

SPEECH ENHANCEMENT WITH MICROPHONE ARRAY AND FOURIER / WAVELET SPECTRAL SUBTRACTION IN REAL NOISY ENVIRONMENTS

Yuki DENDA, Takanobu NISHIURA, and Hideki KAWAHARA

Faculty of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama, 640-8510 Japan
{s045064, nishiura, kawahara}@sys.wakayama-u.ac.jp

ABSTRACT

It is very important to capture distant-talking speech with high quality for teleconferencing systems or voice-controlled systems. For this purpose, microphone array steering and Fourier spectral subtraction, for example, are ideal candidates. A combination technique using both microphone array steering and Fourier spectral subtraction has also been proposed to improve performance. However, it is difficult for the conventional approach to reduce non-stationary noise, although it is easy to robustly reduce stationary noise. To cope with this problem, we propose a new combination technique with microphone array steering and Fourier / wavelet spectral subtraction. Wavelet spectral subtraction promises to effectively reduce non-stationary noise, because the wavelet transform admits a variable time-frequency resolution on each frequency band. As a result of an evaluation experiment in a real room, we confirmed that the proposed combination technique provides better performance of the ASR (Automatic Speech Recognition) and NRR (Noise Reduction Rate) than the conventional combination technique.

1. INTRODUCTION

The high-quality sound capture of distant-talking speech is very important for teleconferencing systems or voice-controlled systems. However, ambient 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 the microphone array, the desired speech signals can be acquired selectively by steering the microphone array in the desired speech direction sensitively.

A delay-and-sum beamformer (DS) [1] is one of the most popular steering techniques for microphone arrays. This beamformer can effectively reduce the undesired noise by steering the sharp directivity to the desired sound source direction. However, it is difficult to completely reduce ambient noise and room reverberations by only using the DS. This is because the DS has frequency dependability, so that it is especially difficult to form sharper directivity in lower frequency bands. On the other hand, Fourier Spectral Subtraction (SS) [2] is also proposed to reduce additive noise effectively. The SS can reduce additive noise by subtracting the long-time average noise spectrum from the spectrum of observed signals on the Fourier space. However, it is difficult for the SS to reduce non-stationary noise (for example, sudden noise) because of noise characteristic mismatches between the noise spectrum and the long-time average noise spectrum.

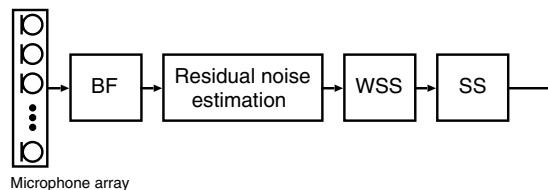


Figure 1: Overview of the proposed method.

To overcome these problems, a combination technique with microphone array steering and SS has been proposed [3]. However, this approach can not sufficiently achieve the effective performance in a non-stationary noisy environment. Therefore, we propose a new combination technique with microphone array steering and Fourier / wavelet spectral subtraction. Wavelet transform admits a variable time-frequency resolution on each frequency band. Therefore, if the noise spectrum is subtracted from the spectrum of the observed signal on the wavelet space, the non-stationary noise reduction performance will be improved. We call this approach the Wavelet Spectral Subtraction (WSS).

In this paper, we try to improve the ASR (Automatic Speech Recognition) performance by the proposed combination technique with microphone array steering and Fourier / wavelet spectral subtraction in real acoustic environments.

2. PROPOSED METHOD

We assume that distant-talking speech, stationary noise and non-stationary noise simultaneously arrive at the microphone array. In this situation, stationary and non-stationary noise reduction is necessary for capturing the distant-talking speech with high quality.

To overcome this problem, we propose a new combination technique with microphone array steering and Fourier / Wavelet Spectral Subtraction (SS / WSS). Figure 1 shows an overview of the proposed combination technique. The wavelet transform admits a variable time-frequency resolution on each frequency band. Therefore, if the noise spectrum is subtracted from the spectrum of the observed signal on the wavelet space, the non-stationary noise reduction performance will be improved.

Every signal is captured with the microphone array. Then, a noise signal is reduced with the DS. However, the residual signal must appear in the output signal of the DS. Therefore, the SS / WSS is conducted to reduce the residual signal based on the characteristics of the residual signal. The SS is a suitable tech-

nique for reducing the stationary noise, and the WSS is a suitable technique for reducing the non-stationary noise. Thus, the characteristic estimation of the residual signal is also conducted on the wavelet space with subtractive beamformer [4] before the SS and WSS is conducted.

2.1. Delay-and-sum beamformer

A delay-and-sum beamformer (DS) [1] is used to form the directivity in the desired sound direction. To capture the signal by microphone array, we assume that the plane wave of the desired sound signal comes from direction θ , the number of transducers is M , and the spacing between the transducers is d . In beamforming, the captured signals $x_1(t), x_2(t), \dots, x_M(t)$ are shown as time delays of $x_1(t)$ in Equation (1).

$$x_m(t) = x_1(t - (m - 1)\tau), \quad \tau = \frac{d \cos \theta}{c}, \quad (1)$$

where m ($m = 1, 2, \dots, M$) is the number of transducers and c is the sound propagation speed. Output signal $y(t)$ of the DS is shown in Equation (2).

$$y(t) = \sum_{m=1}^M x_m(t + (m - 1)\tau), \quad \tau = \frac{d \cos \theta}{c}. \quad (2)$$

In Equation (2), the desired sound signal from direction θ is emphasized M times because the sound signals captured with multiple transducers are added after they are synchronized. On the other hand, no other sound signal is M times as large as the desired sound signal because the directions of the other signals are different from that of the desired sound signal. Thus, the directivity of the DS can only be formed in direction θ . Therefore, the DS can form directivity in the desired talker direction.

2.2. Fourier Spectral Subtraction (SS)

2.2.1. Fourier transform (short-time Fourier transform)

Fourier transform is one of the most popular frequency analyses. The short-time Fourier transform (STFT) [5] is the most popular Fourier transform for time-frequency analysis. STFT is defined as:

$$X(b, \omega) = \int_{-\infty}^{\infty} x(t)w(t - b)e^{-j2\pi ft} dt, \quad (3)$$

where $w(t)$ is a window function and b is a parameter for transforming the window function. In the STFT, the length of the time-frequency window function is fixed. Therefore, the STFT is a suitable technique for stationary noise analysis because the window function always has the same time-frequency resolution.

2.2.2. Fourier spectral subtraction

Fourier Spectral Subtraction (SS) [2] is an effective method for additive noise reduction. SS can reduce the stationary noise by subtracting the long-time average noise Fourier spectrum from the Fourier spectrum of the observed signal on the Fourier space. SS is defined by Equation (4).

$$|\hat{X}(\omega)| = |Y(\omega)| - \alpha |\overline{N(\omega)}|, \quad (4)$$

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. Although SS is a powerful technique for stationary noise reduction, it can not effectively reduce non-stationary noise because of noise characteristic mismatches between the noise Fourier spectrum and the long-time average noise Fourier spectrum.

2.3. Wavelet Spectral Subtraction (WSS)

2.3.1. Wavelet transform

Wavelet transform is also one of the most popular frequency analyses. It has the advantage that it admits a variable time-frequency resolution on each frequency band. The wavelet transform $X(b, a)$ [6] for a one-dimensional signal $x(t)$ is defined by Equation (5).

$$X(b, a) = \int_{-\infty}^{\infty} x(t)\overline{\psi}_{a,b}(t) dt, \quad (5)$$

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \quad (6)$$

where $\psi(t)$ is a function called the mother wavelet, $\overline{\psi}(\cdot)$ represents the complex conjugate of $\psi(\cdot)$, and $\psi_{a,b}(t)$ is obtained by transformation and dilatation of the mother wavelet. The wavelet transform admits varied resolutions of the time-frequency window on each frequency band. Therefore, the wavelet transform is a suitable transform for non-stationary noise (for example, sudden noise) analysis.

2.3.2. Wavelet spectral subtraction

Wavelet Spectral Subtraction (WSS) can reduce the non-stationary noise by subtracting the noise wavelet spectrum from the wavelet spectrum of the observed signal on the wavelet space, because the wavelet transform admits a variable time-frequency resolution on each frequency band. Therefore, the WSS will be a suitable technique for non-stationary noise reduction. The WSS is defined by Equation (7).

$$|\hat{X}(b, a)| = |Y(b, a)| - \alpha |\overline{N(b, a)}|, \quad (7)$$

where $|\hat{X}(b, a)|$ is the wavelet spectrum of the enhanced speech, $|Y(b, a)|$ is the wavelet spectrum of the observed signal, $|\overline{N(b, a)}|$ is the wavelet spectrum of the noise signal and α is the reduction coefficient. However, it is necessary to accurately estimate the wavelet spectrum of the noise signal ($|\overline{N(b, a)}|$). Therefore, we try to estimate the wavelet spectrum of the non-stationary noise signal in Section 2.4. If the wavelet spectrum of the non-stationary noise signal is accurately estimated, the WSS will promises to reduce the non-stationary noise.

2.4. Non-stationary noise estimation for wavelet spectral subtraction

Non-stationary noise wavelet spectrum estimation is necessary for achieving the proposed combination technique. Therefore, we employed the technique based on the subtractive microphone array [4]. We can cancel the target speech signal from the observed signal with the subtractive microphone array. Therefore, the residual signal cancelled by the target speech is almost the mixing-noise signal of the stationary and non-stationary noise, although it is slightly distorted. Therefore, we can roughly estimate the

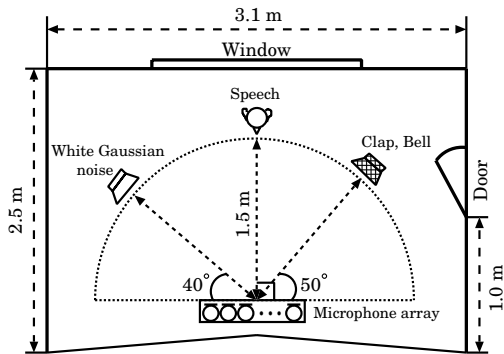


Figure 2: Experimental environment.

non-stationary noise wavelet spectrum by subtracting the long-time average stationary noise wavelet spectrum from the estimated mixing-noise wavelet spectrum as shown in Equation (8).

$$\hat{N}(b, a) = |N(b, a)| - \varepsilon(|\overline{N(b, a)}| + \sigma_{N(a)}), \quad (8)$$

where $\hat{N}(b, a)$ is the estimated non-stationary noise wavelet spectrum, $|N(b, a)|$ is the mixing noise wavelet spectrum acquired using the subtractive microphone array, $|\overline{N(b, a)}|$ is the long-time average stationary noise wavelet spectrum of the pre-estimated non-speech periods, $\sigma_{N(a)}$ is the standard deviation of $|N(b, a)|$ and ε is the admissible error on $|\overline{N(b, a)}| + \sigma_{N(a)}$. We regarded $\hat{N}(b, a)$ as the non-stationary noise wavelet spectrum instead of $|\overline{N(b, a)}|$ in Equation (7). As a result, if $\hat{N}(b, a) \leq 0$, the SS is only conducted for stationary noise reduction. Also, if $\hat{N}(b, a) > 0$, the WSS is conducted for non-stationary noise reduction (however, the SS is always conducted for ambient noise reduction after the WSS).

3. EVALUATION EXPERIMENTS

We carried out an evaluation experiment in a real room, as shown in Figure 2. In this paper, we evaluated the performance of the Automatic Speech Recognition (ASR) and noise reduction with the proposed combination technique.

3.1. Experimental conditions

Figure 2 shows the experimental environment. The desired signal comes from the front direction (90 degrees), the stationary noise comes from the left direction (40 degrees) and the non-stationary noise comes from the right direction (130 degrees). In this paper, the microphone array is steered in the known, desired speech direction. We employed white Gaussian noise as the stationary noise and sudden noise (clap and bell sounds (in the RWCP-DB [7])) as the non-stationary noise. The distance between the sound source and the microphone array is 1.5 meters.

Table 1 shows the recording conditions and Table 2 shows the experimental conditions. We evaluate the performance of the ASR and noise reduction, subject to the stationary Noise to non-stationary Noise Ratio (NNR) of 0 dB and the Signal to Noise Ratio (SNR) of -5 dB, ~, 20 dB, and clean, respectively. We employed the Gabor function as the mother wavelet [6] and conducted the wavelet transform with eight analyzed octaves which include ten divisions in each octave.

Table 1: Recording conditions

Microphone array	8 transducers, 2.125 cm spacing
Sampling frequency	16 kHz
Reverberation time $T_{[60]}$	0.12 sec.
Ambient noise	20 dBA
SNR	-5 dB, ~, 20 dB, clean (NNR: 0 dB)

Table 2: Experimental conditions

Frame length	32 msec. (Hanning window)
Frame interval	8 msec.
Reduction coefficient (α)	SS: 1.0 WSS: 1.0
Admissible error (ε)	WSS: 2.0
ASR	
Phoneme HMM	IPA phoneme model [8]
Feature vector	MFCC, Δ MFCC, and Δ power
Test date	
Speech (open)	216 isolated Japanese words × 2 subjects (1 female and 1 male)
Stationary noise	White Gaussian noise
Non-stationary noise	Clap, bell (RWCP-DB [7])

We employed 216 phoneme-balanced isolated Japanese words × 2 subjects (1 female and 1 male) as the speech test data. The ASR performance is evaluated by the Word Recognition Rate (WRR), and the noise reduction performance is evaluated by the Noise Reduction Rate (NRR). The NRR is calculated by subtracting the input-SNR from the output-SNR.

3.2. Preliminary experimental results

We first evaluate the performance in the stationary or non-stationary noisy environment as preliminary experiments. Figure 3 shows the experimental results in the stationary noise and target speech environment. In Figure 3, 1 ch represents the ASR and NRR results of the captured signal with a single transducer. As a result, we confirm that the proposed method is more effective than the conventional methods on the ASR and NRR. In addition, we also confirm that the combination of the DS and SS is more effective than the combination of the DS and WSS in a stationary noisy environment.

Figure 4 shows the experimental results in the non-stationary noise and target speech environment. As a result, we confirm that the proposed method is more effective than the conventional methods on the ASR and NRR. In addition, we also confirm that the combination of the DS and WSS is more effective than the combination of the DS and SS in a non-stationary noisy environment.

3.3. Experimental results in stationary and non-stationary noisy environment

Finally, we evaluate the proposed method in a stationary noise, non-stationary noise and target speech environment. Figure 5 shows the wave forms for each situation. As shown by the results in Figure 5(e), (f) and (g), we confirm that the proposed method can effectively reduce the stationary and non-stationary noise, although the conventional method is not enough to reduce them. Figure 6 shows the experimental results on the ASR and NRR. As a result, we confirm that the proposed method is more effective than the

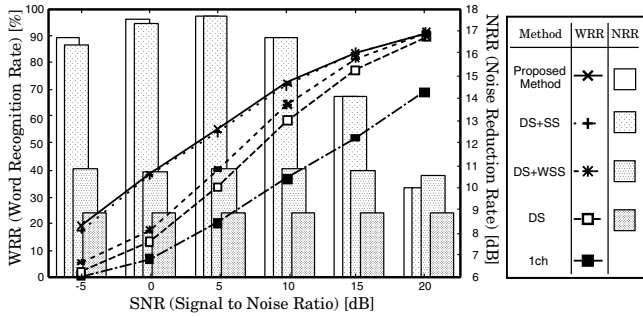


Figure 3: ASR and NRR in the stationary noise and target speech environment.

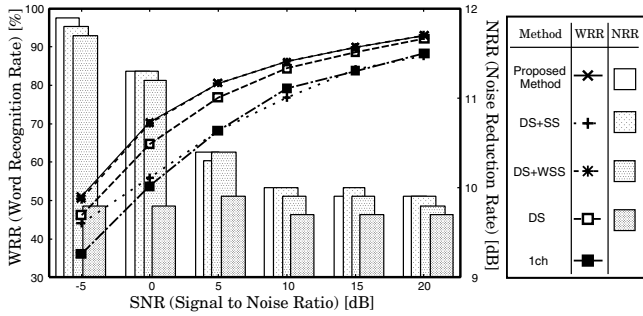


Figure 4: ASR and NRR in the non-stationary noise and target speech environment.

conventional methods on the ASR and NRR. In addition, we also confirm that the combination of the DS and WSS is more effective than the combination of the DS and SS in a non-stationary noisy environment.

4. CONCLUSION

In this paper, we proposed a new combination technique of a microphone array and Fourier / wavelet spectral subtraction for capturing distant-talking speech with high quality. As a result of an evaluation experiment in a real room, we confirmed that the proposed combination technique is more effective than the conventional combination technique on the ASR and NRR. In future work, we will attempt to estimate the noise wavelet spectrum more accurately.

5. 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. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. ASSP*, Vol. ASSP-27, No. 2, pp. 113–120, Apr. 1979.

[3] M. Dahl, I. Claesson and S. Nordebo, "Simultaneous Echo Cancellation and Car Noise Suppression Employing a Microphone Array," *ICASSP97*, Vol.1, pp. 239–242, Apr. 1997.

[4] L. J. Griffiths and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Trans. AP*, Vol. 30, No. 1, pp. 27–34, Jan. 1982.

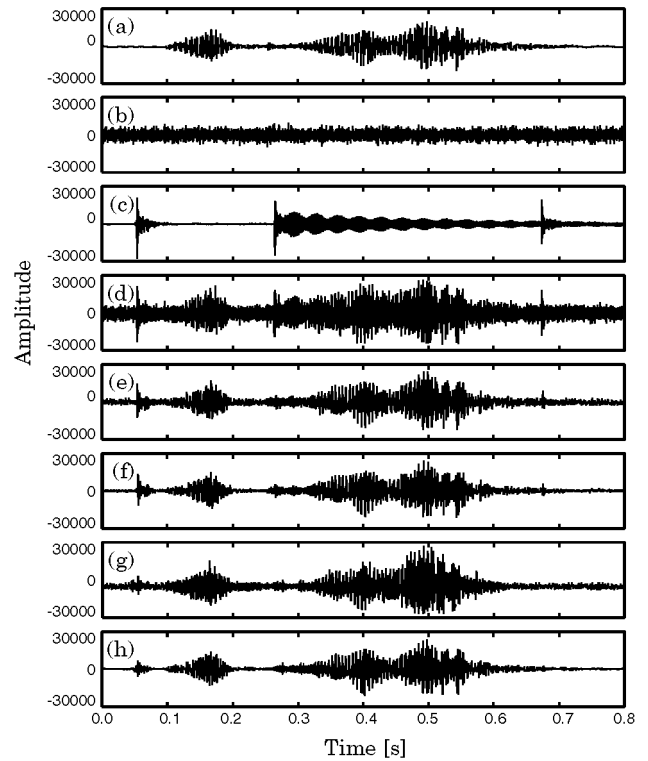


Figure 5: Wave forms: (a) Original speech, (b) Stationary noise, (c) Non-stationary noise, (d) Noise-added speech, (e) Speech enhancement with DS, (f) Speech enhancement with DS+SS, (g) Speech enhancement with DS+WSS, (h) Speech enhancement with proposed method.

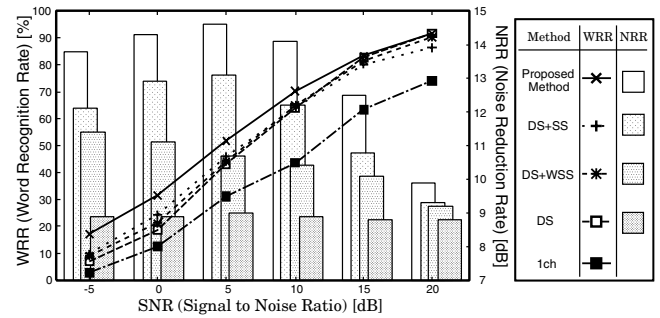


Figure 6: ASR and NRR in the stationary noise, non-stationary noise and target speech environment.

[5] D. Gabor, "Theory of Communications," *J. IEE*, Vol. 93, No. 26, pp. 429–457, Nov. 1946.

[6] I. Daubechies, "The Wavelet Transform, Time-frequency Localization and Signal Analysis," *IEEE Trans. Inf. Theory*, Vol. 36, pp. 961–1005, Sep. 1990.

[7] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, "Data Collection in Real Acoustical Environments for Sound Scene Understanding and Hands-Free Speech Recognition," *Proc. Eurospeech99*, pp. 2255–2258, Sep. 1999.

[8] T. Kawahara, et. al., "Japanese Dictation Toolkit," *J. Acoust. Soc. Jpn. (E)*, Vol. 20, No. 3, pp. 233–239, 1999.