2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Speech enhancement fusing a general microphone and a bone conduction microphone

Junki Kawaguchi and Mitsuharu Matsumoto, *Member, IEEE*

Abstract— **This paper describes a method to reduce noise contained in signal obtained from a general microphone by using sensor fusion with a bone conduction microphone and a general microphone. A bone conduction microphone is a microphone that directly detects the vibrations in the throat when a person speaks. Hence, bone conduction microphones are less susceptible to external noise than general microphones. However, as bone conduction microphones have different acoustic characteristics than general microphones, the sound quality will deteriorate. Based on the above prospects, we are researching the sensor fusion technology with a bone conduction microphone and a general microphone. In this paper, we formulate the framework of the proposal approach and conduct some experiments to check the effectiveness of the proposal approach.**

I. Introduction

With the spread of mobile devices such as smartphones, demand for voice input has increased in recent years. Speech input may be used in noisy as well as quiet environments. Audio signal recorded in such noisy environments may be difficult to hear or understand. Speech enhancement is a technology that extracts desired speech from noise-mixed speech, and is an important issue in acoustic processing.

Speech enhancement can be divided into two types depending on the number of the channels for signal processing. One is speech enhancement that uses only one signal, and the other is speech enhancement that uses multiple signals. Previous survey has classified speech enhancement algorithms that utilize a single signal into three types.

The spectral subtraction (SS) algorithm is a powerful algorithm for suppressing noise from a single signal. Weiss et al. proposed an early approach as spectral subtraction in the correlation domain [2]. Boll also proposed an approach as spectral subtraction in the Fourier transform domain [3]. Spectral subtraction has shown promise and has undergone various improvements [4]. Statistical model-based algorithms are known as another approach for speech enhancement on a signal. For example, McAulay et al. tried to estimate the spectrum of the target signal in the frequency domain by using the maximum likelihood method [5]. The subspace algorithm is the other approach. The algorithm using singular value decomposition and the algorithm using an eigenvalue decomposition of signals are examples of this approach. Speech enhancement using only one signal can remove noise even from a recorded sound source, unlike speech enhancement using multiple signals. However, this approach only allows information about the sound source itself to be used for speech enhancement. Speech enhancement using deep learning has been actively researched in recent years, but issues remain, such as the need for a large amount of computing power [8–10].

Speech enhancement using a microphone array is an example of speech enhancement using multi-channel signals [11,12]. The microphone array utilizes the differences in amplitude and phase of the sound recorded by each microphone to enhance the speech.

There has been a lot of research into microphone arrays, but their implementation requires placing the microphones in different locations and tends to be large.

Famous examples of audio enhancement technology using microphone arrays include sound focus, which aligns the phase in the target direction, and adaptive array, which reduces sensitivity in the direction of noise to zero. Sound separation using sparsity [13,14] and sound separation using independent component analysis [15–17] are famous in the field of blind source separation. However, problems with utilizing a large number of microphones still remain.

Bone conduction microphones record sounds by directly sensing the vibrations of the sound source [18]. For this reason, bone conduction microphones are less susceptible to external noise. This is a great advantage of bone conduction microphones over common microphones. On the other hand, bone conduction microphones use a different recording method than regular microphones, so the sound quality of the recorded signal deteriorates.

In the field of image processing, sensor fusion is used as an approach to improve the quality of obtained images. Cross bilateral filter is a sensor fusion technology and is also called joint bilateral filter [19, 20]. This method combines non-flash and flash images taken at the same location to create a more natural-looking image. Other sensor fusion research is also being conducted to detect various objects such as faces [21], pedestrians [22], and cars using infrared and visible camera images. Based on these studies, we will introduce sensor fusion into the acoustics field. The proposed method uses the audio signal obtained with a bone conduction microphone to identify the position of the audio in frequency space, and emphasizes the audio signal obtained with a general microphone [23]. This approach allows us to efficiently remove the noise contained in regular microphones while obtaining more natural-sounding audio. Although the approach is effective to enhance the speech signal, the noise still remains in the region where signal and noise overlap on the frequency axis. To handle this problem, we propose an improved approach for speech enhancement fusing a general microphone and a bone conduction microphone. Considering

J. Kawaguchi was with the University of Electro-Communications, 1-5-1, Chofugaoka, Chofu-shi, Tokyo, 182-8585, Japan

M. Matsumoto is with the University of Electro-Communications, 1-5-1, Chofugaoka, Chofu-shi, Tokyo, 182-8585, Japan (corresponding author e-mail: mitsuharu.matsumoto@ieee.org).

the simplicity of noise removal processing, this study investigated how much performance could be improved by using spectral subtraction in combination. In the next section, the algorithm of the binary mask is shown to ease the understanding of the proposed approach. The proposed approach in the past study and its problems are shown in Section 3. We also show three approaches using spectral subtraction to improve the past algorithm. We conducted some experiments to show the effectiveness of the proposed approaches in Section 4. Several comparisons are also conducted to check the effectiveness of the approaches. Discussion and conclusion are given in Section 5.

II. Approach Using Binary Mask with Two Microphones

A. Formulate the problem

To clarify the difference between the typical binary mask and our approach, this section formulates the speech enhancement by utilizing the binary mask for two microphones [24–26]. When we use binary masks, two microphones are usually assumed. Let us consider $x_1(t)$ as the mixed signal recorded by microphone 1. Removing the effects of delay and the attenuation for $x_1(t)$ does not result in any loss of generality of the problem.

From the above aspects, we can express $x_1(t)$ as follows:

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

where $s(t)$ and $n_i(t)$ represent the target signal and the ith noise (i = 1, 2, 3, …, N), respectively. Let us consider $x_2(t)$ as the mixed signal recorded by microphone 2. Unlike $x_1(t)$, we need to consider the attenuation and delay of the signal obtained by the microphone 2 relative to the signal obtained by microphone 1. Hence, we can express $x_2(t)$ as follows:

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

where δ represents the delay regarding the target signal. δi represents the delay regarding the ith noise. a is the attenuation between the first microphone and the second microphone regarding the target signal. a_j represents parameters the attenuation between the microphone 1 and the microphone 2 regarding the ith noise.

The following constraints are satisfied when Δ is defined as the maximal possible delay between the microphone 1 and the microphone 2.

$$|\delta| \le \Delta, \tag{3}$$

$$|\delta_i| \le \Delta, \tag{4}$$

When we set the assumption regarding the sparsity, we can consider that the tar-get signal and noise are disjoint in the time-frequency domain. Let us consider $S(\tau, \omega)$ as the short term Fourier transform of $s(t)$. Let us also consider $N_i(\tau, \omega)$ as the short-term Fourier transform of $n_j(t)$. We can express $S(\tau, \omega)$ as follows:

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

where τ represents the time frame. ω represents the angular frequency. $W(\tau)$ represents the window function. When we can consider that the target signal and all the noise are sparse, the following condition can be satisfied:

$$S(\tau,\omega)N_i(\tau,\omega) = 0\ \forall\tau,\omega, \tag{6}$$

B. Speech Enhancement Utilizing Binary Mask

We first estimate the attenuation and the delay in Equation (2) to achieve the speech enhancement. Let us define $X_i(\tau, \omega)$ as a short term Fourier transform of $x_i(t)$. We can express $X_i(\tau, \omega)$ as follows.

$$X_i(\tau,\omega) = \sum_{t=-\infty}^{\infty} x_i(t+\tau)W(t)\exp(-i\omega t)\,, \tag{7}$$

We can express $X_1(\tau, \omega)$ and $X_2(\tau, \omega)$ using matrices as follows.

$$\begin{bmatrix} X_1(\tau,\omega) \\ X_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ a\exp(-i\omega\delta) & a_1\exp(-i\omega\delta_1) & \dots & a_N\exp(-i\omega\delta_N) \end{bmatrix} \begin{bmatrix} S(\tau,\omega) \\ N_1(\tau,\omega) \\ \dots \\ N_N(\tau,\omega) \end{bmatrix} \tag{8}$$

When we can assume the sparsity of target signal and noise in the time-frequency domain, we can obtain the following equation:

$$\begin{bmatrix} X_1(\tau,\omega) \\ X_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ a(\tau_s,\omega_s)\exp(-i\omega_s\delta(\tau_s,\omega_s)) \end{bmatrix} s(\tau_s,\omega_s) \tag{9}$$

Equation (9) means that the target signal exists at (τ_s, ω_s). The ratio of $X_1(\tau_s, \omega_s)$ and $X_2(\tau_s, \omega_s)$ are calculated to estimate (τ_s, ω_s). Let a(τ_s, ω_s) be the relative amplitude. Let δ(τ_s, ω_s) be the relative delay. $a(\tau_s, \omega_s)$ and δ(τ_s, ω_s) can be estimated as

$$\left(a(\tau_s,\omega_s), \delta(\tau_s,\omega_s)\right) = \left(\left|\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}\right|, \frac{1}{\omega}\angle\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}\right) \tag{10}$$

where

$$\angle a\exp(i\phi) = \phi,\ -\pi < \phi < \pi, \tag{11}$$

Let us define $M(\tau, \omega)$ for (τ, ω) as the time-frequency mask. $M(\tau, \omega)$ can be expressed as follows:

$$M(\tau,\omega) = \begin{cases} 1, |\ln a(\tau,\omega) - \ln a| < \frac{\Delta_a}{2} \wedge |\ln\delta(\tau,\omega) - \ln\delta| < \frac{\Delta_\delta}{2} \\ 0. otherwise \end{cases} \tag{12}$$

where Δ_a represents the amplitude resolution width of the histograms. Δ_δ represents delay resolution width of the histograms. We execute the masking process as follows to remove noise included in the mixed signal recorded by the microphone 1.

$$S(\tau,\omega) = M(\tau,\omega)X_1(\tau,\omega), \tag{13}$$

It is known that the binary mask has high ability for speech enhancement. However, the accuracy of speech enhancement depends on how accurately masking information is estimated. Estimating mask information becomes difficult in environments with environmental noise that does not fully satisfy sparsity.

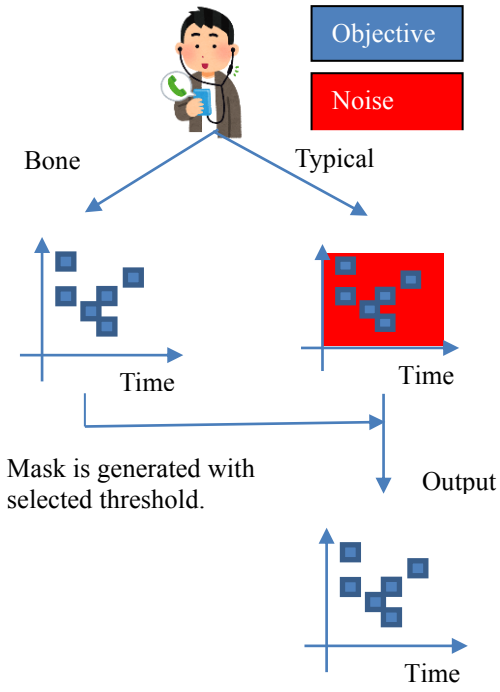

Figure 1. Basic concept of the proposed approach.

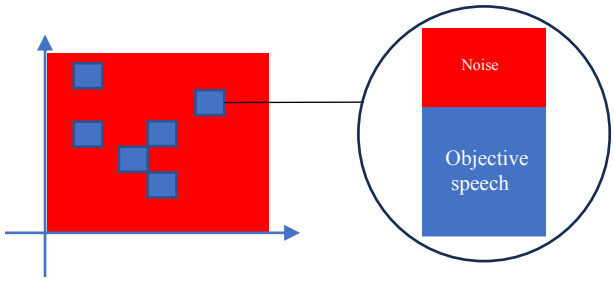

Figure 2. Basic concept of the proposed approach.

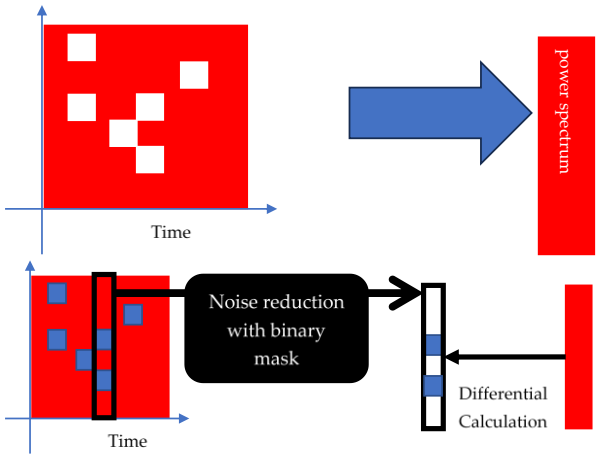

Figure 3. Overview of proposed method 1

III. Method

A. Basic concept of the Proposed 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 con-duction 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.

B. Problem Formulation

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.

$$X_1(\tau,\omega) = S(\tau,\omega) + \sum_{i=1}^{N} N_i(\tau,\omega), \quad (14)$$

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.

$$X_2(\tau,\omega) = B(\omega)S(\tau,\omega) \quad (15)$$

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.

$$M_j(\tau,\omega) = \begin{cases} 1 & |X_2(\tau,\omega)| \geq th_j \\ 0 & |X_2(\tau,\omega)| < th_j \end{cases} \quad (16)$$

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:

$$Y_j(\tau,\omega) = M_j(\tau,\omega)X_1(\tau,\omega) \quad (17)$$

C. Problems with previous methods

In this method, for sounds where the power spectrum of the target signal and the power spectrum of the noise signal

overlap on the frequency axis, as shown in Figure 2, noise remains in the overlapped area. Therefore, in the following proposed method, we will consider an improvement method using speech enhancement based on a combination of the binary mask and spectral subtraction methods.

D. *Spectral subtraction (proposed method 1)*

In this study, we combined the spectral subtraction method to solve the above problems. The spectral subtraction method performs noise reduction by subtracting an estimate of the average power spectrum of the retained noise from the power spectrum of the target speech mixed with noise. Hereafter, to ease the explanation, we call it the mixed speech. Figure 3 shows an overview of the proposed method 1.

With regard to equation (14), the power on the frequency axis is calculated. When we assume that there is no correlation between signal and noise in spectral subtraction, we approximate the power of the mixed signal by using the power of the signal and the power of the noise as follows.

$$
\begin{aligned}
|X_1(\tau,\omega)|^2 &= \left|S(\tau,\omega)+\sum_{i=1}^{n}N_i(\tau,\omega)\right|^2 \\
&= |S(\tau,\omega)|^2 + S(\tau,\omega)\sum_{i=1}^{n}N_i^*(\tau,\omega) + S^*(\tau,\omega)\sum_{i=1}^{n}N_i(\tau,\omega) + \left|\sum_{i=1}^{n}N_i(\tau,\omega)\right|^2 \\
&\approx |S(\tau,\omega)|^2 + \left|\sum_{i=1}^{n}N_i(\tau,\omega)\right|^2
\end{aligned}
\tag{18}
$$

where * denotes the complex conjugation.

Spectral subtraction enhances the target signal by subtracting the estimated noise from the mixture under these assumptions about the noise and signal. In the typical spectral subtraction, the noise is assumed to be stationary. Let $\widehat{SP}(\tau,\omega)$ be the estimated speech spectrum after noise subtraction. $\widehat{SP}(\tau,\omega)$ can be expressed as follows:

$$
\left|\widehat{SP}(\tau,\omega)\right|^2 = |X_1(\tau,\omega)|^2 - \left|\widehat{N}(\tau,\omega)\right|^2 \tag{19}
$$

where $\widehat{N}(\tau,\omega)$ represents the average power spectrum of the estimated noise. In this case, the power of the noise is averaged in areas where it is assumed to be noise.

E. *Spectral subtraction using ratio (proposed method 2)*

The proposed method 1 uses the average power spectrum of the noise as the power spectrum of the estimated noise. Therefore, it is difficult for proposed method 1 to deal with large temporal changes in noise power. In daily life, ambient noise may change in loudness. Therefore, we propose a spectral subtraction method that can robustly cope with changes in ambient noise. Here, we assume that as the loudness of the sound changes, its shape does not change on the frequency axis.

Under the above assumptions, the proposed method 2 estimates the noise as follows.

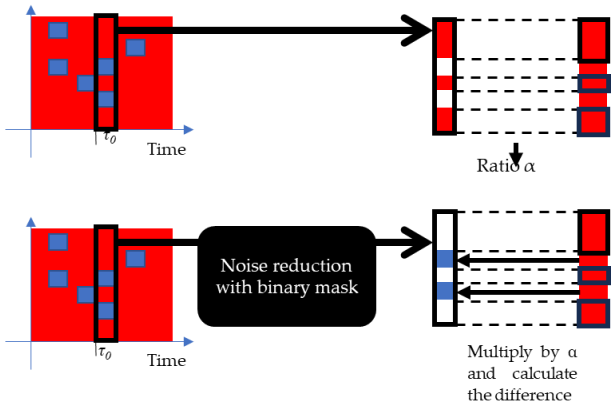

Figure 4. Overview of proposed method 2

The noise area is first estimated by using the signal with the bone conduction microphone. Then, the ratio of the frame of interest to the estimated noise is obtained by comparing the power of the estimated noise to the noise portion in the frame of interest, and the spectral subtraction method is performed in a form that reflects this ratio. Figure 4 shows an overview of the proposed method 2. The specific procedure is as follows.

First, the noise frame is estimated by using the signal from the bone conduction microphone. Then, the power of the noise is averaged by using the mixed signal. We then estimate the noise part of the mixed speech for the processing frame and calculate the average power spectrum of the noise part. After that, we calculate the ratio of the average overall noise power and the power of the noise part of the frame of interest. The ratio is reflected in the subtraction of the spectral subtraction method to cope with changes in the loudness of the sound. Hence, finally, we subtract the noise from the mixed signal at the target frame, considering the calculated ratio.

Let us formulate the above process. Let us define α as the ratio of the loudness of the frame of interest to the other frames. α is estimated from the noise portion using the mask information, and the noise signal is subtracted from the mixed signal using α. α can be estimated as follows.

$$
\alpha = \frac{\sum_{M(\tau,\omega)=0}|X_1(\tau_0,\omega)|}{\sum_{M(\tau,\omega)=0}|X_1(\tau,\omega)|} \tag{20}
$$

where τ_0 indicates the time of the frame of interest and $\sum_{M(\tau,\omega)=0}|X_1(\tau_0,\omega)|$ indicates the sum in the frequency domain where the mask is zero. Spectral subtraction using α can be expressed as follows.

$$
\left|\widehat{SP}(\tau,\omega)\right|^2 = |X_1(\tau,\omega)|^2 - \alpha^2\left|\widehat{N}(\omega)\right|^2 \tag{21}
$$

F. *Spectral subtraction for noise estimation from surroundings (proposed method 3)*

The proposed method 2 assumes that the frequency distribution of noise is the same and its power changes. Since the proposed method 2 computes the average power spectrum of noise, it can cope with changes in the power of noise with the same waveform, but it cannot cope with changes in the type of noise itself.

Moreover, the received speech and mask information are used to compute the average of all frames consisting only of noise, which requires waiting until the end of speech reception, making it unsuitable for real-time processing in a real environment. Here, we propose a method to improve these problems. Assume that the noise in a small interval on the time axis changes little to accommodate the type of noise. Under this assumption, the problem is solved by varying the range over which the average power spectrum of the noise is calculated. Figure 5 shows an overview of the proposed method 3. The average power spectrum of the noise is calculated from the time when the speech mixture was acquired to β frames before. This eliminates the need to wait until all the noise is read in, as in the proposed method 2. Also, since noise estimation is performed from the previous frame, it can cope with changes in noise types if the hypothesis is valid. If the estimated noise is $\widehat{N}(\tau,\omega)$ because the noise changes, the spectral subtraction can be expressed as follows.

$$\left|\widehat{SP}(\tau,\omega)\right|^2 = |X_1(\tau,\omega)|^2 - \alpha^2\left|\widehat{N}(\omega)\right|^2 \quad (22)$$

For noise estimation, the same process as in the proposed method 2 is performed for the β frames prior to the frame of interest, as follows.

$$\alpha = \frac{\sum_{M(\tau,\omega)=0,\tau_0-\beta<\tau<\tau_0}|X_1(\tau_0,\omega)|}{\sum_{M(\tau,\omega)=0,\tau_0-\beta<\tau<\tau_0}|X_1(\tau,\omega)|} \quad (23)$$

where τ_0 indicates the time of the frame of interest and $\sum_{M(\tau,\omega)=0}$ is the sum over the points with zero mask information.

IV. Evaluation by 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 is verified.

We used 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 implement all the programs. Table 1 shows the experimental conditions used in the experiments. We prepared white noise to check the basic performance of the proposed approaches. We also prepared intersection noise, restaurant noise, and station noise as natural noise. Three noises were selected from the Sound Effects Lab [28] and Hashimoto Tech [29]. We prepared the male voice for the target signal. Sound level is expressed in dBFS. We changed the threshold value in 10 dB intervals to check the effect of the parameter. The sampling frequency was set to 48000 Hz, and the frame size was set to 2048. β was set to 40 in the proposed method 3.

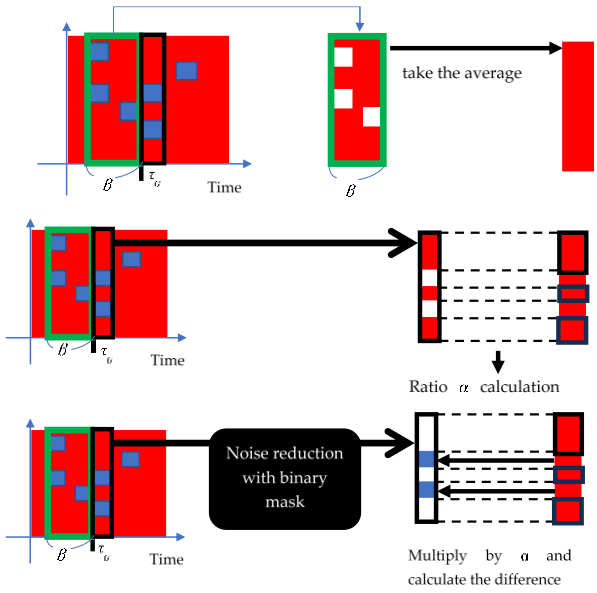

Figure 5. Overview of proposed method 3

TABLE I. Experimental Condition

Speech signal	Speech by a man
Noise signal	White noise Intersection noise, Restaurant noise, Station noise
Threshold for noise reduction	From -80[dB] to -30[dB] with 10[dB] intervals

The window function was set to a Hamming window. We used Signal to Distortion Ratio (SDR) [30] 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.

$$SDR = 10\log_{10}\left(\frac{\sum_{\tau,\omega}|S(\tau,\omega)|}{\sum_{\tau,\omega}\left||S(\tau,\omega)| - \lambda|\hat{S}(\tau,\omega)|\right|}\right) \quad (24)$$

where $\hat{S}(\tau, \omega)$ 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 λ as follows:

$$\lambda = \sqrt{\frac{\sum_{\tau,\omega}|S(\tau,\omega)|}{\sum_{\tau,\omega}|\hat{S}(\tau,\omega)|}} \quad (25)$$

B. Experimental results

Table 2, 3, 4 and 5 show the SDRs of the target signal and the speech after noise reduction using the previous method and the improved methods when we used white noise, intersection noise, restaurant noise, and station noise, respectively.

TABLE II. SDRS WHEN WHITE NOISE WAS USED.

Threshold(dB)	BM	PM1	PM2	PM3
-80	3.199	4.005	4.371	4.356
-70	5.007	5.853	6.203	6.203
-60	6.838	7.409	7.596	**7.602**
-50	6.646	6.816	6.900	6.902
-40	4.236	4.264	4.277	4.259
-30	1.522	1.519	1.517	1.515

TABLE III. SDRS WHEN INTERSECTION NOISE WAS USED.

Threshold(dB)	BM	PM1	PM2	PM3
-80	2.632	3.385	3.665	3.484
-70	8.384	9.015	9.349	9.356
-60	9.593	9.703	**9.838**	9.818
-50	7.393	7.414	7.436	7.441
-40	4.351	4.351	4.351	4.357
-30	1.527	1.526	1.527	1.520

TABLE IV. SDRS WHEN RESTAURANT NOISE WAS USED.

Threshold(dB)	BM	PM1	PM2	PM3
-80	4.948	5.312	5.470	5.475
-70	5.762	6.048	6.244	6.233
-60	6.081	6.257	**6.412**	6.381
-50	5.531	5.629	5.663	5.674
-40	3.653	3.672	3.677	3.705
-30	1.352	1.350	1.338	1.352

TABLE V. SDRS WHEN STATION NOISE WAS USED.

Threshold(dB)	BM	PM1	PM2	PM3
-80	9.144	9.672	9.912	9.913
-70	10.02	10.39	10.55	**10.58**
-60	9.335	9.460	9.500	9.500
-50	7.290	7.322	7.308	7.317
-40	4.386	4.389	4.375	4.380
-30	1.527	1.527	1.517	1.521

The experimental results are expressed as four significant digits. The value with the largest SDR among all the experimental values is bolded. We also show the spectrogram and waveform when we used white noise in the figures. The experimental results show that for all types of noise, the combination of binary mask and spectral subtraction improves speech enhancement performance compared to binary mask alone. As shown in Table 2, for simple white noise, the proposed method 2 has better performance than the method using simple spectral subtraction (the proposed method 1) as we expected. The proposed method 3 performs even better

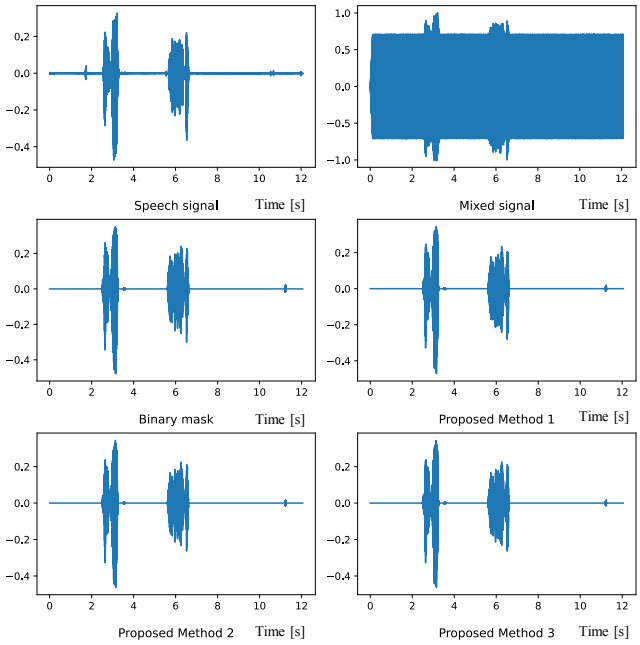

Figure 6. Waveform (White noise)

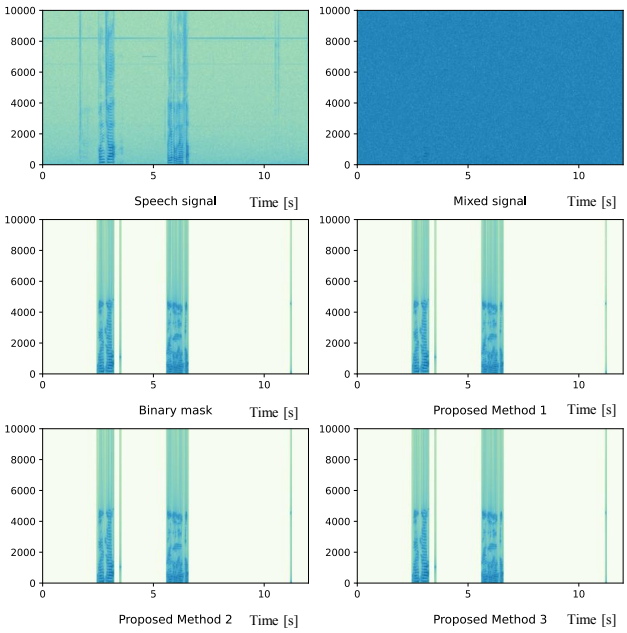

Figure 7. Spectrogram (White noise)

than method 2. The proposed methods 2 and 3 are also better to the proposed method 1 for general noise. Based on these results, the proposed method is considered to have a certain degree of effectiveness over the methods of previous research. The waveforms and spectrograms of these results are shown below. We show two cases (white noise and intersection noise) because the results using other noises have similar characteristics. Figure 6 and 7 show the waveforms and the spectrograms of the target signal, the mixed signal and the outputs of the previous method and the proposed methods when we used white noise as noise signal.

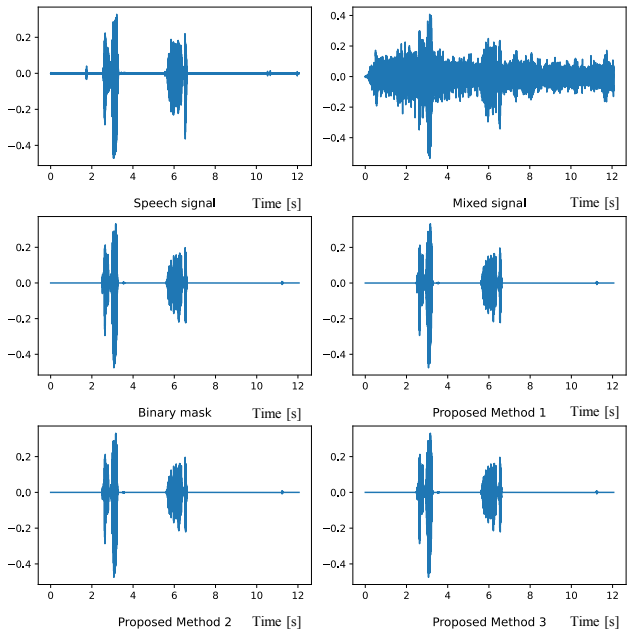

Figure 8. Waveform (Intersection noise)

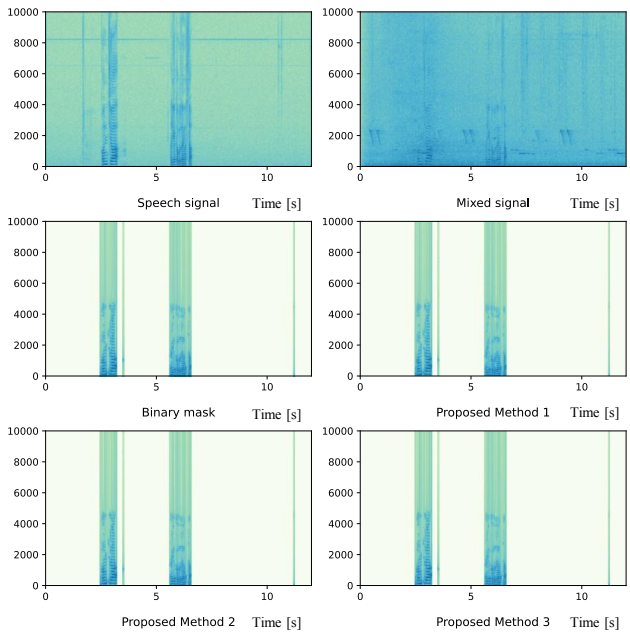

Figure 9. Spectrogram (Intersection noise)

Figure 8 and 9 show the waveforms and the spectrograms of the target signal, the mixed signal and the outputs of the previous method and the proposed methods when we used intersection noise as noise signal. When we observe the resulting waveform, we can see that the main noise is removed when the binary mask is applied in both cases. There is not much change in appearance due to the noise removal with the binary mask.

V. Discussion and Conclusion

Regarding methods to improve noise removal accuracy, we proposed three methods combining the sensor fusion using binary mask proposed in the previous study and spectral subtraction.

We utilized white noise and three types noise; intersection noise, restaurant noise and station noise. To check the sound quality due to the proposed approach, we used SDR to check the performance of the speech enhancement of the proposed approaches. We confirmed that all the approaches could improve the sound quality compared to the previous study. In the previous study, the binary mask is acquired by using the information from the bone conduction microphone, and it is applied to the sound obtained from the general microphone. Due to this process, it was difficult to remove noise in areas where voice and noise overlapped on the frequency axis. Our approach basically aims to solve the problem of the previous method. Through the experiments, we confirmed that all the approaches could improve the sound quality compared to the previous study.

We prepared the speaker voice with a general microphone and the bone conduction microphone simultaneously in the experiments. Although there may be a slight time difference between the two recorded sounds, good results have been obtained in experiments even without taking this time difference into consideration.

Through the experiments, all the results of the proposed methods were superior to the results of the previous method. These results show the effectiveness of the spectral subtraction to the previous approach.

The proposed method still has some problems to be solved. As an example, the frequency range that can be recorded with a typical microphone is wider than the frequency range that can be recorded with a bone conduction microphone. Therefore, in-formation in the high frequency range that cannot be obtained with bone conduction microphones will need to be obtained using other methods.

Sufficient voice enhancement is possible even within the range of frequency information that bone conduction microphones can acquire. However, in order to improve sound quality, it is necessary to consider ways to obtain higher frequency information.

In addition, in order to put the proposed method into practical use, it is necessary to implement it in real time. As it is thought that the algorithm itself can be implemented with some time delay, we would like to consider implementing real-time processing in the future.

References

[1] P. C. Loizou, “Speech Enhancement: Theory and Practice, 2nd Ed.,” CRC Press: London, UK, 2007.

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

[3] S.F.Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process 1979, 27, 113–120.

[4] K. Yamashita, S. Ogata, T. Shimamura, “Improved spectral subtraction utilizing iterative processing,” IEICE Trans Fundamentals 2005, J88-A, 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, 28, 37–145.
[6] M. Dendrinos, S. Bakamides, G. Carayannis, Speech enhancement from noise: A regenerative approach. Speech Commun. 1991, 10, 45–57.
[7] Y. Ephraim, H. L. Van Trees, “A signal subspace approach for speech enhancement,” In Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing, Minneapolis, MN, USA, 27–30 April 1993; pp. 355–358.
[8] E. M. Grais, M. U. Sen, H. Erdogan, “Deep neural networks for single channel source separation,” In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 4–9 May 2014; pp. 3734–3738.
[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 Lett. 2013, 21, 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 the 25th European Signal Processing Conference, Kos, Greece, 28 August–2 September 2017; pp. 1270–1274.
[11] Jarrett, D.P. Theory and Applications of Spherical Microphone Array Processing; Springer: New York, USA, 2017.
[12] J. Benesty, J. Chen, Y. Huang, “Microphone Array Signal Processing,” Springer: New York, USA, 2010.
[13] S. Makino, T.W. Lee, H. Sawada, (Eds.) Blind Speech Separation; Springer: New York, USA, 2007.
[14] M. Taseska, E.A.P. Habets, “Blind Source Separation of Moving Sources Using Sparsity-Based Source Detection and Tracking,” IEEE/ACM Trans. Audio Speech Lang. Processing 2018, 26, 657–670.
[15] 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.
[16] K. Nordhausen, H.Oja. “Independent component analysis: A statistical perspective,” Wires Comput. Stat. 2018, 10, e1440.
[17] S. Addisson, V. Luis, “Independent Component Analysis (ICA): Algorithm, Applications and Ambiguities,” Nova Science Publishers: Hauppauge, NY, USA, 2018.
[18] T. Dekens, W. Verhelst, F. Capman, F. Beaugendre, “Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection,” In Proceedings of the 18th European Signal Processing Conference, Aalborg, Denmark, 23–27 August 2010; pp. 1978–1982.
[19] E. Eisemann, F.Durand, “Flash photography enhancement via intrinsic relighting,” ACM Trans. Graph. 2004, 23, 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 Trans. Graph. 2004, 23, 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 Trans. Multimed. 2010, 12, 682–691.
[22] V. John, S. Tsuchizawa, S. Mita, “Fusion of thermal and visible cameras for the application of pedestrian detection,” Signal Image Video Processing 2017, 11, 517–524.
[23] E. Fendri, R.R. Boukhriss, M. Hammami, “Fusion of thermal infrared and visible spectra for robust moving object detection,” Pattern Anal. Appl. 2017, 20, 907–926.
[24] J. Kawaguchi, M. Matsumoto, “Noise Reduction Combining a General Microphone and a Throat Microphone,” Sensors 2022, 22, 4473. https://doi.org/10.3390/s22124473
[25] S. Rickard, O. Yilmaz, “On the approximate w-disjoint orthogonality of speech,” In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orland, CA, USA, 13–17 May 2002; pp. 529–532.
[26] T. Ihara, M. Handa, T. Nagai, A.Kurematsu, “Multi-channel speech separation and localization by frequency assignment,” IEICE Trans Fundam. 2003, J86-A, 998–1009.
[27] 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,” IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 2005, J88-A, 468–479.
[28] Sound Effect Lab. Available online: https://soundeffect-lab.info/sound/environment/ (accessed on 15/July/2025).
[29] Hashimoto Tech. Available online: https://hashimoto-tech.jp/local/advan/signwav (accessed on 15/July/2025).
[30] M. Fukui, S. Shimauchi, Y. Hioka, A. Nakagawa, Y. Haneda, H. Ohmuro, A. Kataoka, “Noise-power estimation based on ratio of stationary noise to input signal for noise reduction,” J. Signal Processing 2014, 18, 17–28.