Parameter estimation of a method for noise reduction fusing a normal microphone and a bone conduction microphone

Junki Kawaguchi*, Mitsharu Matsumoto†
*Deptartment of Informatics, The University of Electro-Communications., Tokyo, Japan
Email: kawaguchi@mm-labo.org
†Deptartment of Informatics, The University of Electro-Communications., Tokyo, Japan
Email: mitsuharu.matsumoto@ieee.org

***Abstract*— This study proposes a method for setting a parameter used in a noise reduction method fusing a typical microphone and a bone conduction microphone. A bone conduction microphone is a microphone that captures sound by measuring vibrations of the objects directly. When human voice is detected, it is usually attached around the neck. Due to its feature, bone conduction microphones can record the sound without noise compared to typical microphones. On the other hand, bone conduction microphones have different acoustic characteristics from normal microphones, and sound quality deteriorates. The proposal method uses the signal from the bone conduction microphone as reference information. The voice quality is improved while suppressing the noise contained in the signal of the normal microphone by referring to the signal of the bone conduction microphone. This approach requires the selection of an appropriate threshold, which was previously set manually. This paper introduces a method to estimate the parameter automatically.**

Keywords—Parameter estimation, signal noise decorrelation, sensor fusion

I. INTRODUCTION

With the spread of mobile devices such as smartphones, the importance of voice input is increasing. Voice input devices can be used comfortably in a noise-free environment, but are difficult to use in a noisy environment. In such an environment, users have difficulty in hearing voices or using mobile devices. Hence, speech enhancement technology has become an important issue in acoustic processing.

A speech enhancement technique is a technique for enhancing a target signal from a mixed signal in which signals and noise are mixed. There are two approaches to speech enhancement, depending on the number of signals used.

One method is to use a single signal. The other method is to use multiple signals. According to the survey on speech enhancement [1], we can divide the algorithms of a monaural noise reduction into three types:

Spectral subtraction (SS) is a famous approach for single-channel noise reduction. In the first trial, Weiss. et. al., execute the SS method in the correlation domain [2]. Boll et.al also proposed a SS method in the time-frequency domain [3]. The SS method is a simple and useful technique available for single-channel speech enhancement, and various improvements have been proposed [4]. Another approach uses statistical model to achieve the noise reduction of a single channel signal. As an example, some authors aimed to estimate the spectrum of the objective signal utilizing the maximum likelihood method [5]. The other approach uses subspace to realize the noise reduction. Some authors proposed a speech enhancement method utilizing singular value de-composition [6]. Other study realized a speech enhancement method using eigenvalue decomposition of the signal [7]. Speech enhancement from a single signal is an attractive technique because it can be achieved using only recorded sound sources. However, the information available in speech enhancement for a speech signal is limited to in-formation about the sound source itself. Speech enhancement technology using deep learning has been actively researched in recent years, but it requires computational cost [8][9][10].

When we try to reduce the noise with multiple signals, the microphone array is often used [11][12]. It eliminates noise by utilizing the phase difference and gain difference of sounds recorded by the microphones.

A well-known example is sound focus, which emphasizes the sound by aligning the phase with the direction of the target sound. Another well-known example is adaptive arrays that are desensitized to noise directions. Blind source separation is also widely studied in the field of noise reduction using multi-channels. Independent component analysis (ICA) [13][14][15] and sound separation based on sparsity [16][17] have long been studied. However, they require special equipment such as many microphones for implementation.

We can also reduce the effect of noise by making the microphone different from a normal microphone. A bone conduction microphone is a microphone that captures sounds by measuring the vibration of the neck during vocalization, and is less susceptible to ambient noise. There is also research aimed at improving the accuracy of speech recognition by focusing on the characteristics of bone conduction microphones [18]. The characteristics of bone conduction microphones are effective for voice enhancement. However, the bone conduction microphone and the typical microphone have different acoustic characteristics. The sound quality of the bone conduction microphone is not very good. To solve this problem, we consider sensor fusion of bone conduction microphone and normal microphone.

Sensor fusion is actively been studied in image processing fields. Several investigations have been reported on fusing sensors to improve the quality of the obtained images. Cross bilateral filtering is a method of obtaining better quality images by fusing a photograph with flash and a photograph

979-8-3503-5475-1/24/$31.00 ©2024 IEEE

without flash taken at the same location [19][20]. It is also called joint bilateral filtering.

There are also methods that combines infrared camera images and visible camera images assuming the detection of human faces [21], walking persons [22] and vehicles [23].

Based on these studies in the field of image processing, the authors proposed a sensor fusion for speech enhancement. In the proposal method, we refer to the speech signal obtained with a bone conduction microphone and generate the binary mask based on the information, and the unnecessary noise is removed from the mixed signal recorded by a normal microphone [24]. However, it requires a parameter to be determined for speech enhancement. Although it is desirable to be set in advance if possible, the appropriate parameter differs depending on the noise, making it difficult to set the parameters in advance. We investigate an approach for automatic determination of the parameter used in noise reduction methods that combine the bone conduction microphone with the normal microphone in this paper. We consider the automatic parameter setting by using correlation coefficient [25][26]. We demonstrate the effect of the proposal approach through experiments.

We introduce a basic binary mask method with two microphones to make it easier to understand the sensor fusion dealt with in this paper in section 2. In Section 3, the sensor fusion technique and the method for parameter setting are described. In Section 4, we perform experiments utilizing the proposal approach and investigate the appropriateness of the parameters set by the proposal method. We discuss the results and give conclusions and prospects in Section 5.

II. Typical Approach Using Binary Mask with Two Microphones

A. Formatting the Problems

In this section, binary mask using two normal microphones is introduced [27-29]. Two microphones are assumed in the typical binary mask approach. Let us consider $x_1(t)$ and $x_2(t)$ from the microphone 1 and 2, respectively. Regarding $x_1(t)$, we can remove the attenuation and delay without the generality of the problem. In this case, $x_1(t)$ is describe as follows:

$$x_1(t) = s(t) + \sum_{i=1}^{N} n_i(t) \quad (1)$$

where $s(t)$ and $n_i(t)$ are the objective signal and the ith noise ($i = 1, 2, 3, ..., N$), respectively. Regarding $x_2(t)$, the delay and attenuation with respect to $x_1(t)$ should be considered. From the above aspect, we can describe $x_2(t)$ as

$$x_2(t) = as(t - \delta) + \sum_{i=1}^{N} a_i n_i(t - \delta_i), \quad (2)$$

where δ indicates the delay of the target signal with respect to $x_1(t)$. δ_i indicates the delay of the ith noise ($i = 1, 2, 3, ..., N$). a is the relative attenuation of the target signal with respect to $x_1(t)$. ai is the relative attenuation of the ith noise with respect to $x_1(t)$. Let Δ be the maximum delay between $x_1(t)$ and $x_2(t)$. Then, we can constrain δ and δ_i as follows:

$$|\delta| \leq \Delta \quad (3)$$

$$|\delta_i| \leq \Delta, \quad (4)$$

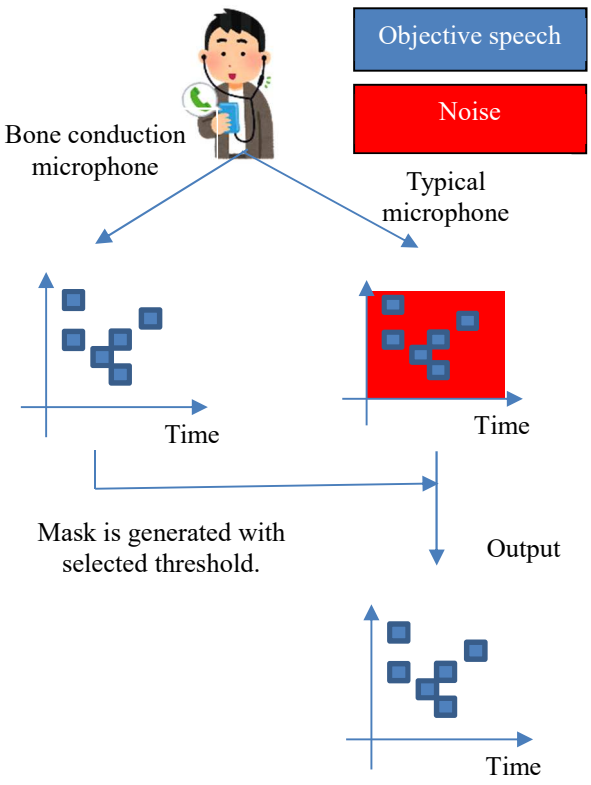

Fig. 1. Basic concept of the proposed approach.

Since we assume the sparsity between the signal and the noise in the time-frequency domain, the time-frequency signals are supposed to be disjoint with respect to the signal and the noise. Let us consider $S(\tau, \omega)$ as the time-frequency signal of the target signal $s(t)$. Let us also consider $N_j(\tau, \omega)$ as time-frequency signal of the jth noise $n_j(t)$, respectively. We can describe $S(\tau, \omega)$ by utilizing the short time Fourier transform as

$$S(\tau, \omega) = \sum_{\tau=-\infty}^{\infty} s(t + \tau) W(\tau) \exp(-i\omega\tau) \quad (5)$$

where $W(\tau)$ indicates the window function. τ indicates the time frame. ω denotes the angular frequency. We also can describe $N_i(\tau, \omega)$ by utilizing the short time Fourier transform as

$$N_j(\tau, \omega) = \sum_{\tau=-\infty}^{\infty} n_j(t + \tau) W(\tau) \exp(-i\omega\tau) \quad (6)$$

When we can assume the signal-noise sparsity, the following constraint is met:

$$S(\tau, \omega) N_i(\tau, \omega) = 0 \ \forall \tau, \omega, \quad (7)$$

B. Noise Reduction Using Binary Mask

When the speech enhancement is executed, we need to estimate the parameters δ, δ_i, a and a_i in Eq. (2) are estimated. Let us define $X_i(\tau, \omega)$ as the time-frequency signal of $x_i(t)$. $X_i(\tau, \omega)$ is described as

$$X_i(\tau, \omega) = \sum_{t=-\infty}^{\infty} x_i(t + \tau) W(t) \exp(-i\omega t), \quad (8)$$

We can rewrite the signals from the microphone 1 and the microphone 2 in the time-frequency domain as follows.

$$\begin{bmatrix} X_1(\tau,\omega) \\ X_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & a\exp(-i\omega\delta) \\ 1 & a_1\exp(-i\omega\delta_N) \\ \vdots & \vdots \\ 1 & a_N\exp(-i\omega\delta_N) \end{bmatrix}^T \begin{bmatrix} S(\tau,\omega) \\ N_1(\tau,\omega) \\ \cdots \\ N_N(\tau,\omega) \end{bmatrix} (9)$$

where T represents transpose. When we can assume the sparsity between the speech signal and the noise, the following constraint is met:

$$\begin{bmatrix} X_1(\tau,\omega) \\ X_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ a(\tau_s,\omega_s)\exp(-i\omega_s\delta(\tau_s,\omega_s)) \end{bmatrix}^T s(\tau_s,\omega_s) (10)$$

where τ_s and ω_s indicate the time and frequency at which the signal is present, respectively. The ratio of $X_1(\tau_s, \omega_s)$ and $X_2(\tau_s, \omega_s)$ is calculated to estimate τ_s and ω_s. Let us consider $a(\tau_s, \omega_s)$ and $\delta(\tau_s, \omega_s)$ as relative amplitude and the relative delay.

$a(\tau_s, \omega_s)$ and $\delta(\tau_s, \omega_s)$ can be estimated as

$$\left(a(\tau_s,\omega_s), \delta(\tau_s,\omega_s)\right) = \left(\left|\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}\right|, \frac{1}{\omega}\angle\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}\right) \quad (11)$$

where $\angle a\exp(i\phi)$ represents the angle of $a\exp(i\phi)$ and is described as follows:

$$\angle a\exp(i\phi) = \phi, \ -\pi < \phi < \pi, \quad (12)$$

The binary mask $M(\tau, \omega)$ for (τ, ω) can be designed as

$$M(\tau,\omega) = \begin{cases} 1, |\ln a(\tau,\omega) - \ln a| < \frac{\Delta_a}{2} \wedge |\ln \delta(\tau,\omega) - \ln \delta| < \frac{\Delta_\delta}{2} \\ 0, otherwise \end{cases} (13)$$

where Δ_a denotes the amplitude resolution. Δ_δ denotes delay resolution. In the typical noise reduction using the binary mask, we apply the masking to the $X_1(\tau, \omega)$ as follows.

$$S(\tau,\omega) = M(\tau,\omega)X_1(\tau,\omega), \quad (14)$$

The process of the binary mask is simple and can eliminate the noise effectively. However, the performance of the binary mask is strongly affected by the mask estimation accuracy. When omnidirectional noise is present in the environment, it is difficult to estimate the mask accurately as the sparsity assumptions are not satisfied.

III. Noise Reduction Fusing a Normal Microphone and a Bone Conduction Microphone

A. Basic Concept of the Proposal Approach

To clarify the concept of noise reduction fusing a normal microphone and a bone conduction microphone, we describe the outline of the proposal approach. The usage scenario of the proposal approach is given in Figure 1. In this method, the sounds were recorded not only by a normal microphone but also a bone conduction microphone simultaneously.

The binary mask is generated by using the signal from the bone conduction microphone. We then apply the created binary mask to the signal from the normal microphone and reduce the noise. For this goal, the audio signals are recorded with both a bone conduction microphone and a normal microphone. The time-frequency signal is obtained from the obtained signal by Fourier transform. To create the binary mask an appropriate threshold is set. We also apply Fourier transform to audio data from a normal microphone and transforms it into frequency domain data. We apply a generated binary mask to frequency data obtained from the normal microphone to enhance speech. The output waveform is obtained by inverse Fourier transforming the voice-enhanced signal. To obtain the adequate parameter automatically, we assume that the target signal and noise are non-correlated.

Under this assumption, the correlation coefficient of the signal after filtering and the difference value between the signal from a typical microphone and the signal after noise reduction is calculated. The parameter is obtained as the signal with the lowest correlation coefficient.

B. Problem formulation

To clarify the concept of noise reduction fusing a normal microphone and a bone conduction microphone, we describe the outline of the proposal approach. The usage scenario of the proposal approach is given in Figure 1. In this method, the sounds were recorded not only by a normal microphone but also a bone conduction microphone simultaneously.

The problem is formulated based on the concept described in the previous section. Let us define $x_1(t)$ as the signal recorded by the normal microphone. Let us define $x_2(t)$ as the signal recorded by a bone conduction microphone. t represents the time. $X_1(\tau, \omega)$ and $X_2(\tau, \omega)$ represent the spectra of $x_1(t)$ and $x_2(t)$ in the frequency domain, respectively. τ denotes the time frame. ω denotes the angular frequency. Let $S(\tau, \omega)$ and $N_i(\tau, \omega)$ be the spectrum of the objective signal is the spectrum of the ith noise. $X_1(\tau, \omega)$ is represented as follows.

$$X_1(\tau,\omega) = S(\tau,\omega) + \sum_{i=1}^{N} N_i(\tau,\omega) \quad (15)$$

When we consider the signal of the bone conduction microphone, we can assume that the signal does not include noise signals. However, bone conduction microphones and normal microphones have different frequency characteristics. Hence, $X_2(\tau, \omega)$ is expressed as follows.

$$X_2(\tau,\omega) = B(\omega)S(\tau,\omega), \quad (16)$$

where $B(\omega)$ denotes the frequency characteristic of a bone conduction microphone to a normal microphone. The binary mask $M_j(\tau, \omega)$ is generated based on the information from the bone conduction microphone as follows.

$$M_j(\tau,\omega) = \begin{cases} 1 & |X_2(\tau,\omega)| \geq th_j \\ 0 & |X_2(\tau,\omega)| < th_j \end{cases} \quad (17)$$

where th_j is a jth threshold.

The obtained binary mask is applied to the time-frequency signal $X_1(\tau, \omega)$ to remove noise of the microphone. The output $Y_j(\tau, \omega)$ can be obtained as follows:

$$Y_j(\tau,\omega) = M_j(\tau,\omega)X_1(\tau,\omega) \quad (18)$$

The output signal $y_j(t)$ is obtained from $Y_j(\tau, \omega)$ by using inverse Fourier transform. To do the above process, we need to set the adequate threshold th_j. Since the acoustic characteristics of a bone conduction microphone and a microphone are different, the threshold value cannot be set using the bone conduction microphone signal. Therefore, it is necessary to set an appropriate threshold value based on the situation where only the mixed sound of the microphone exists. To solve this problem, we assume that the signal and noise are uncorrelated. Figure 2 shows a conceptual diagram of the threshold estimation method when this assumption holds.

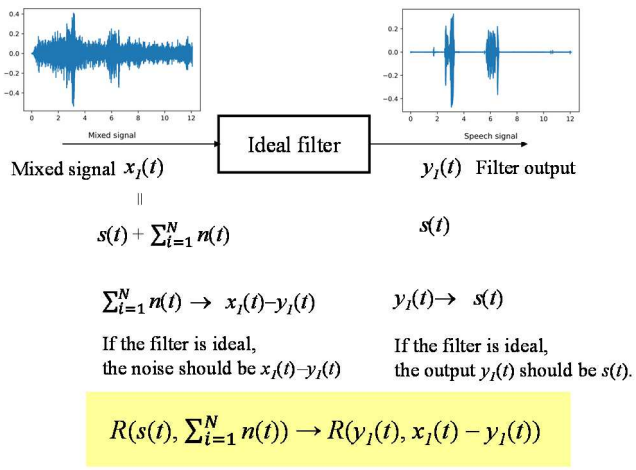

Fig. 2. Basic concept of the parameter estimation under signal-noise decorrelation.

TABLE I. EXPERIMENTAL CONDITION

Speech signal	Speech by a man
Noise signal	Intersection noise, Restaurant noise, Station noise
Thresholds for noise reduction	From -90[dB] to -30[dB] with 10[dB] intervals

TABLE II. SDR (INTERSECTION NOISE)

Threshold (dB)	SDR
-90	0.704
-80	2.632
-70	8.384
-60	9.593
-50	7.393
-40	4.351
-30	1.527

TABLE III. SDR (RESTAURANT NOISE)

Threshold (dB)	SDR
-90	3.370
-80	4.948
-70	5,762
-60	6.801
-50	5.531
-40	3.653
-30	1.352

TABLE IV. SDR (STATION NOISE)

Threshold (dB)	SDR
-90	7.417
-80	9.144
-70	10.02
-60	9.335
-50	7.290
-40	4.386
-30	1.527

TABLE V. CORRELATION COEFFICIENT (INTERSECTION NOISE)

Threshold (dB)	Correlation coefficients
-90	0.053
-80	0.136
-70	0.027
-60	0.033
-50	0.015
-40	0.031
-30	0.051

TABLE VI. CORRELATION COEFFICIENT (RESTAURANT NOISE)

Threshold (dB)	Correlation coefficients
-90	0.067
-80	0.025
-70	0.033
-60	0.053
-50	0.047
-40	0.060
-30	0.034

TABLE VII. CORRELATION COEFFICIENT (STATION NOISE)

Threshold (dB)	Correlation coefficients
-90	0.052
-80	0.014
-70	0.027
-60	0.030
-50	0.017
-40	0.045
-30	0.054

As shown in Fig.2, when we consider the ideal filter with the adequate parameter, the filter output $y_I(t)$ should be $s(t)$. In this case, the noise signal can be obtained by calculating $x_I(t) - y_j(t)$. Hence, if the parameter is adequate, it is expected that the correlation between $x_I(t) - y_j(t)$ and $y_j(t)$ becomes minimal.

Under the above consideration, we delimit the audio signal and calculate the correlation coefficient of the signal after filtering and the difference value between the signal from a typical microphone and the signal after noise reduction. The correlation coefficient $R(x_I(t) - y_j(t), y_j(t))$ is expressed as follows.

$$R(x_1(t) - y_j(t), y_j(t)) = \frac{\sum_{t=1}^{L}(x_1(t) - y_j(t) - \overline{x_1(t) - y_j(t)})(y_i(t) - \overline{y_j(t)})}{\sqrt{\sum_{t=1}^{L}(x_1(t) - y_j(t) - \overline{x_1(t) - y_j(t)})}\sqrt{\sum_{t=1}^{L}(y_j(t) - \overline{y_j(t)})^2}} \quad (19)$$

where L is the amount of data in the signal. $\overline{x_1(t) - y_j(t)}$ and $\overline{y_j(t)}$ are the averages of $x_I(t) - y_j(t)$ and $y_j(t)$, and are described as follows

$$\overline{x_1(t) - y_j(t)} = \frac{1}{L}\sum_{t=1}^{L}(x_1(t) - y_j(t)) \quad (20)$$

$$\overline{y_j(t)} = \frac{1}{L}\sum_{t=1}^{L}(y_j(t)) \quad (21)$$

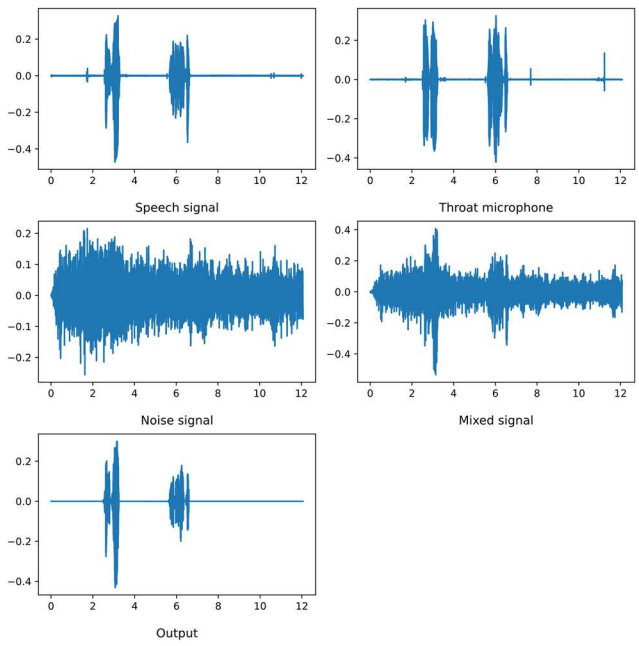

Fig. 3. Waveform of the signal from the typical microphone, the signal from the bone conduc-tion microphone, the noise, the mixed sound, the output signal when using the pa-rameters obtained by the proposal method (Intersection noise)

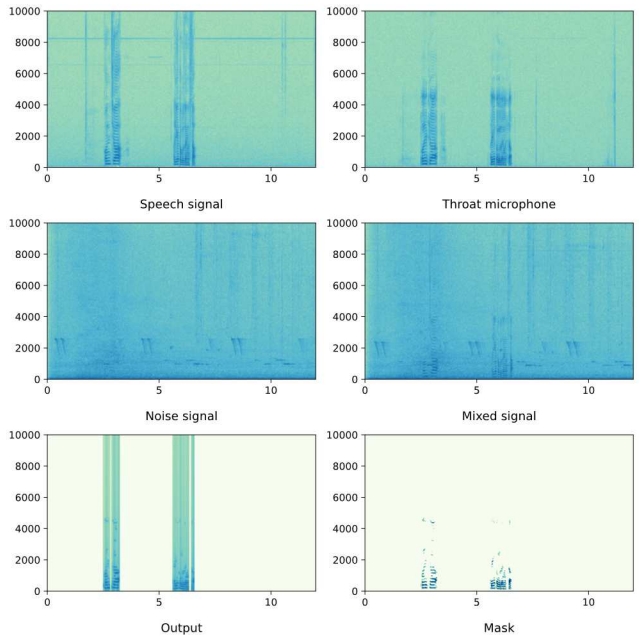

Fig. 4. Spectrogram of the signal from the typical microphone, the signal from the bone con-duction microphone, the noise, the mixed sound, the output signal when using the pa-rameters obtained by the proposal method, and the obtained mask (Intersection noise)

Here, as it is assumed that the speech signal and noise signal are not correlated, $y_j(t)$ is output when the correlation coefficient is value near 0.

Hence, it is expected that we are able to estimate the adequate parameter by using the threshold where the correlation coefficient becomes minimal.

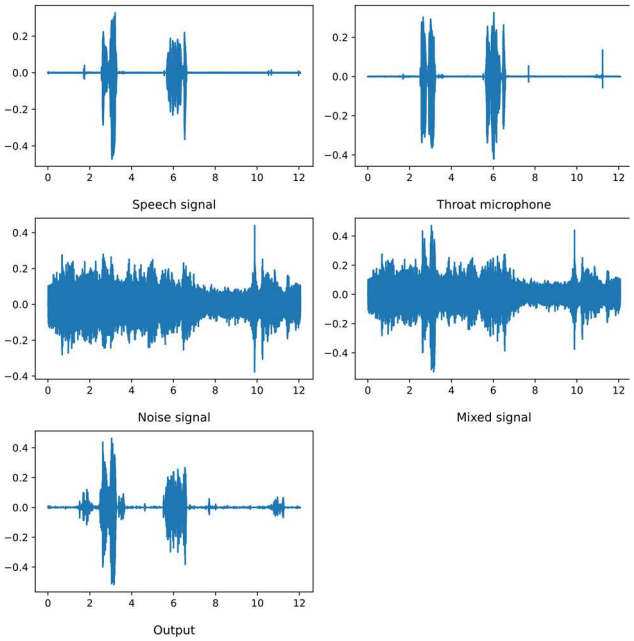

Fig. 5. Waveform of the signal from the typical microphone, the signal from the bone conduc-tion microphone, the noise, the mixed sound, the output signal when using the pa-rameters obtained by the proposal method (Restaurant noise)

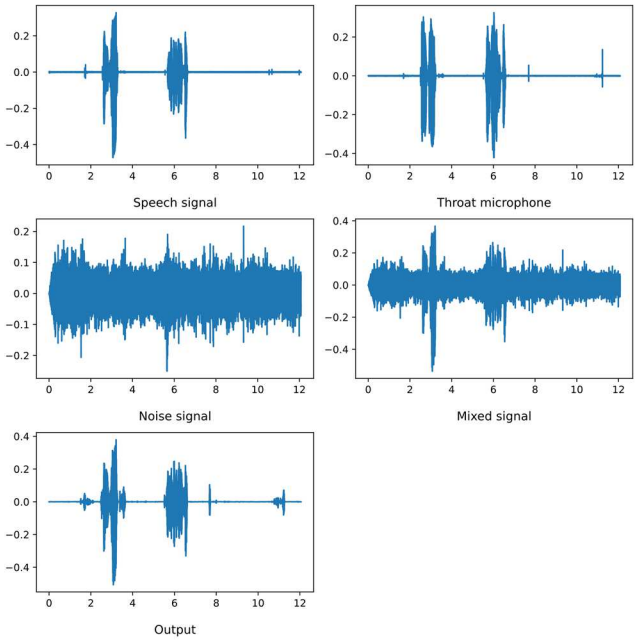

Fig. 7. Spectrogram of the signal from the typical microphone, the signal from the bone con-duction microphone, the noise, the mixed sound, the output signal when using the pa-rameters obtained by the proposal method, and the obtained mask (Station noise)

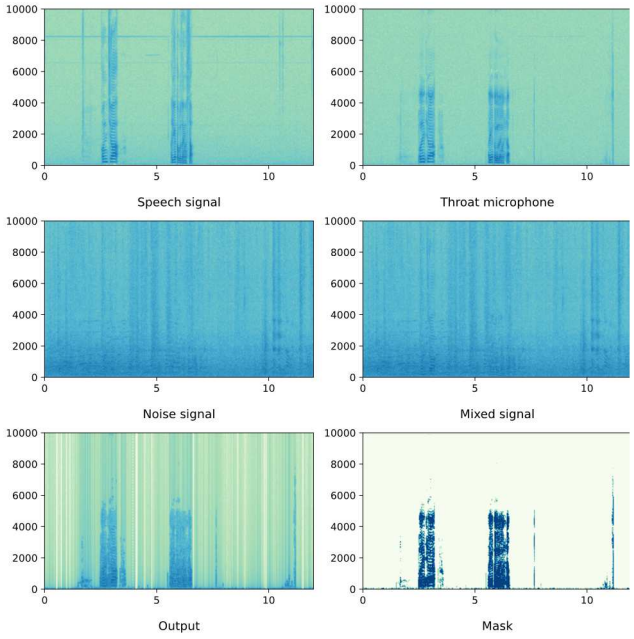

Fig. 6. Spectrogram of the signal from the typical microphone, the signal from the bone con-duction microphone, the noise, the mixed sound, the output signal when using the pa-rameters obtained by the proposal method, and the obtained mask (Restaurant noise)

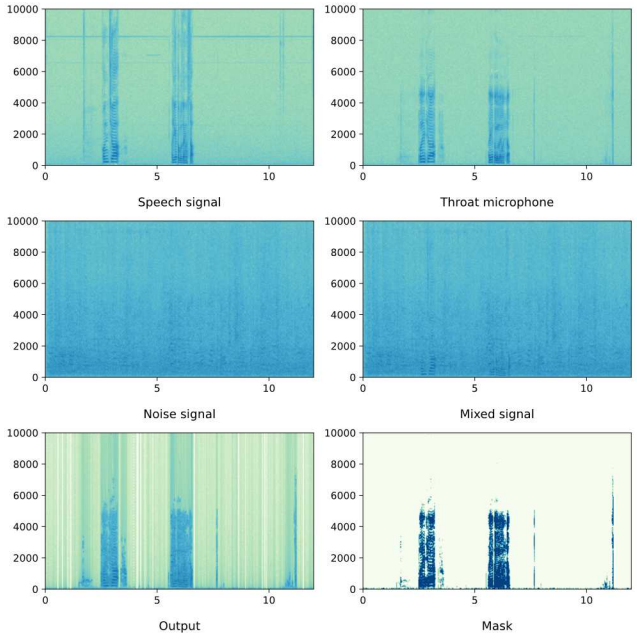

Fig. 8. Spectrogram of the signal from the typical microphone, the signal from the bone con-duction microphone, the noise, the mixed sound, the output signal when using the pa-rameters obtained by the proposal method, and the obtained mask (Station noise)

IV. Experiments

A. Experimental Overview

The performance of the proposal method is investigated based on an experimental basis. In the experiments, the mixed signal is created on a computer from the prepared target signal and noise signal, and the performance of noise removal isverified. Noises were selected from the Sound Effects Lab [30]. We utilized the recorded voice for the voice. We M4U made by inMusic, Inc as a normal microphone.

We also used DN-915129 made by ThirdWave Co., Ltd as a bone conduction microphone. To check the effectiveness of the proposal approach, we generated the mixed signals on a computer. We used Python to write all the programs. Table 1 shows the experimental conditions used in the experiments. We used three types of noise, that is, intersection noise, restaurant noise and station noise. Sound level is expressed in dBFS. We changed the threshold value in 10 dB intervals to check the effect of the parameter.

The window function was set to a Hamming window. We used Signal to Distortion Ratio (SDR) [31] to evaluate the quality of the denoised speech compared to the mixed signal and the target signal. SDR can measure how much the obtained signal after noise reduction is distorted compared to the target speech. We can define SDR as follows.

$$SDR = 10\log_{10}\left(\frac{\sum_{\tau,\omega}|S(\tau,\omega)|}{\sum_{\tau,\omega}\left||S(\tau,\omega)|-\lambda|\hat{S}(\tau,\omega)|\right|}\right) \quad (22)$$

Here $\hat{S}(\tau, \omega)$ represents the signal to be compared to the objective signal. $S(\tau, \omega)$ represents the objective signal. λ indicates a parameter for normalizing the power of $\hat{S}(\tau, \omega)$. We can describe it as

$$\lambda = \sqrt{\frac{\sum_{\tau,\omega}|S(\tau,\omega)|}{\sum_{\tau,\omega}|\hat{S}(\tau,\omega)|}} \quad (23)$$

B. Experimental Results

Tables 2 to 4 show the relation between the SDR values and the threshold value. The results show the results of noise reduction when we employed three types of noise, respectively. The experimental results are described with 4 significant digits. According to Tables 2 to 4, the optimal thresholds under each condition were -60 for intersection noise, -60 for restaurant noise, and -70 for station noise, respectively.

Tables 5 to 7 show the correlation coefficients of the signal after filtering and the difference between the signal from a typical microphone and the signal regarding intersection noise, restaurant noise and station noise, respectively. We give the experimental results in four significant digits.

As it is assumed that the objective signal and noise signal are not correlated, the adequate parameter is expected to be obtained when the correlation coefficient is close to zero. Therefore, -50 for intersection noise, -80 for restaurant noise, and -80 for station noise are optimal from the perspective of the decorrelation criterion. Compared to the results in the case of SDR, the error is -20 at maximum, which is generally good.

To show the appropriateness of the obtained parameter, we also give the waveform of the signal from the typical microphone, the signal from the bone conduction microphone, the noise, the mixed sound, the output signal when we set the obtained parameter to the system. Figure 3 shows the waveform when we utilized intersection noise as the noise. Figure 4 shows the spectrogram when we utilized intersection noise. Figure 5 shows the waveform when we utilized restaurant noise as noise. Figure 6 shows the spectrogram when we utilized restaurant noise. Figure 7 shows the waveform when we utilized station noise as noise. Figure 8 shows the spectrogram when we utilized station noise.

Fig.3, 5 and 7 include the waveform of the signal from the typical microphone, the signal from the bone conduction microphone, the noise, the mixed sound, the output signal when the obtained parameter was used. Fig. 4, 6, and 8 include the spectrogram of the signal from the typical microphone, the signal from the bone conduction microphone, the noise, the mixed sound, the output signal when the obtained parameter was used. To clarify the mask calculated by the obtained parameter, we also show the mask of each case in Fig. 4, 6 and 8.

V. Discussion and Conclusion

In this research, we introduced a method to automatically set the parameter in a noise reduction method that uses a bone conduction microphone and a normal microphone. In the proposal approach, we employed decorrelation criterion to set the parameter.

In the experiments, we used three types of noise, that is, intersection noise, restaurant noise, and station noise. The results of SDRs and correlation values were compared with those obtained with the proposal method. The experimental results on parameter setting show that the proposal method can select a value close to the optimal threshold value. In the approach for noise reduction, the binary mask is generated by using the signal from the bone conduction microphone. It is applied to the signal from a typical microphone to reduce its noise. Although the threshold value should be determined manually in the original approach, we could estimate a good threshold by using the assumption on signal-noise decorrelation.

As the adequate parameter could be obtained by using the proposal method, we would like to develop a real time system for future works. In addition, we would like to proceed with experiments assuming practical applications.

References

[1] P. C. Loizou, Speech Enhancement: Theory and Practice (2 ed.), 2007, CRC Press.

[2] M. Weiss, E. Aschkenasy, T. Parsons, Study and development of the INTEL technique for improving speech intelligibility, Technical Report NSC-FR/4023, 1974.

[3] S. F. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process., 1979, ASSP-27(2), pp.113—120.

[4] K. Yamashita, S. Ogata, T. Shimamura, Improved spectral subtraction utilizing iterative processing, IEICE trans on Fundametals, 2005, J88-A(11), pp.1246-1257.

[5] R. J. McAulay, M. L. Malpass, Speech enhancement using a soft-decision noise suppression filter, IEEE Trans. Acoust. Speech Signal Process., 1980, ASSP-28, pp.37-145.

[6] M. Dendrinos, S. Bakamides, G. Carayannis, Speech enhancement from noise: A regenerative approach, Speech Commun., 1991, 10, pp.45-57.

[7] Y. Ephraim H. L. Van Trees, A signal subspace approach for speech enhancement, In Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing, pp.355--358. Minneapolis, USA, 27-30 April 1993.

[8] E. M. Grais, M. U. Sen, H. Erdogan, Deep neural networks for single channel source separation, In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp.3734-3738, Florence Italy, 4-9 May, 2014.

[9] Y. Xu, J. Du, L-R. Dai, C-H. Lee, An Experimental Study on Speech Enhancement Based on Deep Neural Networks, IEEE Signal Processing Letters, 2013, 21(1), pp.65-68.

[10] Q. Liu, W. Wang, P. B. Jackson, Y. Tang, A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions, In Proceedings of 25th European Signal Processing Conference, pp.1270-1274, Kos Greece, 28 August – 2 September 2017.

[11] D. P. Jarrett, Theory and applications of spherical microphone array processing, 2017, Springer.

[12] J. Benesty, et.al. Microphone array signal processing, 2010, Springer.

[13] Q. Zhao, F. Guo, X. Zu, Y. Chang, B. Li, X. Yuan, An Acoustic Signal Enhancement Method Based on Independent Vector Analysis for Moving Target Classification in the Wild. Sensors 2017, 17, 2224. https://doi.org/10.3390/s17102224

[14] K. Nordhausen, H. Oja. Independent component analysis: A statistical perspective, Wires computational statistics, 2018.

[15] S. Addisson, V. Luis, Independent component analysis (ICA): algorithm, applications and ambiguities. 2018, Hauppauge, NY: Nova Science Publishers.

[16] S. Makino. et.al. Blind speech separation, 2007, Springer.

[17] M. Taseska, E. A. P. Habets, Blind Source Separation of Moving Sources Using Sparsity-Based Source Detection and Tracking, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(3), pp. 657-670.

[18] T. Dekens, T, W. Verhelst, F. Capman, F. Beaugendre, Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection," In Proceedings of 18th European Signal Processing Conference, pp.1978-1982, Aalborg Denmark,23-27 August 2010.

[19] E. Eisemann, F. Durand, Flash photography enhancement via intrinsic relighting, ACM Transactions on Graphics, 2004, 23(3), pp.673-678,

[20] G. Petschnigg, M. Agrawala, H. Hoppe, R. Szeliski, M. Cohen, K. Toyama, Digital photography with flash and no-flash image pairs ACM Transactions on Graphics, 2004, 23(3), pp.664-672.

[21] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, X. Wang, A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference, IEEE Transactions on Multimedia, 2010, 12(7), pp.682-691.

[22] V. John, S. Tsuchizawa, S. Mita, Fusion of thermal and visible cameras for the application of pedestrian detection, Signal, Image and Video Processing, 2017, 11(3), pp.517-524.

[23] E. Fendri, R. R. Boukhriss, M. Hammami, Fusion of thermal infrared and visible spectra for robust moving object detection, Pattern Analysis and Applications, 2017, 20(4), pp.907-926.

[24] J. Kawaguchi, M. Matsumoto, Noise Reduction Combining a Normal microphone and a Throat Microphone. Sensors 2022, 22, 4473. https://doi.org/10.3390/s22124473

[25] T. Abe, S. Hashimoto, M. Matsumoto, Automatic parameter optimization in epsilon-filter for acoustical signal processing utilizing correlation coefficient. The Journal of the Acoustical Society of America, 2010, 127(2), 896-901.

[26] M. Matsumoto, S. Hashimoto, Estimation of optimal parameter in ε-filter based on signal-noise decorrelation, IEICE transactions on Information and Systems, 2009, E92-D(6), 1312-1315.

[27] S. Rickard, O. Yilmaz, On the approximate w-disjoint orthogonality of speech, In Proceedings of IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing, pp.529-532, Orland USA, 13-17 May 2002.

[28] T. Ihara, M. Handa, T. Nagai, A. Kurematsu, Multi-channel speech separation and localization by frequency assignment, IEICE trans on Fundamentals, 2003, J86-A(10), pp.998-1009.

[29] M. Aoki, Y. Yamaguchi, K. Furuya, A. Kataoka, Modifying SAFIA: Separation of the target source close to the microphones and noise sources far from the microphones, The IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2005, J88-A(4), pp.468-479.

[30] Sound effect lab: https://soundeffect-lab.info/sound/environment/

[31] M. Fukui, et.al. Noise-power estimation based on ratio of stationary noise to input signal for noise reduction, Journal of Signal Processing, 2014, 18(1), pp.17-28.