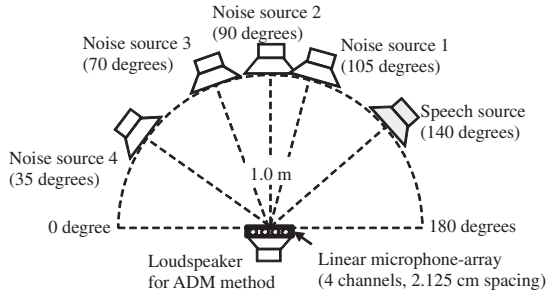


Fig. 4 Simulation environment.

achieved with the normalized least mean square (NLMS). In addition, an output signal of the proposed method is obtained by selecting minimum output signals from multiple-nulls-steering filters. Thus, to evaluate each output, we also evaluate outputs of each nulls-steering filter. Output signals of each steering filters are as follows:

NSF-030 Output of null-steering filter for 30 degrees
NSF-070 Output of null-steering filter for 70 degrees
NSF-090 Output of null-steering filter for 90 degrees
NSF-105 Output of null-steering filter for 105 degrees
NSF-AN Output of null-steering filter for the omni-directional ambient noise

Especially, in this simulation, results of the proposed method show the upper limit performance in the proposed method. In the proposed method, DOA estimation results are utilized to reduce the number of the nulls-steering filters. This is because the proposed method calculates the array output signal by minimizing filter output signals with constraint. Therefore, when various nulls-steering filters are designed in advance, we can obtain results similar to this simulation. However, when various nulls-steering filters are not sufficiently designed, the performance of the proposed method is degraded. Thus, we evaluate NSF-AN which forms nulls to omni-directions except the desired DOA. NSF-AN shows the lower limit performance in the proposed method because this is not affected by performance of DOA estimation.

4.2. Evaluation Index

In this paper, we evaluated the noise reduction rate (NRR), the spectral distortion measure (SD), and the performance of automatic speech recognition (ASR).

NRR is defined as follows:

$$\mathrm{NRR} = 10\log_{10}\frac{\sigma_{\mathrm{i}}^2}{\sigma_{\mathrm{o}}^2}, \tag{8}$$

where σ_{i}^2 is variance of an input noise before noise reduction, and σ_{o}^2 is variance of an output noise after noise reduction.

Distortion of an output speech after noise reduction is evaluated by the SD [5,12] as follows:

$$\mathrm{SD} = \sqrt{\frac{10^2}{N}\sum_{\omega=1}^{N}(\log_{10}|S(\omega)| - \log_{10}|S'(\omega)|)^2}, \tag{9}$$

where N is the number of maximum samples, SD is a spectral distortion measure, $S(\omega)$ is a clean speech spectrum, and $S'(\omega)$ is speech spectrum after beamforming.

ASR is evaluated by the word recognition rate (WRR) of speech recognition using an IPA speaker-independent clean monophone model and word dictionary. Julius [11] is used as a speech recognition engine. In a computer simulation, the WRR of observed speech without noise is 96%.

4.3. Directionality of Nulls-Steering Filters

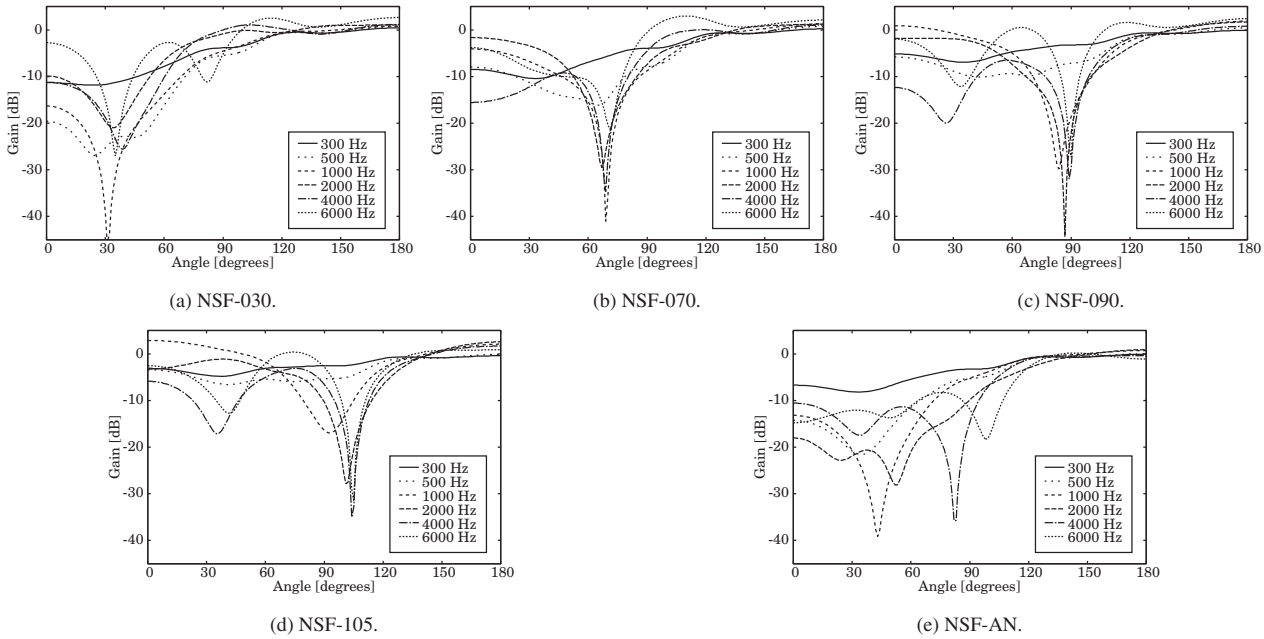
Figure 5 shows directionality patterns of nulls-steering filters. Figures 5(a) through 5(e) are directionality patterns of nulls-steering filters for 35 degrees, 70 degrees, 90 degrees, 105 degrees, and the omni-directions, respectively. As a result of Fig. 5, in forming nulls for the specific direction, noise reduction is not sufficient except at the specific direction.

4.4. Simulation Results

Figures 6(a) and 6(b) show waveforms of clean and noisy observed speech in simulations for male talkers at SNR = 0 dB, respectively. Figures 7(a) through 7(f) show waveforms of output signals to the observed signal shown in Fig. 6(b). Figure 7(a) shows a waveform of output signal with the conventional GJ, Figs. 7(b) through 7(f) show waveforms of output signals with steering filters of NSF-035 through NSF-AN, and Fig. 7(g) shows a waveform of output signal with the proposed method. Figure 7(a) shows the conventional GJ was not able to perform sufficiently because it requires a long training time for adaptive training. Results is Figs. 7(b) through 7(e) show noise cannot be reduced sufficiently under the conditions in which a noise arrives from directions except the direction to form nulls. Figure 7(f) shows noise cannot be reduced compared with the proposed method. Figure 7(g) shows the proposed method was able to reduce noise sufficiently under these simulation conditions.

Figure 8 shows the results of NRR on simulations. NRR omits the SNR axis because NRR performs the same way in each SNR environment. Figure 8 shows the proposed method was able to improve about 7 dB at NRR compared to the conventional GJ.

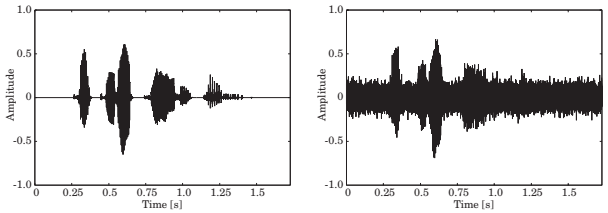
Figure 9 shows a results of SD on simulations in each SNR environment. Figure 9 shows the proposed method tended to improve SD in lower SNR environments and was able to improve SD up to 2.5 dB compared with the conventional GJ. Also, the proposed method had higher SD than the conventional GJ in all SNR environments.



(a) NSF-030. (b) NSF-070. (c) NSF-090.

(d) NSF-105. (e) NSF-AN.

Fig. 5 Directionality patterns of nulls-steering filters.



(a) Clean speech. (b) Noisy speech.

Fig. 6 Waveforms of clean and noisy speech on simulations.

Figure 10 shows WRR of each method on simulations at each SNR environment. Figure 10 shows the proposed method tended to improve WRR in lower SNR environments as in the case with NRR and was able to improve WRR up to 50% compared with the conventional GJ.

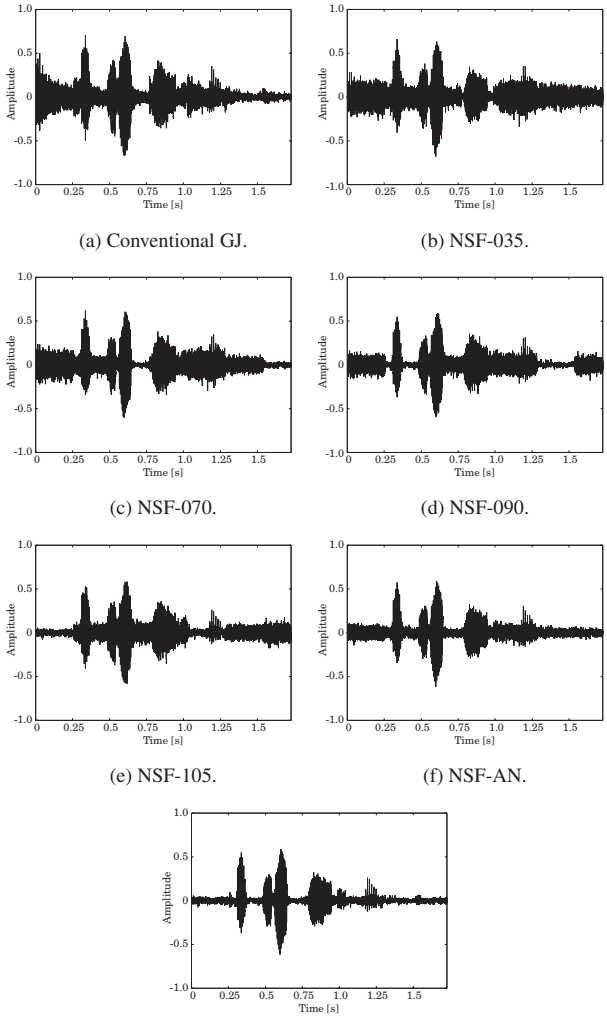
Accordingly, all simulation results confirmed the validity of the proposed method because the proposed method was able to improve NRR, SD, and WRR compared with the conventional GJ.

5. EVALUATION EXPERIMENTS

In evaluation experiments, we evaluated NRR, SD, and ASR the same way as in computer simulations.

5.1. Experimental Conditions

Experimental environments and experimental conditions, design conditions for average speech spectrum, and weight function are similar to simulation conditions in Fig. 4, Table 1, and Table 2. Also, most experimental conditions are the same as those in the computer



(a) Conventional GJ. (b) NSF-035.

(c) NSF-070. (d) NSF-090.

(e) NSF-105. (f) NSF-AN.

(g) Proposed method.

Fig. 7 Waveforms of output signals on simulations.

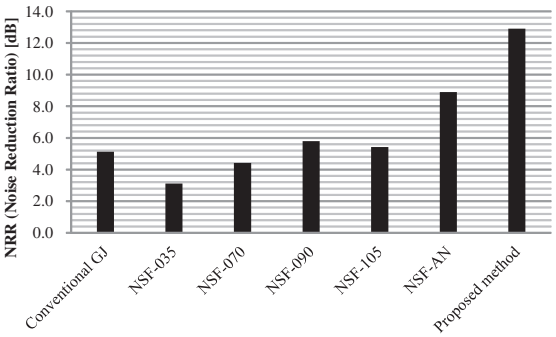


Fig. 8 Simulation results of NRR.

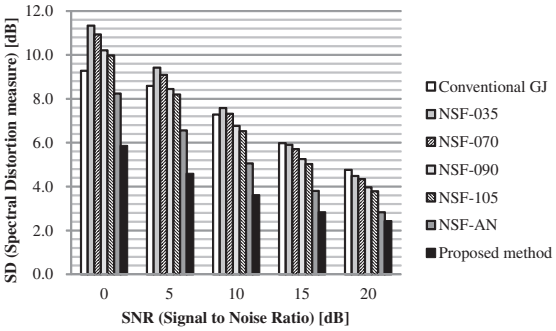


Fig. 9 Simulation results of SD.

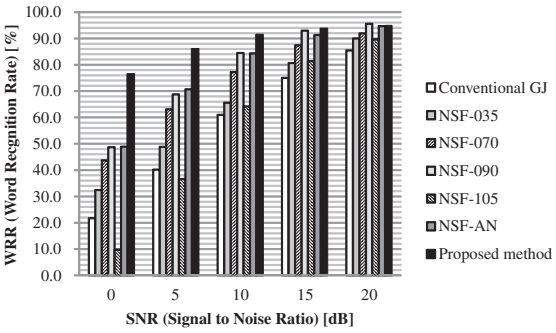
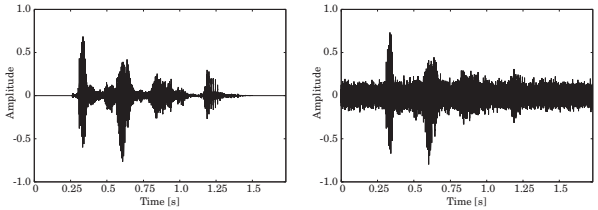


Fig. 10 Simulation results of ASR.

simulation. However, room impulse response was measured by the time stretched pulse (TSP) method [13]. Room reverberation was $T_{[60]} = 0.7$ s, and ambient noise level was $L_{\mathrm{A}} = 28$ dB. In an evaluation experiment, the WRR of observed speech without noise was 88%.

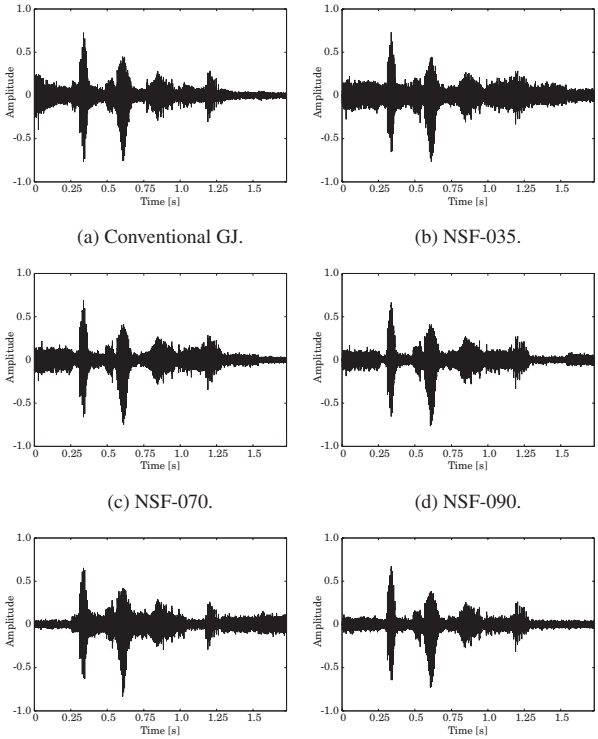
5.2. Experimental Results

Figures 11(a) and 11(b) show waveforms of clean and noisy observed speech in experiments for male talkers at SNR = 0 dB, respectively. Figures 12(a) through 12(g) show waveforms of output signals to the observed signal shown in Fig. 11(b). Figure 12(a) shows a waveform of output signal with the conventional GJ, Figs. 12(b) through 12(f) show waveforms of output signals with steering filters



(a) Clean speech. (b) Noisy speech.

Fig. 11 Waveforms of clean and noisy speech in real environments.



(a) Conventional GJ. (b) NSF-035.

(c) NSF-070. (d) NSF-090.

(e) NSF-105. (f) NSF-AN.

(g) Proposed method.

Fig. 12 Waveforms of output signals in real environments.

of NSF-035 through NSF-AN, and Fig. 12(g) shows a waveform of output signal with the proposed method. Figures 12(a) through 12(f) shows experimental results of waveforms that tend to be similar to those of computer simulations. Thus, from these waveforms, we confirmed that the proposed method was superior to the conventional GJ.

Figure 13 shows the results of NRR in experiments. NRR omits SNR because it performs the same way in

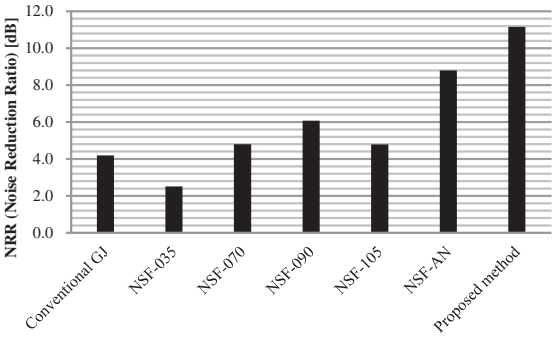


Fig. 13 Experimental results of NRR.

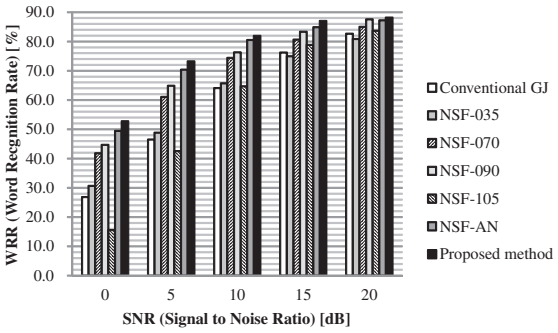


Fig. 15 Experimental results of ASR.

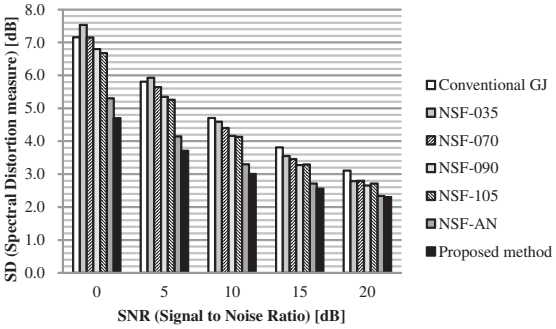


Fig. 14 Experimental results of SD.

each SNR environment as in the case with simulations. Figure 13 shows the proposed method was able to improve about 7 dB at NRR compared with the conventional GJ as in the case with simulations.

Figure 14 shows results of SD on experiments in each SNR environment. Figure 14 shows experimental results of waveforms tend to be similar to those of computer simulations and the proposed method was able to improve SD up to 2.5 dB compared with the conventional GJ as in the case with simulations. Thus, we confirmed that the proposed method was superior to the conventional GJ in terms of SD.

Figure 15 shows WRR of each method in experiments at each SNR. Figure 15 shows experimental results of waveforms tend to be similar to those of computer simulations and the proposed method was able to improve WRR up to 25% at the condition of SNR = 0 dB compared with the conventional GJ as in the case with simulations.

Accordingly, all experimental results confirmed the effectiveness of the proposed method because the proposed method was able to improve NRR, SD, and WRR compared with the conventional GJ.

6. CONCLUSIONS

In this paper, we proposed the multiple nulls-steering beamformer based on positional information by talker and noise localization and their mobility predictions. We have confirmed the validity and effectiveness of the proposed method through computer simulations and evaluation experiments.

ACKNOWLEDGMENTS

The present study was supported in part by a Grant-in-Aid for Scientific Research (C) 23500233, 24560533, and 24700126 from JSPS.

REFERENCES

[1] 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.*, **78**, 1508–1518 (1985).

[2] L. J. Griffith and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, **AP-23**, 27–34 (1982).

[3] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP-34**, 1391–1400 (1986).

[4] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Signal Processing Series (Prentice Hall, Englewood Cliffs, 1993).

[5] M. Nakayama, T. Nishiura and Y. Yamashita, "Noisy speech recognition using adaptive microphone-array based on vowel/consonant-based features," *IEICE Trans. D*, **J92-D**, 1568–1578 (2009) (in Japanese).

[6] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.*, **5**, 288–292 (1997).

[7] M. Nakayama, N. Nakasako, T. Shinohara and T. Uebo, "Talker localization based on interference between transmitted and reflected audible sound," *IEEJ Trans. Electron. Inf. Syst.*, **130-C**, 1994–2000 (2010) (in Japanese).

[8] M. Nakayama, S. Hanabusa, T. Uebo and N. Nakasako, "Acoustic distance measurement method based on phase interference using calibration and whitening processing in real environments," *IEICE Trans. A*, **E94-A**, 1638–1646 (2011).

[9] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi and K. Shikano, "JNAS: Japanese speechcorpus for large vocabulary continuous speech recognition research," *J. Acoust. Soc. Jpn.* (*E*), **20**, 199–206 (1999).

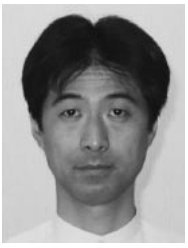
[10] K. Takeda, Y. Sagisaka and S. Katagiri, "Acoustic-phonetic labels in a Japanese speech database," *Proc. Eur. Conf. Speech Technology*, Vol. 2, pp. 13–16 (1987).

[11] A. Lee, T. Kawahara and K. Shikano, "JULIUS — An open source real-time large vocabulary recognition engine," *EUROSPEECH 2001*, pp. 1691–1694 (2001).

[12] S. Furui, *Digital Speech Processing, Synthesis, and Recognition* (Marcel Dekker, New York, 2001).
[13] Y. Suzuki, F. Asano, H. Y. Kim and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am.*, **97**, 1119–1123 (1995).

Masato Nakayama received his B.E. degree from Kinki University in 2001, his M.E. degree from Wakayama University in 2003 and his Dr.Eng. degree from Ritsumeikan University in 2010. He is currently a research associate at College of Information Science and Engineering, Ritsumeikan University, and a researcher at the Faculty of Biology-Oriented Science and Technology, Kinki University. His current research interests include array signal processing, acoustic distance measurement, ultrasonic signal processing, and sound field reproduction. He is a member of IEICE.

Takanobu Nishiura received his B.E. degree from the Nara National College of Technology in 1997 and M.E. and Ph.D. degrees from the Nara Institute of Science and Technology (NAIST) in 1999 and 2001, respectively. From 2001 to 2004, he was a research associate at Wakayama University. He is currently an associate professor at Ritsumeikan University. His current research interests include acoustic sound signal sensor using a microphone array. He received the TELECOM System Technology Award for Students from the Telecommunications Advancement Foundation (TAF) in 2000, and the Best Paper Award from the Virtual Reality Society of Japan (VRSJ) in 2009. He is a member of IEICE, IPSJ, VRSJ and INCE.



Yoichi Yamashita received his B.E., M.E. and Dr.Eng. degrees from Osaka University in 1982, 1984 and 1993, respectively. He has worked for the Institute of Scientific and Industrial Research of Osaka University as a Technical Official, a Research Associate, and an Assistant Professor from 1984 to 1997. In 1997, he joined Ritsumeikan University as an Associate Professor in the College of Science and Engineering. He is currently a Professor in the College of Information Science and Engineering. His research interests include speech understanding, speech synthesis, speech communication, and spoken document processing. He is a member of IEICE, IPSJ, JSAI, ISCA, and IEEE.

Noboru Nakasako received his B.E., M.E., and Dr.Eng. degrees from Hiroshima University in 1982, 1984, and 1990, respectively. He served as a research assistant and an assistant professor in the Department of Electrical Engineering, the Hiroshima Institute of Technology. He joined the Faculty of Biology-Oriented Science and Technology, Kinki University as an associate professor. He has been a professor at Kinki University since 2002. His research interests include acoustic signal processing, sound and vibration control, and independent component analysis. He is a member of IEICE, SICE, ISCIE, and INCE Japan.