Vision-Referential Speech Enhancement with Binary Mask and Spectral Subtraction

Mitsuharu Matsumoto(✉)

University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo, Japan
mitsuharu.matsumoto@ieee.org

Abstract. This paper proposes vision-referential speech enhancement with binary mask and spectral subtraction as a sensor fusion of visual information and audio information. Recently, we can find many smart phones and tablet devices with a camera and a microphone in the worlds. We improve the sound quality of the audio signal by using the mask information from the visual information. Although the frame rate of the camera in such devices. it will be useful to enhance the speech signal if both signals are used adequately. We therefore aim to design a vision-referential speech enhancement. Throughout the experiments, it was confirmed that the speech could be enhanced even when there was high level of real noise in the environments.

1 Introduction

Speech enhancement, which emphasizes the target speech from mixed signals, is one of the important issues in acoustic processing. Multi-channel approach using the microphone array is a typical approach to enhance the speech signal. Speech enhancement is realized by using the difference between the phase and amplitude of the sound entering each microphone [1–3]. Although it is an attractive approach to enhance the speech, it is necessary to set multiple microphones and to estimate the positions of the microphones. When we set the system for speech enhancement to the machines such as robots and vehicles, the background noise makes their performance worse. In order to solve this problem, we focus on speech enhancement technology in sensor fusion that combines image signal and audio signal. P. Duchnowaski et al. [4] proposed a method for extracting speech from the speaker's lips using image information and tracking the speaker's face based on the extracted lips to assist speech enhancement. H. Kulkarni et al. [5] proposed lip reading that can automatically enhance speech in correspondence with a language data set by combining deep learning and lip reading.

In this way, sensor fusion is a technology that has been attracting attention for a long time. However, speech enhancement technology that simultaneously transmits and receives image signal and audio signal has not been studied so far.

Recently, we proposed a speech enhancement technology of cooperative transmission and reception [6]. In this method, the speaker transmits not only the audio signal through the speaker but also the audio information as an image signal through the display. The listener emphasizes the target acoustical signal from the image and audio

L. Barolli et al. (Eds.): 3PGCIC 2020, LNNS 158, pp. 420–428, 2021.
https://doi.org/10.1007/978-3-030-61105-7_42

signals received through the standard camera and microphone mounted on smartphones and tablets. This framework is a framework for sensor-transmission/reception cooperative sensor fusion. The proposed framework makes it possible to implement speech enhancement that is not affected by the type of external noise, which was difficult with conventional speech enhancement techniques. Non-overlapping noise can be removed by the binary mask even for mixed speech with overlap on the frequency axis since the mask information can be obtained as visual information. However, overlapping noise remains on the time-frequency axis after the binary mask. In this paper, we focus on reducing the overlapping noise by using the spectral subtraction and give some experimental results to show the effectiveness of the proposed approach.

2 Problem Formulation

In this section, we describe the supposed situation and formulate the problem. Figure 1 shows an example of how to use the proposed method. It can be used to send voice to an unspecified number of people, such as an election speech in public spaces. The speaker not only generates the voice through the speaker, but also transmits the mask information of the voice to the listener through the display. When the listener would like to hear the speaker's voice, the listener moves the camera of the smartphone or tablet to the display and captures the audio signal and image information. The listener can obtain the voice information of the speaker with high quality even in a noisy situation by using the proposed method. Let us consider a target signal $s(t)$ and ith noise $n_i(t)$. The mixed speech $x_I(t)$ acquired from the microphone array is described as follows.

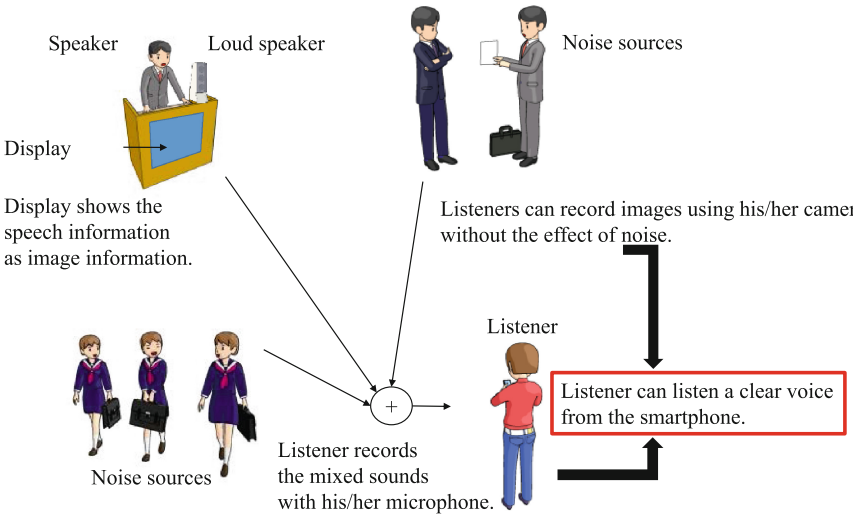

Fig. 1. Assumed usage scenario of the proposed method.

$$x_1(t) = s(t) + \sum_{t=1}^{n} n_t(t) \tag{1}$$

The mask information received as image information is then defined. Even if audio information and image information are sent at the same time, there is a time lag in the signal information received at the receiving side. Therefore, the mask signal X_2 (τ, ω) received by the receiving side can be expressed as follows:

$$X_2(\tau, \omega) = M(\tau - \delta, \omega) \tag{2}$$

where δ is the time delay.

Here, when Δ is defined as the maximum delay between sensors, the following equation is satisfied.

$$|\delta| \leq \Delta \tag{3}$$

Considering that the audio signal received by the microphone is restored by the image signal received by the camera, its output is expected to be maximum when there is no delay. Hence, the delay can be estimated as follows:

$$\tilde{\delta} = \arg \max_{\delta < \Delta} \sum\nolimits_{\tau,\omega} |X_2(\tau + \delta, \omega) X_1(\tau, \omega)| \tag{4}$$

Speech enhancement is performed as follows using the estimated mask information.

$$S(\tau, \omega) = \tilde{M}(\tau, \omega) X_1(\tau, \omega) \forall \tau, \omega \tag{5}$$

where $\tilde{M}$(τ, ω) is the mask information estimated using $\tilde{\delta}$.

3 Proposed Approach

The approach using binary mask works well when the sparseness between the target signal and noise in the time-frequency domain is satisfied. However, if the conventional method is applied to a sound that does not have sparseness, noise remains in the overlapping part. Therefore, we consider an improvement method using speech enhancement by combining a binary mask and a spectral subtraction method as shown in Fig. 1. The spectral subtraction method removes noise by subtracting the estimated value of the average power spectrum of noise from the power spectrum of mixed speech.

Let us suppose the mixed signal $x(t)$ can be written using the signals $s(t)$ and $n(t)$ as

$$x(t) = s(t) + n(t) \tag{6}$$

The signal on the time-frequency axis obtained by Fourier transforming the mixed signal can be described as follows.

$$X(\tau, \omega) = S(\tau, \omega) + N(\tau, \omega) \tag{7}$$

where X(τ, ω), S(τ, ω) and N(τ, ω) are the complex spectra of the $x(t)$, $s(t)$ and $n(t)$, respectively. τ and ω are the time frame and angular frequency, respectively.

In the spectral subtraction method, it is assumed that the signal and noise have no correlation, and is approximated as follows.

$$|X(\tau, \omega)| = |S(\tau, \omega) + N(\tau, \omega)|^2$$

$$|S(\tau,\omega)|^2 + S(\tau,\omega)N^*(\tau,\omega) + S^*(\tau,\omega)N(\tau,\omega) + |N(\tau,\omega)|^2 \tag{8}$$

$$\approx |S(\tau,\omega)|^2 + |N(\tau,\omega)|^2$$

where * indicates a complex conjugate. In the spectral subtraction method, we focus on Eq. (8) and emphasize the target signal by subtracting the estimated noise from the mixed sound. In addition, it is assumed that the noise is stationary in general spectral subtraction. Let the estimated speech spectrum after noise removal be $\tilde{S}(\tau, \omega)$ and the average power spectrum of the estimated noise be $\tilde{N}(\omega)$. $\tilde{S}(\tau, \omega)$ can be estimated as follows:

$$\left|\tilde{S}(\tau,\omega)\right|^2 = |X(\tau,\omega)|^2 - \left|\tilde{N}(\omega)\right|^2 \tag{9}$$

Figure 1 shows the basic concept of the proposed approach in the case that noise remains when the target signal and the noise is overlapped in the time-frequency domain. Figure 2 shows the procedure of the proposed approach to estimate the noise. As shown in Fig. 3, since the mask information is acquired as image information in this method, noise is estimated by averaging the points where the mask is 0 as follows.

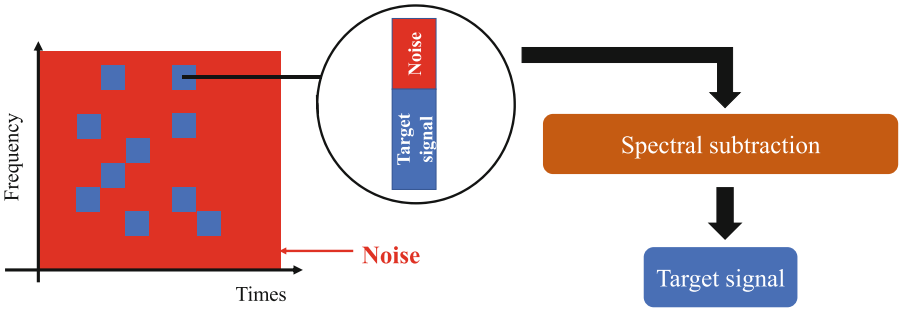

Fig. 2. Noise remains when the target signal and the noise is overlapped in the time-frequency domain. We apply spectral subtraction to reduce the remain noise and extract the target signal.

$$\left|\tilde{N}(\omega)\right|^2 = \frac{1}{T(\omega)} \sum\nolimits_{\tau, M(\tau,\omega)=0} |X(\tau,\omega)| \tag{10}$$

Here, $\sum_{\tau, M(\tau, \omega)=0}$ indicates addition in the τ direction for the point where $M(\tau, \omega) = 0$. In addition, $T(\omega)$ indicates the number of points for which $M(\tau, \omega) = 0$ when added in the τ direction for each ω. Figure 4 shows the process of the proposed method to enhance the target signal. To recover the waveform, not only the power of the signal but also the phase information is needed. Since it is difficult to obtain the phase information of the signal itself from the mixed sound, the phase information of the mixed sound is often used.

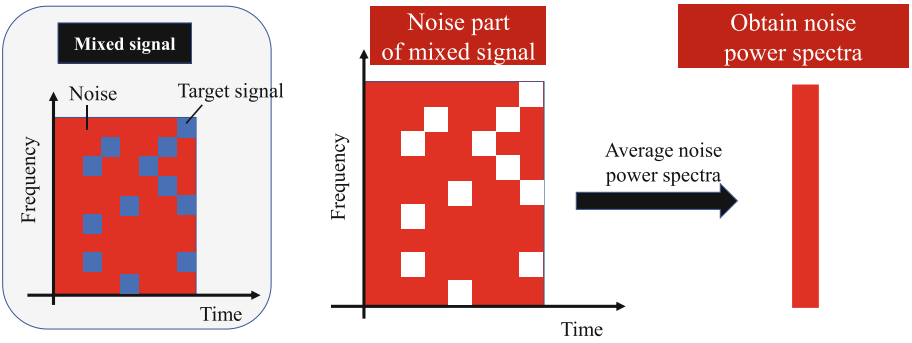

Fig. 3. Procedure of the proposed approach to estimate the noise. Noise parts are estimated by averaging the noise part based on the mask information.

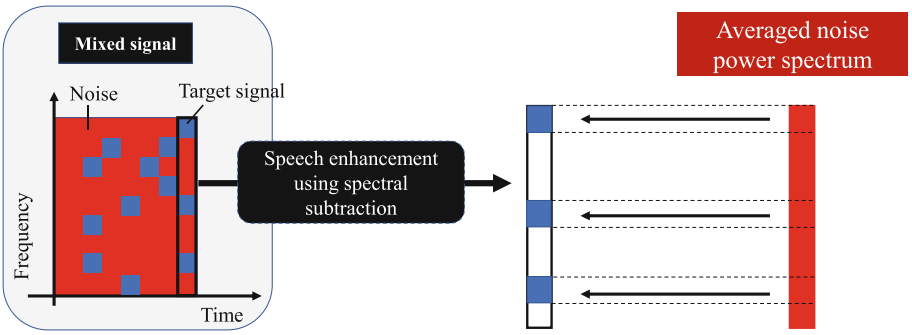

Fig. 4. Procedure of the proposed approach to enhance the speech. Speech is enhanced by subtracting the estimated noise from the mixed noise.

4 Experiments

In this section, we describe the experimental condition and the results. The target voice was a female voice from the ATR newspaper reading database. All programs were created using Python in Visual Studio 2017 Community. As noise, pink noise and white noise were used. Three levels of noise were set: 0, −10, −20 dB.

Table 1 shows the experimental condition. We set the threshold values to create the binary mask from −100 to −30 dB with 10 dB intervals. Signal-to-Distortion Ratio (SDR) was used to evaluate the waveform error between the target speech and the speech after noise removal [7]. Generally, the larger the SDR value, the smaller the distortion for the signal. Let $S\ (\tau,\ \omega)$ be the complex amplitude of the target speech in the time-frequency domain, and let $\tilde{S}(\tau,\ \omega)$ be the power of the signal after speech enhancement.

$$SDR = 10\ \log_{10}\left(\frac{\sum_{\tau,\omega}|S(\tau,\omega)|}{\sum_{\tau,\omega}\left||S(\tau,\omega|-\lambda\left|\tilde{S}(\tau,\omega\right|\right|}\right) \tag{11}$$

Table 1. Experimental condition.

Target signal	Female voice
Noise signal	White noise, Pink noise
Noise level	−20 dB, −10 dB, 0 dB
Threshold	−100 to −30 dB (10 dB interval)

where λ is described as follows:

$$\lambda = \sqrt{\frac{\sum_{\tau,\omega} |S(\tau,\omega)|^2}{\sum_{\tau,\omega} \left|\tilde{S}(\tau,\omega)\right|^2}} \tag{12}$$

Table 2 to Table 4 show the experimental results when various levels of white noise (−20 dB, −10 dB and 0 dB) were used for the experiments, respectively. Table 5 to Table 7 show the experimental results when various levels of pink noise (−20 dB, −10 dB and 0 dB) were used for the experiments. The fond was set to the bold to show the maximum value among all the tested thresholds. From Tables 2, 3, 4, 5, 6 and 7, it can be confirmed that the value of the signal-to-distortion ratio is greatly increased by the proposed method in all the cases, especially when the threshold value is low. When the threshold is high, we succeeded in improving the accuracy when the threshold is low while maintaining the accuracy of the conventional method. The reason why the SDR improves when the threshold is low is thought to be because a lot of noise components remain in the mixed signal.

Table 2. Experimental results when white noise (−20 dB) was used.

Threshold (dB)	Mixed signal (dB)	Conventional approach (dB)	Proposed approach (dB)
−90	−3.129	2.064	3.078
−80	−3.129	4.791	5.537
−70	−3.129	6.036	**6.126**
−60	−3.129	5.258	5.239
−50	−3.129	3.766	3.765
−40	−3.129	2.086	2.087
−30	−3.129	0.4700	0.4701

Table 3. Experimental results when white noise (−10 dB) was used.

Threshold (dB)	Mixed signal (dB)	Conventional approach (dB)	Proposed approach (dB)
−90	−7.024	−2.829	−2.000
−80	−7.024	0.02104	0.8847
−70	−7.024	3.201	3.768
−60	−7.024	4.190	**4.249**
−50	−7.024	3.438	3.418
−40	−7.024	2.053	2.053
−30	−7.024	0.4701	0.4703

Table 4. Experimental results when white noise (0 dB) was used.

Threshold (dB)	Mixed signal (dB)	Conventional approach (dB)	Proposed approach (dB)
−90	−8.626	−5.824	−5.625
−80	−8.626	−3.972	−3.677
−70	−8.626	−1.298	−0.8659
−60	−8.626	1.103	1.506
−50	−8.626	2.222	**2.298**
−40	−8.626	1.735	1.698
−30	−8.626	0.4708	0.4722

Table 5. Experimental results when pink noise (−20 dB) was used.

Threshold (dB)	Mixed signal (dB)	Conventional approach (dB)	Proposed approach (dB)
−90	−0.9651	3.804	4.680
−80	−0.9651	5.563	6.080
−70	−0.9651	6.050	**6.109**
−60	−0.9651	5.188	5.174
−50	−0.9651	3.721	3.719
−40	−0.9651	2.081	2.082
−30	−0.9651	0.4696	0.4697

Table 6. Experimental results when pink noise (−10 dB) was used.

Threshold (dB)	Mixed signal (dB)	Conventional approach (dB)	Proposed approach (dB)
−90	−4.953	−1.114	−0.2876
−80	−4.953	1.116	1.906
−70	−4.953	3.350	3.813
−60	−4.953	3.968	4.038
−50	−4.953	3.293	3.271
−40	−4.953	2.018	2.017
−30	−4.953	0.4692	0.4696

Table 7. Experimental results when pink noise (0 dB) was used.

Threshold (dB)	Mixed signal (dB)	Conventional approach (dB)	Proposed approach (dB)
−90	−6.459	−4.045	−3.701
−80	−6.459	−2.318	−1.896
−70	−6.459	−0.03450	0.4462
−60	−6.459	1.724	2.052
−50	−6.459	2.319	2.352
−40	−6.459	1.718	1.690
−30	−6.459	0.4712	0.4733

5 Conclusion

In this research, we proposed a vision-referential speech enhancement with binary mask and spectral subtraction. To check the validity of the proposed methods, we prepared white noise and pink noise with different noise level and checked the effectiveness of the proposed method. From the experimental results, the proposed method could effectively remove noise that is not sparse compared to the previous method. Since the proposed method is not effective against fluctuating noise, we consider a method that can cope with fluctuating noise. We also consider real-time processing in the future.

Acknowledgments. This research was supported by the research grant of Support Center for Advanced Telecommunications Technology Research and by the research grant of Foundation for the Fusion of Science and Technology. I also would like to thank Mr. Suzuki.

References

1. Jarrett, D.P.: Theory and Applications of Spherical Microphone Array Processing. Springer, Berlin (2017). https://doi.org/10.1007/978-3-319-42211-4

2. Yu, C., Su, L.: Speech enhancement based on the generalized sidelobe cancellation and spectral subtraction for a microphone array. In: Proceedings of 8th International Congress on Image And Signal Processing, pp. 1318–1322 (2015)
3. Doclo, S., Moonen, M.: Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics. IEEE Trans. Sign. Process. **51**(10), 2511–2526 (2003)
4. Duchnowski, P., Hunke, M., Busching, D., Meier, U., Waibel, A.: Toward movement-invariant automatic lip-reading and speech recognition. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 109–112 (1995)
5. Kulkarni, A.H., Kirange, D.: Artificial intelligence: a survey on lip-reading techniques. In: International Conference on Computing, Communication and Networking Technologies, p. 45670 (2019)
6. Matsumoto, M.: Vision-referential speech enhancement of an audio signal using mask information captured as visual data. J. Acoust. Soc. Am. **145**(1), 338–348 (2019)
7. Fukui, K., Shimauchi, S., Nakagawa, A., Hioka, Y., Haneda, Y., Ohmuro, H., Kataoka, A.: Noise-power estimation based on ratio of stationary noise to input signal for noise reduction. J. Sign. Process. **18**(1), 17–28 (2014)