AN EVALUATION OF ADAPTIVE BEAMFORMER BASED ON AVERAGE SPEECH SPECTRUM FOR NOISY SPEECH RECOGNITION

Takanobu Nishiura[†,‡]*, Masato Nakayama*[‡]*, and Satoshi Nakamura*[†]

[†] ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto, 619-0288 Japan
[‡] Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama, 640-8510 Japan

ABSTRACT

Distant-talking speech recognition in noisy environments 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 as an adaptive beamformer for capturing the desired distant signals in noisy environments by Kaneda et al. Although the AMNOR has been proven effective, it can be further improved if we know the spectrum characteristics of the desired distant signals in advance. Therefore, we regarded speech as a desired distant signal and designed an AMNOR based on the average speech spectrum. In this paper, we particularly focused on the performance of 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 confirmed that the ASR (Automatic Speech Recognition) performance was improved 5 – 10% by using an AMNOR based on the average speech spectrum in noisy environments. In addition, the proposed AMNOR provides better noise reduction performance than that of conventional AMNOR.

1. INTRODUCTION

To capture and recognize distant-talking speech is one of the most important functions for achieving natural interfaces for machines such as self-moving robots. However, background noise and room reverberations seriously degrade the sound capture quality in 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 selectively acquired 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 the delay-and-sum beamformer.

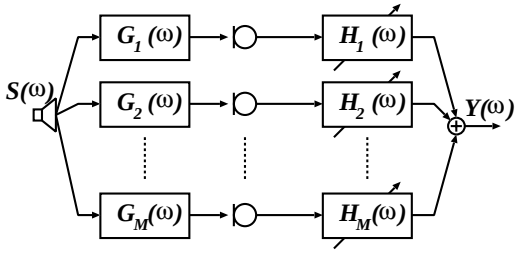


Figure 1: Block diagram of adaptive beamformer.

AMNOR (Adaptive Microphone-array for NOise Reduction) is an adaptive beamformer proposed by Kaneda et al. in 1986[4]. This 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 when capturing the desired distant signal.

However, if we knew the spectrum characteristics of desired distant signals before designing the adaptive filter of an AMNOR, we could further improve its performance. The conventional AMNOR is designed to suppress the spectrum distortion of the desired distant signal on all frequency bands. However, in many cases, the purpose of signal capture is limited to speech capture. Therefore, in this paper we regarded speech as the desired distant signal and designed an 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. The $G_m(\omega)$ 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

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

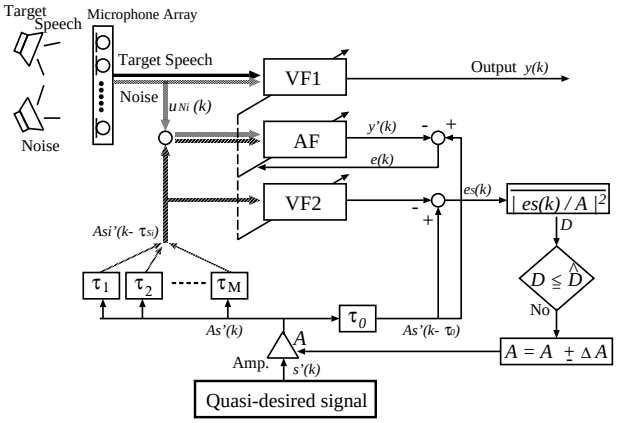


Figure 2: Overview of AMNOR.

where M is the number of microphone elements. The concept of the adaptive beamformer is to minimize the output noise energy while constraining $F(\omega)$ to the desired frequency response. An AMNOR [4] has the constraint shown in Equation (2):

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

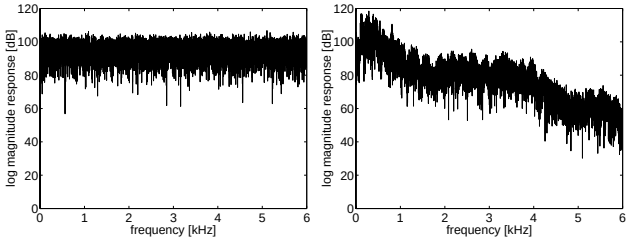
This constraint attains maximum noise reduction while allowing a small distortion D in the frequency response to the desired signal. In this paper, we focus on suitable control of the admissible distortion $\hat{D}$ in the frequency response for noisy speech recognition. Figure 2 shows a general overview of AMNOR. In Figure 2, each VF1, AF, and VF2 is a FIR filter with M-input and 1-output. The AF is an 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 an AMNOR because an AMNOR attains maximum noise reduction with a quasi-desired signal and a 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 a simulation in which signal $As'(k)$ arrives from the desired sound source with a 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 input to the adaptive filter AF after adding it to the quasi-desired signal $As'_i(k - \tau_{si})$. The 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. The $es(k)$ 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 $D \leq \hat{D}$. In the above algorithm, AMNOR attains higher noise reduction performance in real acoustic environments.



(a) White Gaussian spectrum (b) Average speech weighted spectrum

Figure 3: Spectrum of quasi-desired signal.

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. However, 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 might 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 an 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 (Automatic Speech Recognition) performance by using the average speech spectrum weighted quasi-desired signal for the proposed AMNOR, and this modified system was named S-AMNOR.

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

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

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
Number of state	3 state
Feature vector	MFCC (16 orders, 4 mixtures), ΔMFCC (16 orders, 4 mixtures), Δpower (1 order, 2 mixtures)
Average speech spectrum weight	
Speech DB	ATR speech DB SetA [5] and ASJ continuous speech corpus [6]
Speech (L)	2620 words $\times$ 4 subjects and 150 sentences $\times$ 64 subjects
Consonants (L_c, N_c)	L_c: 28156 phonemes, N_c: 94315 frames (= $\sum N_{l_c}$)
Vowels (L_v, N_v)	L_v: 23440 phonemes, N_v: 107085 frames (= $\sum N_{l_v}$)
Test data (Open)	
Desired speech signal	Speech: 216 words $\times$ 2 subjects (1 female and 1 male)
Noise signal	Female speech, male speech or white Gaussian noise
SNR	3 dB

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

4. EVALUATION EXPERIMENTS

We performed the evaluation experiments in a real acoustic room. Particularly in this paper, we focus on the ASR (Automatic Speech Recognition) performance and directivity pattern of each beamformer.

4.1. Experimental conditions

Table 1 shows the experimental conditions, and Figure 4 shows the experimental environment. 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. The microphone array consists of 14 transducers and 2.83 cm spacing. In this environment and under these conditions, we evaluated the ASR performance and the directivity pattern of each beamformer.

4.2. Experimental results for ASR performance

The ASR performance was evaluated by using WRR (Word Recognition Rate). In addition, ASR performance was also evaluated

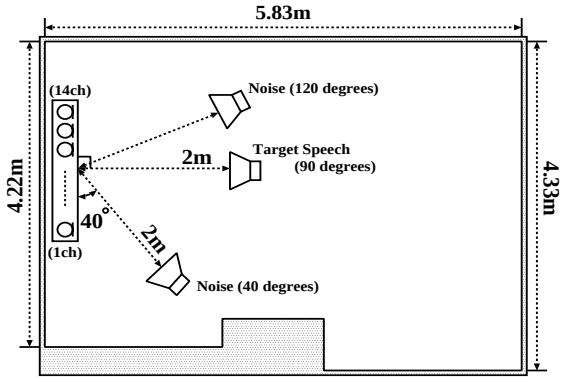


Figure 4: Experimental environment.

by using variations in the admissible distortion $\hat{D}$ (Equation (2)). Figure 5 shows the maximum ASR performance with the optimum admissible distortion $\hat{D}$, which we manually selected from experimental results (these results were reported in [7]). In this experiment, the sound source position was known in advance of designing the adaptation filter. In Figure 5, the three left bars show the results in an environment of one desired speech [90 degrees DOA(Direction Of Arrival)]. The three center bars show the results in an environment of one desired speech [90 degrees DOA] and one noise (female speech [40 degrees DOA]).The three right bars show 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]).

The results of our evaluation experiments, we could confirm that the average speech spectrum weighted AMNOR (S-AMNOR) provides higher ASR performance than the conventional AMNOR. 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 the Normalized S-AMNOR has a more greatly optimized energy balance between vowels and consonants than that of S-AMNOR.

In Figure 5, we show that if we estimate the optimum admissible distortion $\hat{D}$ in advance, the ASR performance is improved by 5 – 10% by using the normalized speech spectrum weight in a noisy environment.

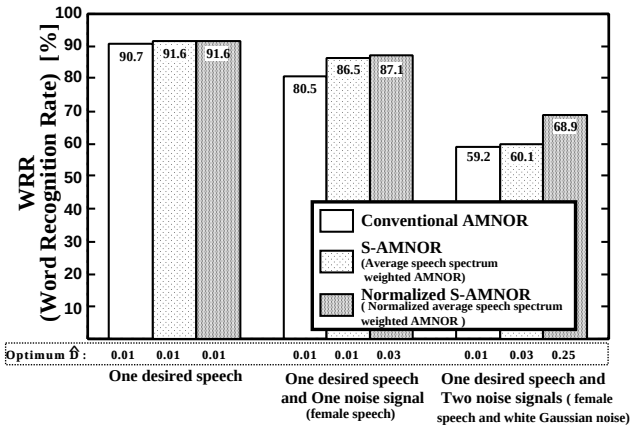
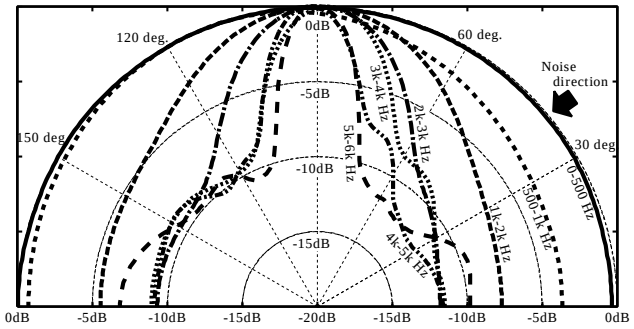
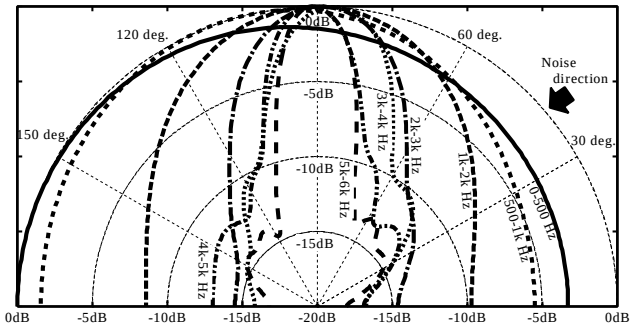


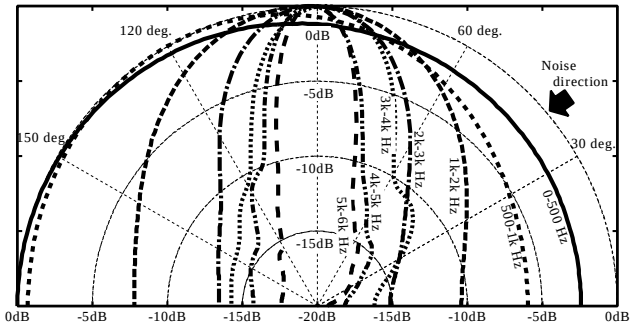
Figure 5: ASR performance with optimum admissible distortion $\hat{D}$.



(a): Conventional AMNOR.



(b): S-AMNOR.



(c): Normalized S-AMNOR.

Figure 6: Directivity pattern with optimum admissible distortion $\hat{D}$ (Desired direction [desired speech]: 90 degrees, Null direction [undesired noise (female speech)]: 40 degrees).

4.3. Experimental results for directivity pattern

Figure 6 shows the directivity pattern obtained with conventional AMNOR, S-AMNOR, and Normalized S-AMNOR. In this evaluation experiment, the sound source position was known in advance of designing the adaptation filter. The directivity pattern results were also for an environment of one desired speech [90 degrees DOA] and one noise (female speech [40 degrees DOA]). Also, the admissible distortion $\hat{D}$ was the same as that of the center three bars in Figure 5, which is the optimum admissible distortion $\hat{D}$ for ASR. We calculated the directivity pattern from signals made with frequency band pass filters and white Gaussian noise. Frequency band pass filters were designed for 0 Hz – 500 Hz, 500 Hz – 1 kHz, 1 kHz – 2 kHz, 2 kHz – 3 kHz, 3 kHz – 4 kHz, 4 kHz – 5 kHz, and 5 kHz – 6 kHz, respectively.

As a result of our evaluation experiments, we could confirm that the average speech spectrum weighted AMNOR (S-AMNOR) provides better reduction performance of the noise signal [40 degrees DOA] than the conventional AMNOR. Moreover, by comparing the directivity pattern of S-AMNOR and Normalized S-AMNOR in Figure 6(b) (c), we could confirm that Normalized S-AMNOR has a sharper directivity pattern than basic S-AMNOR in higher frequency bands.

Through the described above evaluation experiments, we could confirm that normalized average speech spectrum weighted AMNOR (Normalized S-AMNOR) is more effective than the conventional AMNOR and basic S-AMNOR.

5. CONCLUSIONS

In this paper, we proposed a new AMNOR(Adaptive Microphone-array for NOise Reduction) with an average speech spectrum weight to improve ASR performance in noisy environments. As a result of evaluation experiments in real acoustic environments, we confirmed that the ASR performance was improved and the directivity pattern was much sharper as a result of 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.

Acknowledgements: This research was partially supported by The Telecommunications Advancement Organization of Japan 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] 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.

[6] 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.

[7] T. Nishiura, S. Nakamura, Y. Okada, T. Yamada, and K. Shikano, "Suitable Design of Adaptive Beamformer Based on Average Speech Spectrum for Noisy Speech Recognition, ", Proc. ICSLP2002, pp. 1789–1792, Sept. 2002.

[8] 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.