SPEECH ENHANCEMENT USING SQUARE MICROPHONE ARRAY FOR MOBILE DEVICES

Shintaro Takada, Tetsuji Ogawa, Kenzo Akagiri, and Tetsunori Kobayashi

Department of Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

ABSTRACT

In this paper, we propose a new type of speech enhancement method that is suitable for mobile devices used in noisy environments. For the sake of achieving high-performance speech recognition and auditory perception in the mobile devices, disturbance noises have to be removed under the requirements of a space-saving microphone arrangement and a low computational cost. The proposed method can reduce both the directional and the diffuse noises under the requirements for the mobile devices by applying the square microphone array and the low-cost processing that consists of multiple null beamforming, their minimum power channel selection and Wiener filtering. The effectiveness of the proposed method is clarified for speech recognition accuracies and speech qualities under the condition in which both the directional and the diffuse noises exist simultaneously: it reduced 40% of recognition errors and improved PESQ-based MOS value by 0.75 point.

Index Terms— array signal processing, Wiener filtering, speech enhancement, speech recognition, mobile devices

1. INTRODUCTION

Mobile devices such as cellular phones and personal digital assistants (PDAs) are used at a distance from a user's mouth for the case in which they are used as TV phones and voice applications with a display. The speech signals coming from distant sound sources are weakly received by microphones of the devices. In this case, signal-to-noise ratios (SNRs) are seriously degraded compared to those under close-talking conditions. Thus, environmental noises have to be removed in order to achieve high-performance speech recognition and auditory perception for the mobile devices.

Noise reduction for the mobile devices has specific requirements such as space-saving microphone arrangements and low computational costs because of limited space and processing power. In addition, since the mobile devices are used in various environments, various types of noises, which include directional noises and diffuse noises, have to be removed.

Beamforming, blind source separation (BSS) and Wiener filtering [1][2] are frequently applied to noise reduction. However, these methods generally aim at reducing either only directional noises (e.g. BSS) or only diffuse noises (e.g. Wiener filtering). Moreover, these methods often require large number of microphones and large scale of microphone arrangement or high computational costs. Thus, these methods are not suitable for the speech enhancement system of the mobile devices.

In the present paper, we propose a new type of noise reduction method suitable for the mobile devices using the square microphone array. Four microphones are arranged at the apexes of a small square (one side is 4 cm). This arrangement is appropriate to mount the microphones on the mobile devices. By applying this square microphone array, the proposed method can suppress both directional noises and diffuse noises with the same microphone arrangement and the low computational cost. Firstly, four spatial filters are formed by multiple null beamformers, each of which is developed by a pair of two microphones. Then, using outputs of the spatial filters, minimum power channel selection is performed for directional noise reduction. In parallel with that, multi-channel Wiener filtering that uses the outputs of the spatial filters is performed for diffuse noise reduction. For the case in which the distance between two microphones is small (e.g. 4cm), the performance of diffuse noise reduction is generally poor because inter-channel correlation of the diffuse noises included in the microphone observations is not low especially in a low frequency domain. In the proposed method, on the other hand, this problem can be solved by using not microphone observations but four spatial filter outputs. As the result, the correlation of diffuse noises becomes lower and more precise multi-channel Wiener filter can be achieved. Finally, we perform single-channel Wiener filtering for reducing residual diffuse noise and then obtain enhanced target speech.

The rest of the present paper is organized as follows. In section 2, the algorithm of the proposed method is described. In section 3, the effectiveness of the proposed method is evaluated for continuous speech segregation and recognition on the basis of word accuracy and perceptual evaluation of speech quality (PESQ) [3]. We give the conclusions in section 4.

2. PROPOSED METHOD

A microphone arrangement is described in Fig. 1. Four omni-directional microphones are arranged at the apexes of the square. The distances between two microphones are 4 cm. The di-

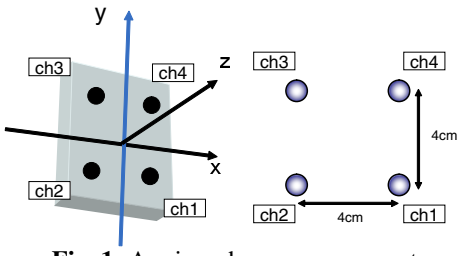

Fig. 1. A microphone arrangement.

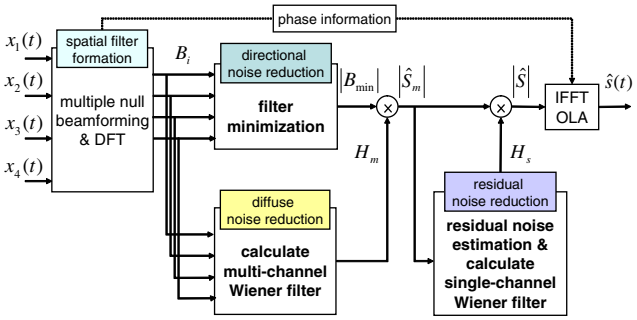

Fig. 2. A block diagram of the proposed method. (ω, k) is abbreviated for processing of frequency domain.

rection of the desired target source is assumed as that of the z-axis in Fig. 1. This assumption is adequate for the mobile applications. In the present paper, we define that interfering noises consist of directional noises and diffuse noises. Here , we assume the sound field in which one disturbance speech exists as the directional noise and the almost stationary noise (e.g. server room noise) exists as the diffuse noise. Two types of noises are arrived at the microphones simultaneously.

Figure 2 illustrates a diagram of the proposed noise reduction method. The proposed method consists of four stage signal processing: 1) multiple null beamforming for generating four spatial filters, 2) spatial filter minimization for directional noise reduction, 3) multi-channel Wiener filtering for diffuse noise reduction and 4) single-channel Wiener filtering for residual noise reduction.

2.1. Spatial filter formation

Four spatial filters ϕ_1, ϕ_2, ϕ_3 and ϕ_4 are formed by multiple null beamformers, each of which is developed by a pair of two microphones. Figure 3 illustrates the directivity patterns generated by the spatial filters ϕ_i. This stage applies fixed null beamformer and requires no adaptive techniques. Four spatial filter outputs are described as follows.

$$b_1(t) = x_2(t-\tau) - x_1(t) \quad (1)$$
$$b_2(t) = x_3(t) - x_4(t-\tau) \quad (2)$$
$$b_3(t) = x_2(t-\tau) - x_3(t) \quad (3)$$
$$b_4(t) = x_1(t) - x_4(t-\tau) \quad (4)$$

where $b_i(t)$ denotes the output of i-th spatial filter ϕ_i at each discrete time t, $x_j(t)$ denotes the microphone observation of

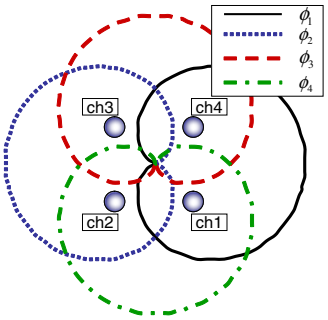

Fig. 3. Directivity patterns of spatial filtering.

j-th channel, and τ denotes the delay added for generating cardioid patterns as shown in Fig. 3. In this case, τ equals d/c, where c denotes the sound velocity and d denotes the distance between two microphones. In the present paper, $B_i(\omega, k)$ represents a short-time spectral component of the spatial filter output $b_i(t)$ at each frame k and each discrete frequency ω.

2.2. Directional noise reduction

The output of the spatial filter $B_i(\omega, k)$ consists of three spectral components, $S_i^B(\omega, k)$, $N_i^{dir}(\omega, k)$ and $N_i^{dif}(\omega, k)$, where $S_i^B(\omega, k)$, $N_i^{dir}(\omega, k)$ and $N_i^{dif}(\omega, k)$ denote the target source component, the directional noise component and the diffuse noise component, respectively. Here, each component is assumed to be uncorrelated with the others. $B_i(\omega, k)$ is described as follows.

$$B_i(\omega, k) = S_i^B(\omega, k) + N_i^{dir}(\omega, k) + N_i^{dif}(\omega, k) \quad (5)$$

Since the direction of the target source is assumed as that of the z-axis in Fig. 1 and the diffuse noise has no directional components, $|S_i^B(\omega, k)|$ and $|N_i^{dif}(\omega, k)|$ are respectively the same for every spatial filter as follows.

$$|S_i^B(\omega, k)| = |S^B(\omega, k)|, \quad (i = 1, 2, 3, 4) \quad (6)$$
$$|N_i^{dif}(\omega, k)| = |N^{dif}(\omega, k)|, \quad (i = 1, 2, 3, 4) \quad (7)$$

Thus, $|B_i(\omega, k)|$ depends on only $|N_i^{dir}(\omega, k)|$. From the above discussion, the least-directional-noise component $|B_{\min}(\omega, k)|$ can be estimated by selecting the minimum of $|B_1(\omega, k)|$, $|B_2(\omega, k)|$, $|B_3(\omega, k)|$ and $|B_4(\omega, k)|$ as follows.

$$|B_{\min}(\omega, k)| = \min_i \left[|B_i(\omega, k)| \right], \quad (i = 1, 2, 3, 4) \quad (8)$$

2.3. Diffuse noise reduction

In the second stage, the spectrum of $N_i^{dif}(\omega, k)$ is suppressed by multi-channel Wiener filtering. Zelinski computed the multi-channel Wiener filter using observations of omni-directional microphones [1]. However, the performance of diffuse noise reduction is poor for the case in which the distance between two microphones is small. Aiming at solving this problem, we attempt to use not microphone observations but four spatial filter outputs. The Wiener filter we applied is described as follows.

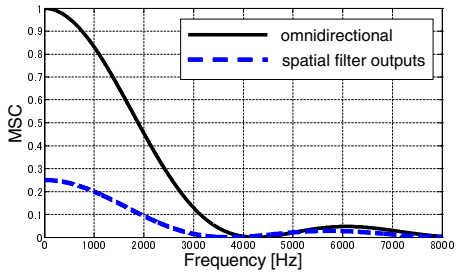

Fig. 4. Theoretical magnitude-squared coherences as a function of frequencies for the case in which the distances between two microphones are 4 cm.

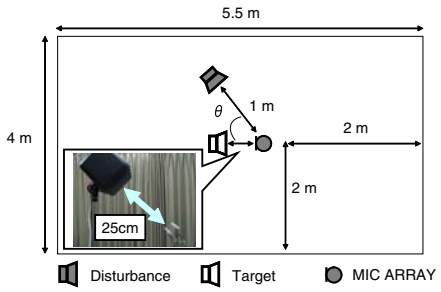

Fig. 5. Recording condition of target speech and directional noise. θ = 30°, 60°, 90°, 120°, 150°, 180°.

$$H_m(\omega,k) = \frac{\frac{1}{2}\sum[\mathrm{Re}\{B_p(\omega,k)B_q^*(\omega,k)\}]}{\frac{1}{4}\sum[B_r(\omega,k)B_r^*(\omega,k)]} \quad (9)$$

where p, q, and r, are selected as $(p,q) = \{(1,2),(3,4)\}$ and $r = \{1,2,3,4\}$. By selecting the pair of spectral components that have adverse directivity patterns (difference between them is just 180 degrees) in the numerator, the correlation among the diffuse noise components is expected to be minimized. Figure 4 illustrates theoretical magnitude-squared coherence (MSC) for the case in which omni-directional microphone observations are used and that for the case in which the spatial filter outputs are used.

The estimated target signal component $|\hat{S}_m(\omega,k)|$, in which both the directional noise and the diffuse noise are suppressed, is described as follows.

$$|\hat{S}_m(\omega,k)| = H_m(\omega,k)\cdot|B_{\min}(\omega,k)| \quad (10)$$

2.4. Residual noise reduction

As shown in Fig. 4, MSC estimated from the output of the spatial filter is not completely zero for all frequencies. This means that $|\hat{S}_m(\omega,k)|$ includes the residual diffuse noise components especially in a low frequency domain. Thus, we attempt to remove the residual noise by single-channel Wiener filter computed by using the noise components estimated as follows.

$$|\hat{N}_m(\omega,k)|^2 = \lambda|\hat{N}_m(\omega,k-1)|^2+(1-\lambda)|\hat{S}_m(\omega,k)|^2 \quad (11)$$

where $|\hat{N}_m(\omega,k)|^2$ denotes the estimate of the residual noise spectrum and λ denotes a forgetting factor, which is determined on the basis of ideal speech presence probabilities. Using estimated residual noise power spectrum $|\hat{N}_m(\omega,k)|^2$, we compute the Wiener filter $H_s(\omega,k)$ for residual noise reduction as follows.

$$H_s(\omega,k) = \frac{\mathrm{SNR}_{\mathrm{priori}}(\omega,k)}{\mathrm{SNR}_{\mathrm{priori}}(\omega,k)+1} \quad (12)$$

where $\mathrm{SNR}_{\mathrm{priori}}(\omega,k)$ is defined as follows.

$$\mathrm{SNR}_{\mathrm{priori}}(\omega,k) = \frac{E[|S(\omega,k)|^2]}{E[|N_m(\omega,k)|^2]} \quad (13)$$

$E[.]$ represents an expectation operator. Here, $\mathrm{SNR}_{\mathrm{priori}}(\omega,k)$ is approximated by using two-step noise reduction [4].

The final estimate of the target spectrum $|\hat{S}(\omega,k)|$ is obtained as follows.

$$|\hat{S}(\omega,k)| = H_s(\omega,k)\cdot|\hat{S}_m(\omega,k)| \quad (14)$$

Here, a distortion in frequency of $|\hat{S}(\omega,k)|$, which is caused by null beamforming, is compensated by [5], and an appropriate phase function has to be given in order to recover the time-domain estimate $\hat{s}(t)$.

3. SPEECH ENHANCEMENT EXPERIMENTS

Experimental comparisons were conducted for speech segregation and recognition under the condition that both the directional and the diffuse noise exist. The performances were evaluated using the measures of word accuracies and PESQ to derive mean opinion scores (MOS).

3.1. Experimental setup

In this experiment, the target speech, the directional noise and the diffuse noise were recorded separately and then mixed. Figure 5 shows the positions of the target speech and the interfering directional noise. Time stretched pulses (TSPs) were played back through a loudspeaker that acts as a replacement of people from six different directions. Impulse responses were computed using the TSPs [6]. Here, the reverberation time (RT_{60}) was 240 ms. As shown in Fig. 5, the microphone array was placed in a state inclined, the loudspeaker representing the target source was placed in front of the microphones, and the other loudspeaker representing the disturbance was placed in a direction of θ degree from the target source. The distance between the target source and the microphones was 25 cm and that between the disturbance source and the microphones was 100 cm. Altitudes of the target and the disturbance loudspeaker were 140 cm. For the target and the disturbance utterances, we selected a total of 100 Japanese newspaper article sentences spoken by 23 male speakers. Then, we computed the convolution of these sentences and the impulse response. On the other hand, the diffuse noise are recorded in a server room using the same microphone array as used for recordings of the directional signals.

Table 1. Setup for speech enhancement.

sampling frequency	32 kHz
frame length	1024 points
frame shift	256 points
analysis window	Hamming window
added delay τ	$\tau = 0.04/343$ sec
analysis range of frequencies	300 Hz $-$ 7500 Hz

This noise was almost stationary. The target utterance was mixed with the disturbance utterance and the diffuse noise. SNR between the target and the disturbance utterance was 10 dB, and that between the target utterance and the diffuse noise was 15 dB. Thus, the final SNR between the target utterance and the interfering noises was about 6 dB. Experimental setup for speech enhancement is shown in Table 1.

The setup for speech recognition is described as follows. Acoustic features were represented by 25-dimensional parameters that consist of 12-dimensional MFCCs, 12-dimensional ΔMFCCs, and Δpower. Acoustic models were trained with 20414 sentences, which were spoken by 133 male speakers and recorded with close-talking microphones. We adopted state-tied triphones in which the number of states was 2000 and the distribution function in each state was represented by a 16-mixture Gaussian distribution with diagonal covariances. As for a language model, we used word trigrams that were constructed using a lexicon of 20K vocabulary.

3.2. Experimental results

The effectiveness of the proposed method was evaluated on the basis of speech recognition accuracies and PESQ-based MOS values (PESQ-MOS). PESQ-MOS was calculated using the reference signal that was the observations of the microphones for the case in which only the target source exist. The signals were down-sampled into 16 kHz because of the implementation of the PESQ. We evaluated the performances of three kinds of evaluation items as follows: 1) directional noise reduction processing by only the filter minimization (directional NR), 2) directional and diffuse noise reduction processing by filter minimization and multi-channel Wiener filter (directional & diffuse NR) and 3) residual noise estimation and reduction by single-channel Wiener filter in addition to 2) (residual NR). The word accuracies and the PESQ-MOS were shown in Fig. 6 and Fig. 7, respectively.

As shown in Fig. 6, the word accuracy is about 35 % without any noise reduction processing. Only the directional noise reduction based on filter minimization achieved a word accuracy of 60%. Moreover, diffuse noise reduction based on multi-channel Wiener filtering and additional residual noise reduction based on single-channel Wiener filtering improved the word accuracies to 70% and 75%, respectively. These improvements were also found in the evaluation based on the PESQ-MOS as shown in Fig. 7. These results indicate that the proposed method achieves high performances from the viewpoints of both speech recognition accuracies and speech qualities.

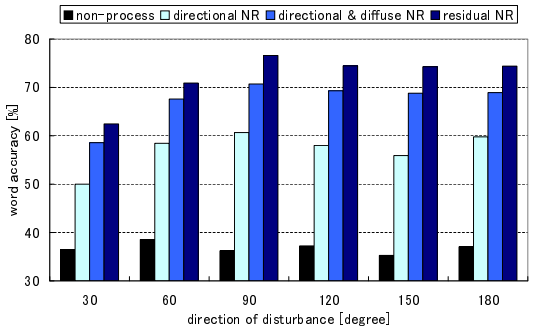

Fig. 6. Word accuracy as a function of disturbance directions.

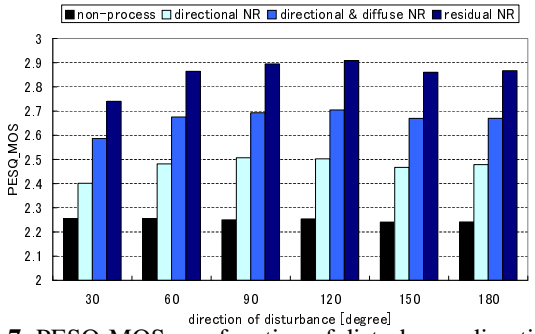

Fig. 7. PESQ-MOS as a function of disturbance directions.

4. CONCLUSION

We proposed a new type of speech enhancement method that targets both the directional and the diffuse noises by using multiple null beamforming, their minimum power channel selection , multi-channel Wiener filtering and single-channel Wiener filtering. We revealed the effectiveness of the proposed method for continuous speech segregation and recognition on the basis of word accuracies and speech qualities. The proposed method improved the word accuracy of 40% and the PESQ-based MOS of about 0.7 point compared with non-processing under the conditions in which both the directional and the diffuse noises exist simultaneously.

5. REFERENCES

[1] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," Proc. ICASSP, vol. 5, pp. 2578-2581, 1988.

[2] Y. Ephraim *et al.*, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," IEEE Trans. Speech Audio Process., vol. 33, no. 2, pp. 443-445, 1985.

[3] ITU-T Recommendation P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," 2001.

[4] C. Plapous *et al.*, "Two-Step Noise Reduction Technique," Proc. ICASSP, vol. 1, pp. 289-292, May 2004.

[5] S. Takada *et al.*, "Sound Source Separation Using Null-Beamforming and Spectral Subtraction for Mobile Devices," Proc. WASPAA, pp.30-33, Oct. 2007.

[6] Y. Suzuki *et al.*, "An Optimum Computer-Generated Pulse Signal Suitable for the Measurement of Very Long Impulse Responses," J. Acoust. Soc. Am., vol.97(2), pp.1119-1123, 1995.