ADAPTIVE BEAMFORMER BASED ON AVERAGE VOWEL / CONSONANT SPECTRUM WITH PHONEME IDENTIFICATION

Masato Nakayama ,Takanobu Nishiura and Hideki Kawahara

Graduate School of Systems Engineering, Wakayama University,
930 Sasaedani, Wakayama, 640-8510 Japan

1. INTRODUCTION

For tele-conference systems or voice-controlled systems, the high-quality sound capture of distant-talking speech is very important. However, background noise and room reverberations seriously degrade the sound capture quality in real acoustical environments. A microphone array is an ideal candidate for capturing distant-talking speech. With a microphone array, the desired speech signals can be acquired selectively by steering the directivity. Accordingly, a super-high directivity is necessary to reduce noise signals.

To form directivity, delay-and-sum beamformer [1] and adaptive beamformers [2] [3] have been proposed as the conventional beamformers. A delay-and-sum beamformer forms the 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).

AMNOR (Adaptive Microphone-array for NOise Reduction) [3] is an adaptive beamformer proposed by Kaneda et al. in 1986. It promises a high quality sound-capture performance even in real acoustic environments. S-AMNOR [4] has also proposed in ICSLP2002. The S-AMNOR is the modified AMNOR based on a long time speech spectrum for capturing distant-talking speech with high quality. However, the S-AMNOR is not so suitable technique for recognizing the distant-talking speech, because speech has different characteristics as vowels and consonants.

Therefore in this paper, we attempt to improve the speech recognition performance of the S-AMNOR with two adaptive filters based on vowel / consonant spectrum with phoneme identification.

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

Figure 1 shows a block diagram of an adaptive beamformer. In Fig. 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 between the desired sound source and the m-th transducer, 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

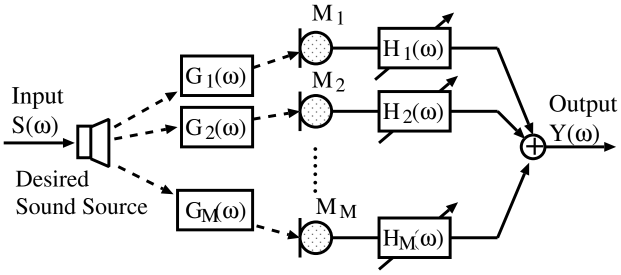


Figure 1: Block diagram of adaptive beamformer.

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

where M is the number of transducers. The concept of the adaptive beamformer is to minimize the output noise energy while constraining $F(\omega)$ to the desired frequency response.

AMNOR [3] has the constraint shown in Equation (2):

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

where $F(\omega)$ is frequency response of adaptive beamformer to the desired signal. This constraint achieves a maximum noise reduction while allowing a small distortion D in the frequency response to the desired signal. The AMNOR assumes two conditions. One, the desired sound source's DOA (Direction Of Arrival) is known. The others, the microphone only captures the noise signal without the desired signal. The AMNOR achieves the maximum noise reduction with a quasi-desired signal and an environmental noise signal from the environment.

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 Fig.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 achieves maximum noise reduction with a quasi-desired signal and an environmental noise signal from the environment. The quasi-desired signal $s'(k)$ derives $As'_i(k - \tau_{s_i})$ from amplifier and time delay $\tau_{si}, i = 1, ..., M$, which is calculated subject

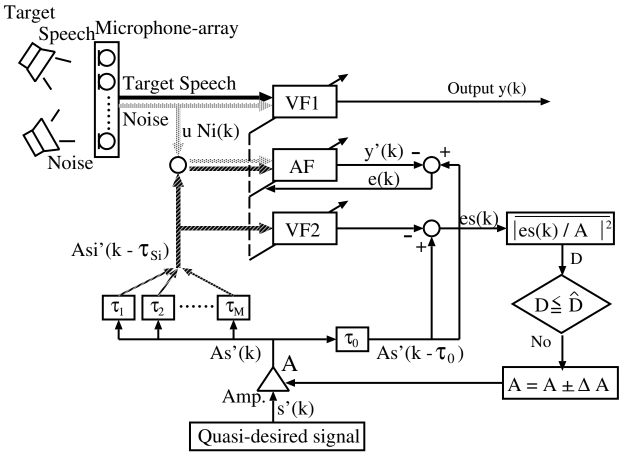


Figure 2: Overview of AMNOR.

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 $u'_{N_i}(k), i = 1, ..., M$ (not including the desired signal), and it is inputted in the adaptive filter AF after adding it to quasi-desired signal $As'_i(k - \tau_{s_i})$. AF controls the filter coefficients based on $e(k)$ as the following Eq. (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 Eq. (4).

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

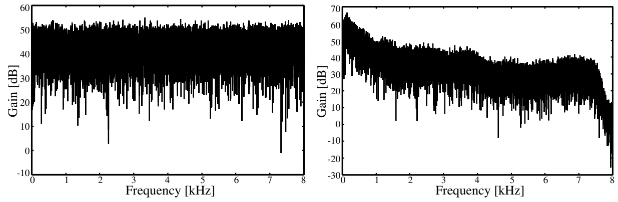
By comparing current distortion D and admissible distortion $\hat{D}$, amplitude A is renewed with the amplifier until $D \leq \hat{D}$. In the above algorithm, AMNOR achieves higher noise reduction performance in real acoustic environments.

2.1. S-AMNOR (average speech spectrum-based AMNOR)

The conventional AMNOR employed a white Gaussian signal which has flat frequency characteristics as a quasi-desired signal for suppressing a spectrum distortion of the desired signal on every frequency band. 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 may be possible to improve the performance of AMNOR by designing a suitable adaptive filter in the environment. Thus, the S-AMNOR [4] regards speech as the desired distant signal and designs by using the long time average speech spectrum for distant-talking speech capture and recognition. The long time average speech spectrum weight is calculated by (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 signal $sp_l(t)$, and $W_{sp}(\omega)$ represents the average speech spectrum weight. The S-AMNOR consists of the quasi-desired signal which is weighted the average speech spectrum to white Gaussian noise.



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

Figure 3: Spectrum of quasi-desired signal.

2.2. Spectral subtraction

Spectral Subtraction (SS) [5] is an effctive method for additive noise reduction. SS can reduce the stationary noise by subtracting the long-time average noise Fourier spectrum form the Fourier spectrum of the observed signal on the Fourier space. SS is defined by Equation (6).

$$|\hat{X}(\omega)| = |Y(\omega)| - \alpha\overline{|N(\omega)|}, \tag{6}$$

where $|\hat{X}(\omega)|$ is the Fourier spectrum of the enhanced speech, $|Y(\omega)|$ is the Fourier spectrum of the observed signal, $\overline{|N(\omega)|}$ is the long-time average noise Fourier spectrum and α is the reduction coefficient.

3. PROPOSED APPROACH

The S-AMNOR employed a long time average speech weighted spectrum for a quasi-desired signal. However, speech has different characteristics as vowels and consonants. Therefore, if the adaptive filters for the S-AMNOR can be independently designed based on a long time vowel and consonant spectrum instead of the long time speech spectrum, the performance of the S-AMNOR will be improved more effectively. Thus, we attempt to improve the speech recognition performance with the two adaptive filters based on vowel / consonant spectrum.

Figure 4 shows an overview of the proposed approach with the two adaptive filters based on vowel / consonant spectrum. In Figure 4, each VF_a, VF_v, and VF_c is FIR filter with M-input and one-output. VF_a is an adaptive filter designed by the AMNOR based on a white Gaussian noise as a quasi-desired signal, VF_v and VF_c are adaptive filters designed by the proposed approach based on an average vowel and consonant spectrum.

The vowel / consonant identification is shown as (A) and (B) in Figure 4. The vowel / consonant identification is very important for realizing this approach. Thus, the proposed approach conducts the speech enhancement process twice. First time is the speech enhancement with a conventional AMNOR and SS for the vowel / consonant identification. Second time is the speech enhancement with the S-AMNOR by switching the two adaptive filters for vowel and consonant based on the result of the vowel / consonant identification. The two-class identification for vowel / consonant identification will be conducted accurately by Equations (7)

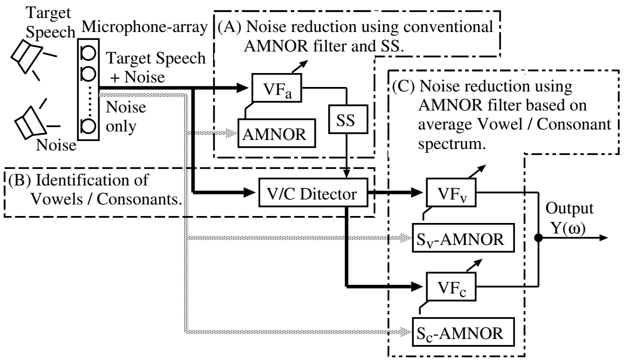


Figure 4: Overview of proposed approach with adaptive filters based on vowel / consonant spectrum

and (8).

$$\text{if } pow[N_L] > A\,\sigma^2, \quad N_{L_v} = N_L, \tag{7}$$
$$\text{else}, \qquad N_{L_c} = N_L, \tag{8}$$

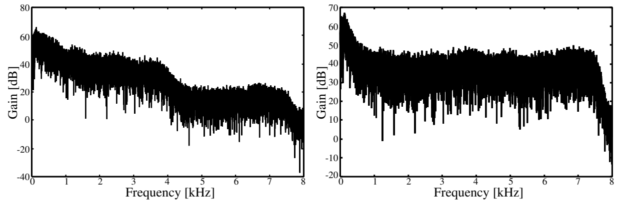
where σ^2 represents a background noise energy, A represents amplifier (in this case, $A = 10$), L represents the number of speech, L_v represents the number of vowels, L_c represents the number of consonants, N_L represents the number of speech frame, $pow[N_L]$ represents the energy of speech frame, 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 proposed approach is necessary to calculate the average vowel / consonant spectrum weights in advance by Equations (9) and (10).

$$W_v(\omega) = \frac{1}{L_v}\sum_{l_v=1}^{L_v}\frac{1}{N_{L_v}}\sum_{n=1}^{N_{l_v}} SP_{l_v}(\omega;n), \tag{9}$$

$$W_c(\omega) = \frac{1}{L_c}\sum_{l_c=1}^{L_c}\frac{1}{N_{L_c}}\sum_{n=1}^{N_{l_c}} SP_{l_c}(\omega;n), \tag{10}$$

where L_v 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), $SP_{l_v}(\omega;n)$ represents the Fourier transform of vowel signal $sp_{l_v}(t)$, $SP_{l_c}(\omega;n)$ represents the Fourier transform of consonant signal $sp_{l_c}(t)$, W_v represents the average vowel spectrum weight, W_c represents the average consonant spectrum weight. The proposed approach designs the adaptive filters with the quasi-desired spectrum by weighting the vowel / consonant spectrum weight (W_v and W_c) to the white Gaussian noise. In this proposed approach, we attempt to improve the performance of the conventional AMNOR and the S-AMNOR by switching the two adaptive filters for vowel and consonant based on the result of the vowel / consonant identification as (C) shown in Figure 4.

Figure 5 shows the average vowel / consonant weighted spectrum for the proposed approach. The average consonant weighted spectrum is enhanced at lower frequency bands (0-500Hz) compared with the average speech weighted spectrum in Figure 3 (b). The average vowel weighted spectrum is enhanced at lower-middle frequency bands (0-4000Hz). Compared Figure 5 (a) with Figure 5 (b), we confirm that the average vowel weighted spectrum is more



(a) Average vowel weighted spectrum
(b) Average consonant weighted spectrum

Figure 5: Spectrum of quasi-desired signal.

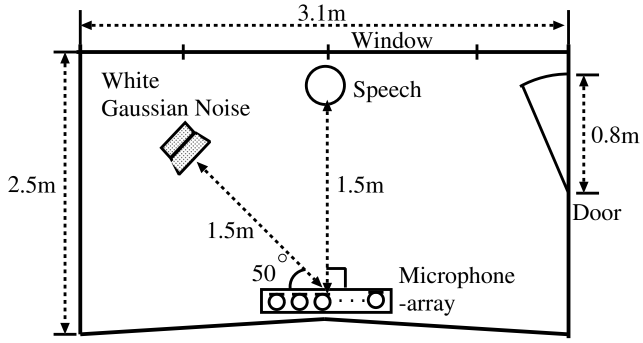


Figure 6: Experimental environment.

enhanced than the average consonant weighted spectrum in lower frequency bands.

4. EVALUATION EXPERIMENTS

We evaluated ASR (Automatic Speech Recognition) performance in a real acoustic room. Figure 6 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 directions (50 degrees). The distance between the sound source and the microphone array is 1.5 meters. The ASR performance was evaluated by the WRR (Word Recognition Rate). In this paper, we conducted the evaluation experiments in the condition of known sound source positions, known vowel / consonant periods and unknown vowel / consonant periods.

4.1. Experimental results

Figure 7 shows the ASR performance with the optimum admissible distortion $\hat{D}$. In Figure 7, in the condition of known vowel and consonant periods, we confirm that if we can estimate the optimum admissible distortion $\hat{D}$ in advance, the ASR performance of the proposed approach improves 10-13 % compared with the conventional AMNOR, and improves 8 % compared with the S-AMNOR. Also, in Figure 7, in the condition of unknown vowel and consonant periods, we confirm that if we can estimate the optimum admissible distortion $\hat{D}$ in advance, the ASR performance of the proposed approach improves 10-11 % compared with the conventional AMNOR, and improves 5-7 % compared with the S-AMNOR by the vowel /consonant identification. In addition,

Table 1: Experimental conditions

Recording conditions	
Reverberation time	$T_{[60]} = 0.12$ sec.
Microphone array	Linear type 8 transducers 2.125cm spacing
Sampling frequency	16 kHz (Quantization: 16 bit)
Average speech spectrum weight	
Speech DB	ATR speech DB SetA [6]
Speech	503 sentences × 5 subjects (males)
Vowels	72128 phonemes
Consonants	54810 phonemes
Frame length	32 msec. (Humming window)
Frame interval	8 msec.
Test data (Open)	
HMM	IPA phoneme model [7]
Desired speech signal	Speech: 216 words × 1 subject (male)
Noise signal	White Gaussian noise
SNR	0 dB, 10 dB, 20 dB

the proposed approach was always more effective than the conventional AMNOR and the S-AMNOR in each SNR environment. However, compared with the condition of known vowel and consonant periods, in the condition of unknown vowel and consonant periods, the ASR performance degrades 1-3 % in higher noisy environments (SNR=0 dB,10 dB), and improves 2 % in lower noisy environment (SNR=20 dB). This result shows two points: the vowel / consonant filter is applied as the mismatch with phoneme periods of captured speech, because the vowel / consonant identification can not be achieved accurately in higher noisy environment. On the other hand, in lower noisy environment, the vowel / consonant identification can be achieved effective performance.

Through above evaluation experiments, we confirm that the proposed approach with the two adaptive filters for vowel and consonant is more effective than the conventional AMNOR and the S-AMNOR. This is because a signal distortion will be reduced by switching the two adaptive filters for vowel and consonant.

In this paper, we use vowel / consonant classification as speech sub-categories, because the vowel and the consonant widely differ in a spectral perspective. In addition, speech sub-categories identification performance and ASR performance may degrade to use a lot of classification. Therefore, in future work, we would investigate optimum number of classification and more suitable speech sub-categories from a spectral perspective. In addition, to improve the performance, we will automatically identify the vowels and consonants from output signal of the AMNOR by using a GMM (Gaussian Mixture Model).

5. CONCLUSION

In this paper, we attempt to improve a performance of the conventional AMNOR and the S-AMNOR by the proposed approach based on the average vowel / consonant spectrum weights in noisy environments. As a result of evaluation experiment in real acoustic environment, we confirmed that the speech recognition performance is more improved than the conventional AMNOR and the S-AMNOR. In proposed approach, we use vowel / consonant classification as speech sub-categories, because the vowel and the consonant widely differ in a spectral perspective. In addition, speech sub-categories identification performance and ASR performance may degrade to use a lot of classification. Therefore, in future work, we would investigate optimum number of classification and more suitable speech sub-categories from a spectral perspective. In addition, to improve the performance, we will automatically identify the vowels and consonants from output signal of the AMNOR by using a GMM (Gaussian Mixture Model).

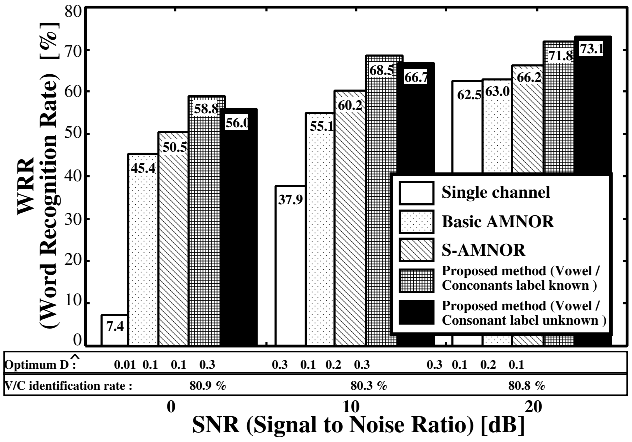


Figure 7: ASR performance

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] L. J. Griffiths, C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE. Trans. Antennas Propag., vol. AP-23, no. 1, pp. 27-34, Jan 1982.

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

[4] 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.

[5] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. ASSP, Vol.ASSP-27, No. 2, pp. 133-120, Apr. 1979.

[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] A.Lee, T.Kawahara, K.Shikano, "Julius — An Open Source Real-Time Large Vocabulary Recognition Engine," EUROSPEECH2001, pp.1691-1694, Sept 2001.