SUITABLE DESIGN OF ADAPTIVE BEAMFORMER BASED ON AVERAGE SPEECH SPECTRUM FOR NOISY SPEECH RECOGNITION

Takanobu Nishiura[†,‡], *Satoshi Nakamura*[‡], *Yuka Okada*[§], *Takeshi Yamada*[*], *and Kiyohiro Shikano*[§]

[†] Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama, 640-8510 Japan
[‡] ATR Spoken Language Translation Research Laboratories
[§] Graduate School of Information Science, Nara Institute of Science and Technology
[*] Institute of Information Sciences and Electronics, University of Tsukuba

ABSTRACT

Recognition of distant-talking speech is indispensable for self-moving robots or tele-conference systems. However, background noise and room reverberations seriously degrade the sound capture quality in real acoustic environments. A microphone array is an ideal candidate as an effective method for capturing distant-talking speech. AMNOR (Adaptive Microphone-array for NOise Reduction) was proposed an adaptive beamformer for capturing the desired distant signals in noisy environments by Kaneda et al. Although AMNOR has proven itself effective, it could be further improved if we knew the spectrum characteristics of desired distant signals in advance. Therefore, in this paper we regard speech as a desired distant signal and design AMNOR based on the average speech spectrum for distant-talking speech capture and recognition. As a result of evaluation experiments in real acoustic environments, we could confirm that the ASR (Automatic Speech Recognition) performance was improved 5 $\sim$ 10% by AMNOR based on average speech spectrum in noisy environments.

1. INTRODUCTION

It is very important for the natural interfaces of machines like self-moving robots to capture and recognize distant-talking speech with high accuracy. However, background noise and room reverberations seriously degrade the sound capture quality in the real acoustic environments. A microphone array is an ideal candidate for capturing distant-talking speech. With a microphone array, a desired speech signal can be acquired selectively by steering the directivity. Accordingly, super-high directivity is necessary to reduce noise signals.

To form directivity, delay-and-sum beamformers [1, 2] and adaptive beamformers [3, 4] have been proposed as conventional beamformers. A delay-and-sum beamformer forms super-high directivity to the desired signal, and an adaptive beamformer forms null directivity to the noise signal. However, delay-and-sum beamformers have two serious drawbacks: the performance is not good enough to capture the desired signal without a sufficient number of transducers, and performance degrades in highly reverberant rooms. On the other hand, adaptive beamformers can form null directivity with a small number of transducers. Furthermore, they can form sharper directivity than delay-and-sum beamformer. Consequently, adaptive beamformers are often used for the front-end processing of ASR (Automatic Speech Recognition) [5].

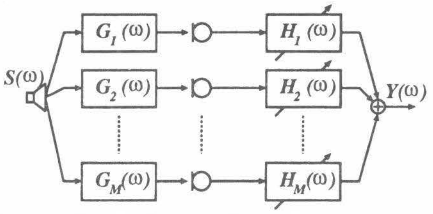


Figure 1: Block diagram of adaptive beamformer.

AMNOR (Adaptive Microphone-array for NOise Reduction) [4] is an adaptive beamformer proposed by Kaneda et al. in 1986. AMNOR is an effective beamformer for capturing and recognizing desired distant signals in noisy environments. Also, it can be easily designed with an adaptive filter for noise reduction in real environments, because it only allows small distortion for capturing the desired distant signal.

However, if we knew the spectrum characteristics of desired distant signals when designing the adaptive filter of AMNOR, its performance could be further improved. The conventional AMNOR is designed to suppress the spectrum distortion of the desired distant signal on all frequency bands, but in many cases, the purpose of signal capture is limited to speech capture. Therefore, in this paper we regard speech as the desired distant signal and design AMNOR by using the speech spectrum for distant-talking speech capture and recognition.

2. AMNOR (ADAPTIVE MICROPHONE-ARRAY FOR NOISE REDUCTION)

Figure 1 shows a block diagram of the adaptive beamformer. In Figure 1, $S(\omega)$ is the Fourier transform of the desired signal and $Y(\omega)$ is the Fourier transform of the output signal. $G_m(\omega)$ is the acoustic transfer function from the desired sound source to the m-th microphone element and $H_m(\omega)$ is the frequency response of the m-th filter. The frequency response $F(\omega)$ of the adaptive beamformer to the desired signal is represented as

$$F(\omega) = \sum_{m=1}^{M} G_m(\omega)H_m(\omega), \qquad (1)$$

where M is the number of microphone elements. The concept of the adaptive beamformer is to minimize the output noise energy

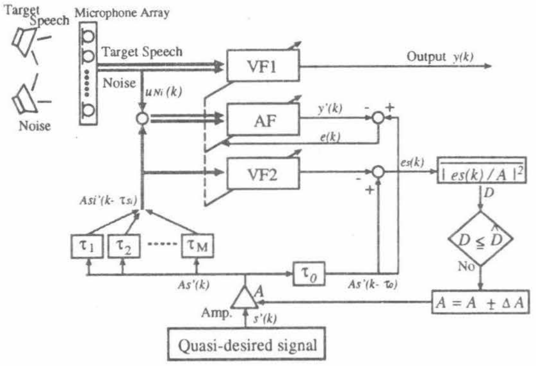


Figure 2: Overview of AMNOR.

while constraining $F(\omega)$ to the desired frequency response. AMNOR [4] has the constraint shown in Equation (2):

$$D = \int |1 - F(\omega)|^2 d\omega \ \le \hat{D}. \tag{2}$$

This constraint attains maximum noise reduction while allowing a small distortion D in the frequency response to the desired signal. In this paper, we focus on suitable control of the admissible distortion $\hat{D}$ in the frequency response for noisy speech recognition. Figure 2 shows a general overview of AMNOR. In Figure 2, each VF1, AF, and VF2 is a FIR filter with M-input and 1-output. AF is the adaptive filter, and VF1 and VF2 are variable filters that have the same filter coefficients as AF. A quasi-desired signal $s'(k)$ is indispensable for designing the adaptive filter of AMNOR because AMNOR attains maximum noise reduction with a quasi-desired signal and an environmental noise signal from the environment. The quasi-desired signal $s'(k)$ derives $As_i'(k - \tau_{si})$ from amplifier and time delay $\tau_{si}, i = 1, \ldots, M$, which is calculated subject to the known desired sound source's DOA (Direction Of Arrival). This situation assumes the simulation where signal $As'(k)$ arrives from the desired sound source with known DOA to the microphone array. In addition, the microphone only captures the noise signal $u_{Ni}(k), i = 1, \ldots, M$ (not including the desired signal), and it is inputted in the adaptive filter AF after adding it to quasi-desired signal $As_i'(k - \tau_{si})$. AF controls the filter coefficients based on $e(k)$ as the following Equation (3).

$$e(k) = As'(k - \tau_0) - y'(k), \tag{3}$$

where τ_0 is the constant delay for cause and effect. $es(k)$ is calculated by using VF2 after designing the filter coefficients by AF, and current distortion D is derived from Equation (4).

$$D = \overline{|\ es(k)/A\ |^2}. \tag{4}$$

By comparing current distortion D and admissible distortion $\hat{D}$, amplitude A is renewed with the amplifier until $D \le \hat{D}$. In the above algorithm, AMNOR attains higher noise reduction performance in real acoustic environments.

3. SUITABLE DESIGN OF AMNOR BASED ON AVERAGE SPEECH SPECTRUM

The conventional AMNOR uses a white Gaussian signal that has flat frequency characteristics as a quasi-desired signal in order to suppress the spectrum distortion of the desired signal on all frequency bands. But in many cases, the purpose of signal capture is limited to speech capture. Therefore, if we knew the spectrum characteristics of desired distant signals in advance, it may be possible to improve the performance of AMNOR by designing a suitable adaptive filter for the environment. In this paper, we regard speech as the desired distant signal and design AMNOR by using the speech spectrum for distant talking speech capture and recognition. First, we calculate the average speech spectrum weight by Equation (5).

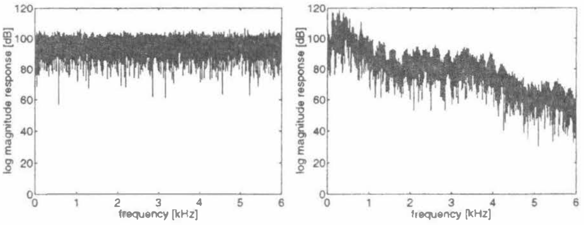
$$W_{sp}(\omega) = \frac{1}{L \cdot N} \sum_{l=1}^{L} \sum_{n=1}^{N} SP_l(\omega; n), \tag{5}$$

where L represents the number of speech (words), N represents the number of frames, $SP_l(\omega; n)$ represents the Fourier transform of speech signed $sp_l(t)$, and $W_{sp}(\omega)$ represents the average speech spectrum weight. The quasi-desired signal based on the average speech spectrum is derived from weighting the white Gaussian spectrum with the average speech spectrum weight $W_{sp}(\omega)$. Figure 3 shows the spectrum of white Gaussian as the quasi-desired spectrum for the conventional AMNOR and the spectrum of average speech weighted as quasi-desired spectrum for the proposed AMNOR. Compared with the spectra in Figure 3, the average speech weighted spectrum is enhanced at lower frequencies. We attempted to improve the ASR performance by using the average speech spectrum weighted quasi-desired signal for AMNOR, and this modified system was named S-AMNOR.

In addition, we also investigated whether the average speech spectrum weight is normalized to keep the energy ratio equivalent between vowels and consonants on each frame when estimating $W_{sp}(\omega)$ in Equation (5). We further consider a new spectrum weight defined by Equation (6). This weight is capable to balance the occurrence of vowel and consonant frames.

$$\begin{aligned} W_{sp}(\omega) \quad = \quad & \frac{1}{2}\Bigg(\frac{1}{L_c} \sum_{l_c=1}^{L_c} \frac{1}{N_{l_c}} \sum_{n=1}^{N_{l_c}} SP_{l_c}(\omega; n) \\ & + \frac{1}{L_v} \sum_{l_v=1}^{L_v} \frac{1}{N_{l_v}} \sum_{n=1}^{N_{l_v}} SP_{l_v}(\omega; n) \Bigg), \end{aligned} \tag{6}$$

where L_c represents the number of vowels, L_c represents the number of consonants, N_{l_v} represents the number of vowel frame on each speech (word), and N_{l_c} represents the number of consonant frame on each speech (word). The system using this modified W_{sp} was named Normalized S-AMNOR.



(a) White Gaussian spectrum

(b) Average speech weighted spectrum

Figure 3: Spectrum of quasi-desired signal.

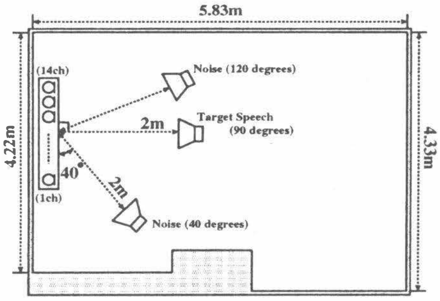


Figure 4: Experimental environment.

Table 1: Experimental conditions

Recording conditions	
Reverberation time	$T_{[60]}$=180 msec.
Microphone array	Linear type 14 transducers, 2.83 cm spacing
Sampling frequency	12 kHz (Quantization: 16 bit)
Experimental conditions for ASR	
Frame length	32 msec. (Frame interval: 8 msec.)
HMM	Gaussian mixture density(3 states)
Feature vector	MFCC (16 orders, 4 mixtures), ΔMFCC (16 orders, 4 mixtures), Δpower (1 order, 2 mixtures)
Average speech spectrum weight	ATR speech DB SetA [6] (2620 words × 4 subjects) and ASJ continuous speech corpus [7] (150 sentenses × 64 subjects)
Test data (Open)	
Desired speech signal	Speech: 216 words × 2 subjects (1 female and 1 male)
Noise signal	Female speech, male speech or white Gaussian noise
SNR	3 dB

4. EVALUATION EXPERIMENTS

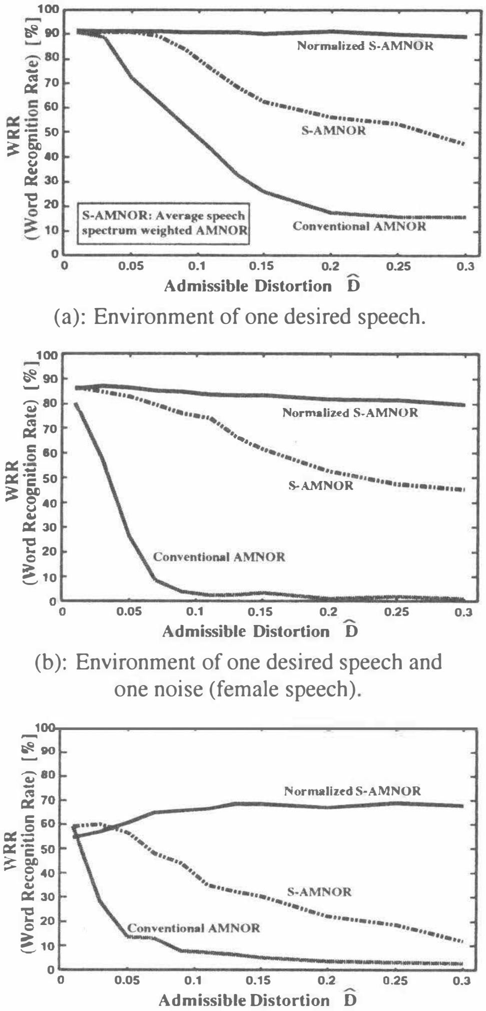
4.1. Experimental conditions

We evaluated the ASR performance in a real acoustic room. Figure 4 shows the experimental environment, and Table 1 shows the experimental condition. The desired distant signal arrives from the front direction (90 degrees), and the noise signal arrives from the right and left directions (40 degrees and 120 degrees, respectively). The distance between the sound source and the microphone array is two meters. In this situation, the ASR performance was evaluated by variations in the admissible distortion $\hat{D}$ as Equation (2). ASR performance was also evaluated by the WRR (Word Recognition Rate).

4.2. Experimental results for ASR performance

Figure 5 shows the ASR performance in the evaluation environment. In this experiment, the sound source position is known for designing the adaptation filter. In Figure 5, (a) shows the results in an environment of one desired speech [90 degrees DOA], (b) shows the results in an environment of one desired speech [90 degrees DOA] and one noise (female speech [40 degrees DOA]), and (c) shows the results in an environment of one desired speech [90 degrees DOA] and two noises (female speech [40 degrees DOA] and white Gaussian signal [120 degrees DOA]).

As a result of our evaluation experiments, we could confirm that the average speech spectrum weighted AMNOR (S-AMNOR) provides higher ASR performance than the conventional AMNOR. The effectiveness of S-AMNOR is also confirmed with large admissible distortion $\hat{D}$. It also showed the same tendency in the environment of one desired speech [90 degrees DOA] and one noise (male speech [40 degrees DOA]) and in that of one desired speech [90 degrees DOA] and one noise (white Gaussian noise [40 degrees DOA]). In addition, we could confirm that the normalized speech spectrum weighted AMNOR (Normalized S-AMNOR) is more effective than the basic S-AMNOR. This is because the adaptive filter of Normalized S-AMNOR has a more greatly optimized energy balance between vowels and consonants than that of S-AMNOR.



(a): Environment of one desired speech.

(b): Environment of one desired speech and one noise (female speech).

(c): Environment of one desired speech and two noises (female speech and white Gaussian noise).

Figure 5: ASR performance.

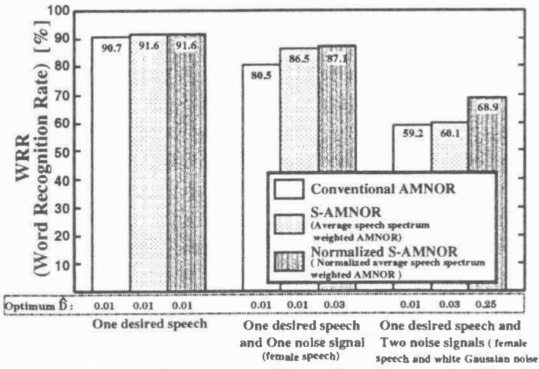
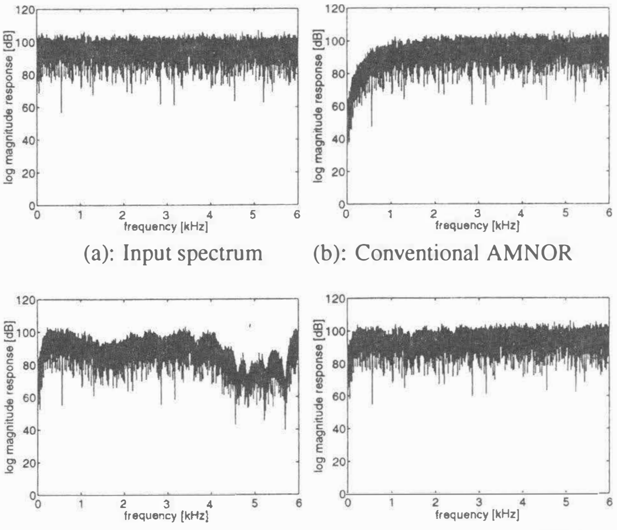


Figure 6: ASR performance with optimum admissible distortion $\hat{D}$.



(a): Input spectrum (b): Conventional AMNOR

(c): S-AMNOR (d): Normalized S-AMNOR

Figure 7: Spectrum characteristics of adaptive filter based on average speech spectrum weight when admissible distortion $\hat{D} = 0.1$.

Next, we show the maximum ASR performance with the optimum admissible distortion $\hat{D}$ in Figure 6, which we manually selected. In Figure 6, we confirm that if we estimate the optimum admissible distortion $\hat{D}$ in advance, ASR performance improves by $5 \sim 10\%$ with the normalized speech spectrum weight in a noisy environment.

4.3. Experimental results for the spectrum characteristics of designed adaptive filter

Figure 7 shows the spectrum characteristics of adaptive filters. We investigated performance with the signal as a flat spectrum characteristic and the designed adaptive filters. In Figure 7, (a) shows the input spectrum with a white Gaussian signal as the flat spectrum characteristic, (b) shows the output spectrum with an adaptive filter based on the conventional AMNOR, (c) shows the output spectrum with an adaptive filter based on the average speech spectrum weight (S-AMNOR), and (d) shows the output spectrum with an adaptive filter based on the normalized average speech spectrum weight (Normalized S-AMNOR). By comparing the results from Figure 7(b) and (d), we could confirm that the adaptive filter based on the normalized average speech spectrum weight (Normalized S-AMNOR) shows almost no distortion on any frequency band although the adaptive filter based on the white Gaussian (Conventional AMNOR) shows severe distortion in the lower frequency bands which are indispensable for speech recognition. In addition, we could also confirm, by comparing the results using Figure (c), and (d), that normalization of the average speech spectrum (Normalized S-AMNOR) improves the signal capturing performance. In the above evaluation experiments, we confirmed that AMNOR based on the normalized average speech spectrum weight (Normalized S-AMNOR) is more effective than the conventional AMNOR for noisy speech recognition.

5. CONCLUSIONS

In this paper, we proposed a method to improve ASR performance by AMNOR (Adaptive Microphone-array for NOise Reduction) with the average speech spectrum weight in noisy environments. As a result of evaluation experiments in real acoustic environments, we confirmed that ASR performance is improved by using the normalized average speech spectrum weighted AMNOR (Normalized S-AMNOR). In the future, we will improve ASR performance by integrating the proposed AMNOR with talker localization [8] and automatically estimating the optimum admissible distortion $\hat{D}$ for ASR in noisy environments.

Acknowledgement: 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.

6. 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., Vol. 78, No. 5, pp. 1508–1518, Nov. 1985.

[2] S.U. Pillai, "Array Signal Processing," Springer-Verlag, New York, 1989.

[3] L.J. Griffiths and C.W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beam-forming," IEEE Trans. AP, Vol. AP-30, No. 1, pp. 27–34, 1982.

[4] Y. Kaneda and J. Ohga, "Adaptive Microphone-array System for Noise Reduction," IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-34, No. 6, pp. 1391–1400, Dec. 1986.

[5] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone Array based speech recognition with different talker-array position," Proc. ICASSP97, pp. 227–230, 1997.

[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] T. Kobayashi, S. Itahashi, and T. Takezawa, "ASJ continuous speech corpus for research," J. Acoust. Soc. Jpn., Vol. 48, No. 12, pp. 888–893, 1992.

[8] 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.