Communication

Noise Reduction Combining a General Microphone and a Throat Microphone

Junki Kawaguchi and Mitsuharu Matsumoto *

Graduate School of Informatics and Engineering, University of Electro-Communications, 1-5-1, Chofugaoka, Chofu-shi 182-8585, Japan; k2230046@edu.cc.uec.ac.jp
* Correspondence: mitsuharu.matsumoto@ieee.org

Abstract: In this study, we propose a method to reduce noise from speech obtained from a general microphone using the information of a throat microphone. A throat microphone records a sound by detecting the vibration of the skin surface near the throat directly. Therefore, throat microphones are less prone to noise than ordinary microphones. However, as the acoustic characteristics of the throat microphone differ from those of ordinary microphones, its sound quality degrades. To solve this problem, this study aims to improve the speech quality while suppressing the noise of a general microphone by using the information recorded by a throat microphone as reference information to extract the speech signal in general microphones. In this paper, the framework of the proposed method is formulated, and several experiments are conducted to evaluate the noise suppression and speech quality improvement effects of the proposed method.

Keywords: noise reduction; throat microphone; sensor fusion

1. Introduction

In recent years, with the development of information and communication devices, there has been an increase in the use of devices that use voice input, such as mobile devices, e.g., smartphones, for making calls. Although the use of voice input devices is comfortable in a noise-free environment, they may be used in a noisy environment. In such a case, the voice may be difficult to hear or unbearable to use. Therefore, speech enhancement is one of the important issues in acoustic processing to emphasize the target speech from the mixed signal.

The approach for speech enhancement can be categorized into two types. One is the approach using a single-channel signal. The other is the approach using multi-channel signals. The survey on speech enhancement classifies algorithms for speech enhancement of single-channel signals into three categories [1].

The spectral subtraction (SS) algorithm is one of those approaches. Early approaches include spectral subtraction in the correlation domain proposed by Weiss et al. [2] and that in the Fourier transform domain proposed by Boll [3]. Spectral subtraction is a powerful approach to single-channel speech enhancement and has undergone various improvements [4]. Another approach is a statistical model-based algorithm. For example, McAulay et al. used the maximum likelihood method to estimate the Fourier transform spectrum of the target signal [5]. The other approach is a subspace algorithm. Dendrinos et al. proposed an algorithm that uses singular value decomposition [6]. Ephraim et al. proposed an algorithm that uses an eigenvalue decomposition of signals [7]. The single-channel signal approach is attractive in that it uses only a single recording source. However, information that can be used for speech enhancement is limited to that related to the sound source itself. In recent years, deep learning-based approaches have also been actively studied but require a lot of computational power [8–10].

The microphone array is a typical method using multi-channel signals [11,12]. A microphone array uses multiple microphones to achieve noise reduction using the phase
and amplitude differences of the sound entering each microphone. The microphone array is implemented in various ways. However, in principle, its implementation requires special equipment, such as a large number of microphones and complex processing. There have been various approaches to the study of noise reduction techniques. In general, noise reduction techniques using microphones have been widely studied. Well-known examples are sound focusing by phase homomorphism and adaptive microphone array by generating the zero point in the noise direction. In the field of blind source separation, independent component analysis [13–15] and sparsity-based source separation [16,17] are well known, but they require special equipment, such as a large number of microphones, for implementation.

One way to use voice input devices comfortably without being affected by noise is to use a microphone that is less susceptible to noise. A throat microphone is an example and records sound by detecting the vibration of the skin near the throat directly. Compared to general microphones that use air vibration to record, the throat microphone is airless and, therefore, more resistant to noise. Some studies paid attention to the characteristics and aimed to improve the accuracy of speech recognition by using a throat microphone that records by directly detecting the vibration of the skin near the throat [18]. Although the feature of throat microphones is useful, as described above, their acoustic characteristics differ from those of ordinary microphones, resulting in degraded speech quality.

In image processing fields, there are some studies on sensor fusion to obtain better image quality. Cross bilateral filter (Joint bilateral filter) is a filter technique that combines non-flash images with flash images taken at the same location [19,20]. The filter inputs are flash and non-flash images, resulting in a more natural quality image as the output. Other research cases have been reported that combine infrared and visible camera images to detect various objects, such as faces [21], pedestrians [22] and cars [23].

Inspired by these studies, we apply these concepts to the acoustical field and aim to obtain more natural speech by eliminating noise in speech signals obtained by a general microphone by using speech signals obtained by a throat microphone as a reference.

In this study, we propose a noise reduction method for general microphones based on speech signals from throat microphones. We also evaluate the effectiveness of the proposed method by conducting experiments. The rest of the paper is summarized as follows: In Section 2, we describe the typical binary mask using two microphones to clarify the proposed method. In Section 3, we describe the proposed approach and the usage scenario. In Section 4, some experiments are conducted using the proposed method to evaluate the noise reduction performance of the proposed method and the sound quality compared to a throat microphone. In Section 5, we present a discussion and conclusions on the results of the experiment and future perspectives.

2. Binary Mask Using Two Microphones
2.1. Problem Formulation

To clarify our approach, we first show the typical method using the binary mask with two general microphones [24–26]. A typical approach using binary masks assumes two microphones. For the mixed signal \( x_1(t) \) obtained from microphone 1, removing the effects of attenuation and delay does not lose the generality of the problem. Hence, \( x_1(t) \) can be written as follows:

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

where \( s(t) \) is the objective signal, and \( n_i(t) \) is the \( i \)th noise (\( i = 1, 2, 3, \ldots, N \)). On the other hand, it is necessary to consider the delay and attenuation of microphone 2’s signal with respect to the signal acquired by microphone 1. Therefore, the second mixed signal \( x_2(t) \) can be written as:

\[
x_2(t) = as(t - \delta) + \sum_{i=1}^{N} a_i n_i(t - \delta),
\]
where \( \delta \) and \( \delta_i \) are the delays from the objective signal and the delay from the noise \( j \), respectively. \( a \) and \( a_j \) are parameters that indicate the relative attenuation between the sensors for the objective signal and noise, respectively.

When we define \( \Delta \) as the maximum possible delay between sensors, the following constraints are obtained:

\[
|\delta| \leq \Delta, \tag{3}
\]

\[
|\delta_i| \leq \Delta, \tag{4}
\]

The sparse approach assumes that the target signal and noise are sparse in the time-frequency domain. Let \( S(\tau, \omega) \) and \( N_i(\tau, \omega) \) be the short-time Fourier transform of \( s(t) \) and \( n_j(t) \), respectively. \( S(\tau, \omega) \) can be described as follows:

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

where \( \tau \) and \( \omega \) are the time frame and the angular frequency, respectively. \( W(\tau) \) is the window function. If the sparsity between the objective signal and all the noise is satisfied, we can obtain the following condition:

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

### 2.2. Noise Reduction Using Binary Mask

In order to perform denoising, the delay and attenuation parameters in Equation (2) are estimated. If \( X_i(\tau, \omega) \) is a short-time Fourier transform of \( x_i(t) \), it can be written as follows.

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

The signals obtained by microphones 1 and 2 can be rewritten in the time-frequency domain as follows.

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

If the target signal and noise are sparse in the time-frequency domain, the following equation is obtained:

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

where the objective signal exists at \( (\tau_s, \omega_s) \). To estimate \( (\tau_s, \omega_s) \), we calculate the ratio of \( X_1(\tau_s, \omega_s) \) and \( X_2(\tau_s, \omega_s) \). Let us define \( a(\tau_s, \omega_s) \) and \( \delta(\tau_s, \omega_s) \) as the relative amplitude and delay parameters. We can estimate \( a(\tau_s, \omega_s) \) and \( \delta(\tau_s, \omega_s) \) as follows:

\[
(a(\tau_s, \omega_s), \delta(\tau_s, \omega_s)) = \left( \frac{X_2(\tau, \omega)}{X_1(\tau, \omega)} \frac{1}{\omega}, -\frac{X_2(\tau, \omega)}{X_1(\tau, \omega)} \right) \tag{10}
\]

where

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

We can design the time-frequency mask \( M(\tau, \omega) \) for \( (\tau, \omega) \) as follows:

\[
M(\tau, \omega) = \begin{cases} 
1, & |\ln a(\tau, \omega) - \ln a| < \frac{\Delta}{2} \land |\ln \delta(\tau, \omega) - \ln \delta| < \frac{\Delta}{2} \\
0, & \text{otherwise}
\end{cases} \tag{12}
\]
where $\Delta_a$ and $\Delta_\delta$ are defined as the amplitude and delay resolution width of the histograms, respectively. In order to remove noise from the signal obtained by microphone 1, the masking process is executed as follows.

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

Although the noise reduction performance of the binary mask is high, its quality depends on the accuracy of the estimated masking information. In the presence of environmental noise, the noise does not fully meet the sparsity assumptions, making it difficult to estimate mask information.

3. Proposed Approach

3.1. Overview of the Proposed Method

Figure 1 shows an overview of the proposed method. In our supposed scenario, we record the sound with a general microphone and a throat microphone simultaneously. We then create a binary mask based on the speech data obtained from the throat microphone to remove noise from the speech data obtained from the general microphone.

![Figure 1. Overview of the proposed method.](image-url)
To achieve this goal, speech data is first captured by both a throat microphone and a general microphone. We then transformed the sound into frequency domain data by Fourier transform. The next step is to set a threshold value for the power spectrum of the speech data from the throat microphone and create a binary mask with 1 for the power above the threshold value and 0 for the rest. The audio data from the general microphone is also subjected to Fourier transform and converted to frequency domain data. The binary mask is applied to the obtained frequency data from the general microphone to remove noise. Finally, the noise-removed signal is inverse Fourier transformed to obtain the output signal.

3.2. Formulation by Equation

We formulate the proposed approach in this section. Let \( x_1(t) \) be the mixed signal obtained from a general microphone and \( x_2(t) \) be the signal obtained from a throat microphone. \( T \) is time. The spectra of \( x_1(t) \) and \( x_2(t) \) are denoted as \( X_1(\tau, \omega) \) and \( X_2(\tau, \omega) \), respectively, when the signals are mapped into frequency space by Fourier transform. \( \tau \) is the time frame, and \( \omega \) is the angular frequency. When the spectrum of the objective signal is \( S(\tau, \omega) \) and the spectrum of the \( i \)th noise is \( N_i(\tau, \omega) \), \( X_1(\tau, \omega) \) is expressed as follows.

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

On the other hand, the signal of the throat microphone does not contain noise signals, while the frequency response of the sound recorded by the throat microphone is different from that of the microphone. \( X_2(\tau, \omega) \) is expressed as follows.

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

where \( B(\omega) \) is the frequency response of a throat microphone to a general microphone. For noise reduction, a threshold is set, and a binary mask \( M(\tau, \omega) \) is generated as follows. \( th \) is a threshold.

\[
M(\tau, \omega) = \begin{cases} 
1 & \text{if } |X_2(\tau, \omega)| \geq th \\
0 & \text{if } |X_2(\tau, \omega)| < th
\end{cases} \quad (16)
\]

Based on the mask information, the following process is applied to the spectral signal \( X_1(\tau, \omega) \) of a general microphone to remove noise and obtain the output signal spectrum \( Y(\tau, \omega) \).

\[
Y(\tau, \omega) = M(\tau, \omega)X_1(\tau, \omega) \quad (17)
\]

Finally, the obtained signal \( Y(\tau, \omega) \) is inverse Fourier transformed to obtain the output signal \( y(t) \). Binary mask estimation using the microphone ratio is difficult if sparse assumptions are not met.

On the other hand, our method obtains mask information directly from the throat microphone. Therefore, the proposed method can realize efficient noise removal even for a voice that does not satisfy the assumption of sparsity.

We also note that the delay between the throat microphone and the general microphone is not taken into account. The main reason is due to the difference between the typical binary mask usage scenario and the proposed method usage scenario. In the typical binary mask usage scenario, the speaker is apart from the listener, and the listener records the sound from the speaker. The proposed method aims to record the speaker’s voice clearly in a situation where the surroundings are extremely noisy. Since this work is performed on the speaker side, it is assumed that recording with a microphone and a throat microphone will be performed at the same time.

4. Experiment

4.1. Experimental Method

The effectiveness of the method proposed in Section 2 is evaluated by experiments. In this study, the accuracy of the noise removal is verified by creating the target speech, noise speech and mixed speech in advance. For the noise, we used sound effects from...
the Sound Effects Lab [27] and Hashimoto Research Institute [28]. Recorded speech was used as the voice. Recording with the general microphone and throat microphone was performed at the same time in a normal quiet room. The used general microphone was M4U made by inMusic, Inc. (Cumberland, RI, USA). The used throat microphone was DN-915129 made by ThirdWave Co., Ltd. (Tokyo, Japan). To evaluate the noise reduction performance quantitatively, the mixtures were created on a computer. All programs were written in Python.

The experimental conditions are shown in Table 1. Three types of noise signals were synthesized for each target signal, for a total of three types of signals. The noise level was described by dBFS. The threshold was changed in 10 dB steps for noise removal. A Hamming window was used as the window function. In this study, Signal to Distortion Ratio (SDR) [29], Signal to Noise Ratio (SNR) and Signal to Interference Ratio (SIR) [24] were used for evaluation between the target speech and the denoised speech.

Table 1. Experimental condition.

| Target Speech Male Voice |
|--------------------------|
| Noise signal Intersection noise, white noise (0, −15 [dB]) |
| Noise reduction threshold −90 to −30 [dB] (in 10 [dB] steps) |

The Signal to Distortion Ratio (SDR) is a measure of how much the speech is distorted relative to the target speech after noise reduction. $\hat{S}(\tau, \omega)$ is the signal to be evaluated for the target signal, and $S(\tau, \omega)$ is the target signal. SDR is defined as follows.

$$\text{SDR} = 10 \log_{10} \left( \frac{\sum_{\tau,\omega} |S(\tau, \omega)|}{\sum_{\tau,\omega} |S(\tau, \omega)| - \lambda |\hat{S}(\tau, \omega)|} \right)$$ (18)

$\lambda$ is a parameter used to normalize the power of the signal and is expressed as follows.

$$\lambda = \sqrt{\frac{\sum_{\tau,\omega} |S(\tau, \omega)|}{\sum_{\tau,\omega} |\hat{S}(\tau, \omega)|}}$$ (19)

The signal-to-noise ratio (SNR) is the ratio of the power of the signal to that of the noise, which is used to evaluate how much noise is included in the signal mixture. The higher the value of SNR is, the smaller the effect of noise is. If $S(\tau, \omega)$ is the original target signal and $N(\tau, \omega)$ is the noise signal, SNR is defined as follows.

$$\text{SNR} = 10 \log_{10} \left( \frac{\sum_{\tau,\omega} |S(\tau, \omega)|}{\sum_{\tau,\omega} |N(\tau, \omega)|} \right)$$ (20)

On the other hand, the signal to interference ratio (SIR) can be used to evaluate how much noise is contained in the output after applying the binary mask [11]. Assuming that the binary mask is $M(\tau, \omega)$, SIR is defined as follows:

$$\text{SIR} = 10 \log_{10} \left( \frac{\sum_{\tau,\omega} |M(\tau, \omega)S(\tau, \omega)|}{\sum_{\tau,\omega} |M(\tau, \omega)N(\tau, \omega)|} \right)$$ (21)

By comparing these results, it is possible to understand the effect of noise on the target signal of the speech after denoising.
4.2. Quantitative Evaluation of Experiments

The SDR of the throat microphone to the signal obtained by a standard microphone was 0.474 dB in noise-free conditions. Tables 2–4 show the SDR values between the target signal and the speech after noise reduction and the SIR between the noise signal and the speech after noise reduction when three types of noise are mixed, respectively.

Table 2. Intersection noise.

| Threshold (dB) | SDR  | SIR  |
|----------------|------|------|
| −90            | 0.704| −3.347|
| −80            | 2.632| 0.008068|
| −70            | 8.384| 7.795 |
| −60            | 9.593| 12.97 |
| −50            | 7.393| 15.50 |
| −40            | 4.351| 17.10 |
| −30            | 1.527| 20.83 |

Table 3. White noise (0 dB).

| Threshold (dB) | SDR  | SIR  |
|----------------|------|------|
| −90            | 1.665| −9.767|
| −80            | 3.199| −7.361|
| −70            | 5.007| −4.756|
| −60            | 6.838| −1.369|
| −50            | 6.646| 2.083 |
| −40            | 4.236| 5.033 |
| −30            | 1.522| 8.821 |

Table 4. White noise (−15 dB).

| Threshold (dB) | SDR  | SIR  |
|----------------|------|------|
| −90            | 7.510| 2.233 |
| −80            | 9.579| 4.639 |
| −70            | 10.66| 7.244 |
| −60            | 9.764| 10.63 |
| −50            | 7.534| 14.08 |
| −40            | 4.415| 17.03 |
| −30            | 1.540| 20.82 |

The experimental results are expressed in four significant digits. The background color is yellow when the SDR value is the highest. Tables 2–4 show that the SDR values of the noise-eliminated speech are larger than those of the throat microphone in all cases, indicating that the speech recorded by the noise-eliminated speech is closer to that recorded by a conventional microphone. Next, the SNRs of each noise signal and the target signal are shown in Table 5.

Table 5. SNR results.

| Noise Speech        | SNR  |
|---------------------|------|
| Intersection noise   | −4.158|
| White noise (0 dB)   | −22.96|
| White noise (−15 dB) | −10.96|
Noise is unlikely to be mixed into the voice recorded by the throat microphone. However, the frequency response of the sound recorded by the throat microphone is different from that of the microphone. SDR shows the similarity to clear voice recorded with a general microphone. Hence, the fact that the SDR at the output of the proposed method is higher than that of the throat microphone means that the proposed method can be used to obtain a sound closer to the sound recorded by a normal microphone than by recording with a throat microphone.

Comparing the SNR values in Table 5 with the SIR values in Tables 2–4, it was confirmed that under the same noise conditions, the speech after denoising was larger, and the effect of noise was reduced. The waveform and spectrogram of the results obtained in the experiment are shown in Figures 2–7. The spectrograms and waveforms for the maximum SDR among all the threshold values are shown in the figures. Figures 3, 5 and 7 also show the figure of the binary mask. It is noted that the output spectrogram is colored even in the high-frequency range, but the actual power is almost zero. This is due to the rounding error caused by outputting the spectrogram again after obtaining the output waveform once by the proposed method.

Figures 2 and 3 show the waveform and the spectrogram, respectively, when intersection noise is used as noise.

Figures 4 and 5 show the waveform when we used white noise whose noise level was 0 dB. Figures 6 and 7 show the waveform when we used white noise whose noise level was −15 dB. As shown in Figures 2–7, even if the signal is buried in noise, the objective signal can be extracted well.

![Figure 2. Waveform (Intersection noise).](image-url)
Figure 2. Waveform (Intersection noise).

Figure 3. Spectrogram (Intersection noise).

Figure 4. Waveform (White noise (0 dB)).

Figure 5. Spectrogram (White noise (0 dB)).

Figure 3. Spectrogram (Intersection noise).

Figure 4. Waveform (White noise (0 dB)).
Figure 5. Spectrogram (White noise (0 dB)).

Figure 6. Waveform (White noise (−15 dB)).
5. Discussion and Conclusions

In this study, we proposed a noise reduction method using a combination of a throat microphone and a conventional microphone. In the proposed method, a throat microphone is used as a reference signal to remove the noise in the signal obtained from a general microphone.

In the experiments, white noise and general noise were used as mixtures, and SDR and SIR were used as evaluation criteria to evaluate the noise suppression performance and speech quality. The results showed that the proposed method was able to remove noise. The proposed method applies a mask created based on the information of the throat microphone to the sound recorded by a general microphone, and it is considered that noise is mixed in the remaining frequency range. However, since voice information is sparse in the frequency domain, it is considered that the SN ratio after masking is improved even if the SN ratio is bad in the time domain. As a result, it is considered that masked sound from the general microphone has better audio quality than the sound from the throat microphone.

In the experiment, the voice of the speaker was recorded simultaneously using both a microphone and a throat microphone. Strictly speaking, there may be some time lag, but good results have been obtained experimentally without considering it.

However, there are some problems to be solved in the proposed method. For example, the frequency range in which a throat microphone can record is lower than the frequency range in which a general microphone can record. One solution is to apply the proposed method only to the frequency domain where the throat microphone can record. However, even with this approach, noise remains in the high-frequency components. The inability to record high-frequency components is a problem caused by the use of piezoelectric elements in the throat microphone. Hence, it is necessary to improve the sensor in order to substantially improve this problem.

In order to put this method to practical use, we are thinking of implementing the proposed method in real-time. We would like to proceed with experiments using the
signals recorded in real noisy environments in the future. We also would like to investigate a method to estimate the threshold value automatically.

**Author Contributions:** Conceptualization, J.K. and M.M.; methodology, M.M.; validation, J.K.; investigation, J.K. and M.M.; writing—original draft preparation, J.K.; writing—review and editing, M.M.; supervision, M.M.; project administration, M.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Loizou, P.C. *Speech Enhancement: Theory and Practice*, 2nd ed.; CRC Press: London, UK, 2007.
2. Weiss, M.; Aschkenasy, E.; Parsons, T. *Study and Development of the INTEL Technique for Improving Speech Intelligibility*, Technical Report NSC-TR/4023; Nicolet Scientific Corporation: Northvale, NJ, USA, 1975.
3. Boll, S.F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process* **1979**, *27*, 113–120. [CrossRef]
4. Yamashita, K.; Ogata, S.; Shimamura, T. Improved spectral subtraction utilizing iterative processing. *IEICE Trans Fundamentals* **2005**, *J88-A*, 1246–1257. [CrossRef]
5. McAulay, R.J.; Malpass, M.L. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust. Speech Signal Process* **1980**, *28*, 37–145. [CrossRef]
6. Dendrinos, M.; Bakamides, S.; Carayannis, G. Speech enhancement from noise: A regenerative approach. *Speech Commun.* **1991**, *10*, 45–57. [CrossRef]
7. Ephraim, Y.; Van Trees, H.L. 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. Grais, E.M.; Sen, M.U.; Erdogan, H. 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. Xu, Y.; Du, J.; Dai, L.-R.; Lee, C.-H. An Experimental Study on Speech Enhancement Based on Deep Neural Networks. *IEEE Signal Processing Lett.* **2013**, *21*, 65–68. [CrossRef]
10. Liu, Q.; Wang, W.; Jackson, P.B.; Tang, Y. 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, NY, USA, 2017.
12. Benesty, J.; Chen, J.; Huang, Y. *Microphone Array Signal Processing*; Springer: New York, NY, USA, 2010.
13. Zhao, Q.; Guo, F.; Zu, X.; Chang, Y.; Li, B.; Yuan, X. An Acoustic Signal Enhancement Method Based on Independent Vector Analysis for Moving Target Classification in the Wild. *Sensors* **2017**, *17*, 2224. [CrossRef] [PubMed]
14. Nordhausen, K.; Oja, H. Independent component analysis: A statistical perspective. *Wires Comput. Stat.* **2018**, *10*, e1440. [CrossRef]
15. Addisson, S.; Luis, V. *Independent Component Analysis (ICA): Algorithm, Applications and Ambiguities*; Nova Science Publishers: Hauppauge, NY, USA, 2018.
16. Makino, S.; Lee, T.W.; Sawada, H. (Eds.) *Blind Speech Separation*; Springer: New York, NY, USA, 2007.
17. Taseska, M.; Habets, E.A.P. Blind Source Separation of Moving Sources Using Sparsity-Based Source Detection and Tracking. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 657–670. [CrossRef]
18. Dekens, T.; Verhelst, W.; Capman, F.; Beaugendre, F. 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. Eisemann, E.; Durand, F. Flash photography enhancement via intrinsic relighting. *ACM Trans. Graph.* **2004**, *23*, 673–678. [CrossRef]
20. Petschnigg, G.; Agrawala, M.; Hoppe, H.; Szeliski, R.; Cohen, M.; Toyama, K. Digital photography with flash and no-flash image pairs. *ACM Trans. Graph.* **2004**, *23*, 664–672. [CrossRef]
21. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; Wang, X. A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference. *IEEE Trans. Multimed.* **2010**, *12*, 682–691. [CrossRef]
22. John, V.; Tsuchizawa, S.; Mita, S. Fusion of thermal and visible cameras for the application of pedestrian detection. *Signal Image Video Process.* **2017**, *11*, 517–524. [CrossRef]
23. Fendri, E.; Boukhrris, R.R.; Hammami, M. Fusion of thermal infrared and visible spectra for robust moving object detection. *Pattern Anal. Appl.* **2017**, *20*, 907–926. [CrossRef]
24. Rickard, S.; Yilmaz, O. 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.
25. Ihara, T.; Handa, M.; Nagai, T.; Kurematsu, A. Multi-channel speech separation and localization by frequency assignment. *IEICE Trans Fundam.* 2003, J86-A, 998–1009.
26. Aoki, M.; Yamaguchi, Y.; Furuya, K.; Kataoka, A. 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.
27. Sound Effect Lab. Available online: https://soundeffect-lab.info/sound/environment/ (accessed on 23 December 2021).
28. Hashimoto Tech. Available online: https://hashimoto-tech.jp/local/advan/signwav (accessed on 23 December 2021).
29. Fukui, M.; Shimauchi, S.; Hioka, Y.; Nakagawa, A.; Haneda, Y.; Ohmuro, H.; Kataoka, A. Noise-power estimation based on ratio of stationary noise to input signal for noise reduction. *J. Signal Processing* 2014, 18, 17–28. [CrossRef]