Noise Suppression Using a Differential-type Microphone Array and Two-dimensional Amplitude and Phase Spectra

Koichiro Shiozawa*, Kenji Ozawa*, and Tomohiko Ise†
* University of Yamanashi, Kofu, Japan
E-mail: [t15cs030, ozawa]@yamanashi.ac.jp Tel/Fax: +81-55-220-8586
† Alps Alpine Co., Ltd., Iwaki, Japan
E-mail: tomohiko.ise@alpsalpine.com

***Abstract*—This study aims to achieve noise suppression by processing the output of a microphone array with artificial neural networks (NNs). A differential-type array is used to avoid nonlinear distortions produced by a nonlinear system, such as an NN. The output of the array is considered an image, and it is transformed into a 2-dimensional (2D) spectrum. In the 2D spectrum, the frequency components of a noise are perfectly localized as direct current (DC) components along the spatial frequency axis. In this study, noise suppression was performed by spectral subtraction, after the DC components of noise were instantaneously estimated from the amplitude and phase spectra using independent NNs. As a result, the noise reduction performance of the proposed method with a 16-cm-long array was approximately 24 dB. Although the NNs were trained by white noise, the system was effective for speech and music signals as well as for white noise.**

I. Introduction

Noise suppression using a microphone array is effective when recording a target sound in a noisy environment [1]. If an artificial neural network (NN) is used, which is one of machine learning techniques, a small microphone array having a sharp directivity can be achieved [2], [3], [4], [5], [6], [7], [8].

A microphone-array-based noise suppression method using an NN that managed the waveform of a target and noise was proposed in one of the previous studies [8]. A differential-type microphone array [2], [3] was utilized to avoid the production of distortion components of the target. Despite this advantage, there were two outstanding issues: 1) unexpected broadband noises that were not related to the input noise spectrum were produced due to the error in the NN processing; 2) versatility was insufficient because the system was not effective for sounds that were different from the sound used in training the NN. To overcome these issues, this study proposes a novel noise-suppression method based on the two-dimensional (2D) amplitude and phase spectra of the outputs from a microphone array.

Though the 2D spectrum of the output signals from a microphone array has been previously used for spatial filtering [7], [9], [10], [11], [12], [13], [14], [15], only the amplitude spectrum was addressed; however, this study considers the phase spectrum as well. Moreover, the method proposed in this study differs from those in other studies [9], [10], [11], [12], [13], [14] in that the noise is suppressed by the spectral subtraction based on the instantaneous estimation of the noise spectrum [7], [15].

II. Fundamentals of the Proposed Method

A. Differential-type array

Generally, distortion components of the input signal are produced when a signal is processed with a nonlinear system, such as an NN. However, a differential-type array proposed by Kobatake *et al.* [2], [3] does not produce any distortion component with respect to the target sound because the target was subtracted before the processing in an NN. Figure 1 presents a block diagram of the proposed system where the input part consists of a differential-type array.

In a differential-type array, $M+1$ microphones are arranged on a straight line; one of them is called a reference microphone (RM), and the other M microphones are called sensor microphones (SMs). The output from the RM is subtracted from the outputs of the SMs to obtain differential signals. A target sound is assumed to come from the vertical direction of the array, and this direction is defined as 0° with respect to the direction of arrival (DOA) of a sound. The differential signals do not consist of target components because the target sound arrives at all the microphones at exactly the same time. These differential signals are processed to estimate the noise signal at the RM, then it is subtracted from the RM output; as a result, the target signal is obtained. The target is guaranteed to be distortion-free because the target sound does not go through the processor, including an NN.

When the output from the RM is denoted as $x_{\mathrm{RM}}(n)$, and the output and differential signals from the m-th SM are given as $x(n,m)$ and $d(n,m)$, respectively, the differential signals are given by the following equation:

$$\begin{aligned} d(n,m) &= x(n,m) - x_{\mathrm{RM}}(n) \\ &= s(n) + u(n-\tau_m) - \{s(n) + u(n)\} \\ &= u(n-\tau_m) - u(n),\ m = 0, 1, \ldots, M-1 \end{aligned} \quad (1)$$

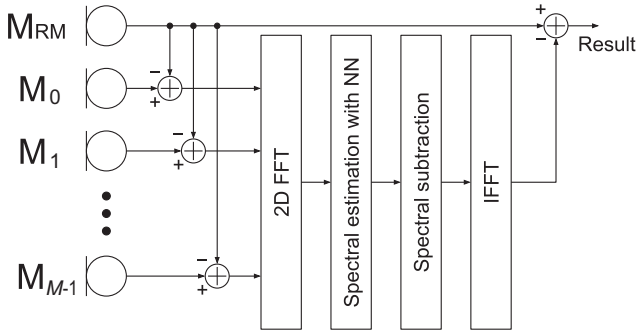


Fig. 1. Block diagram of the proposal system.

where $s(n)$ is a target signal with the DOA of 0° and $u(n)$ is a noise with the DOA of θ observed at the RM. When the microphones are assumed to be uniformly arranged and spaced with l, the temporal delay τ_m at the m-th SM from the RM is given by $\tau_m = (m+1)l\sin\theta/c$ where c is the sound velocity of 340 m/s. Note, $d(n,m)$ does not include $s(n)$.

B. Noise suppression based on a differential image

The authors have defined a *spatio-temporal sound pressure distribution image* by a 2D image consisted of temporal sequences of luminance, which are produced by transforming the instantaneous sound pressure of microphone outputs into luminance, arranged in parallel, corresponding to the microphone positions in an array [14]. In their previous study [8], they attempted to suppress a noise based on such an image created from the differential signals. Hereafter, this image is referred to as a *differential image*. Figure 2 presents an example differential image when white noise arrives from 90° to an array consisted of nine microphones (8 SMs and 1 RM) on a straight line with the uniform intervals of 2 cm. The abscissa is the time and the ordinate corresponds to the index of the eight SMs from 0 to 7 where the RM was placed next to M_0 as shown in Fig. 1. This image was fed into an NN in which the noise components included in $x_{\mathrm{RM}}(n)$ were estimated. Finally, the estimated values were subtracted from $x_{\mathrm{RM}}(n)$ to obtain the target signal.

In the previous study [8], the estimation of the noise components was conducted by recalling the noise waveform by the NN. In this study, however, the estimation is conducted by recalling the noise spectrum.

III. Properties of the 2D spectrum of a differential image

A. 2D spectrum of a differential image

By applying the 2D fast Fourier transform (FFT) to a differential image, its 2D spectrum is obtained. In this study, the abscissa and ordinate of a 2D spectrum are called the temporal and spatial frequencies, respectively. The direct current (DC) component of the spatial frequency is referred to as spatial #0 bin.

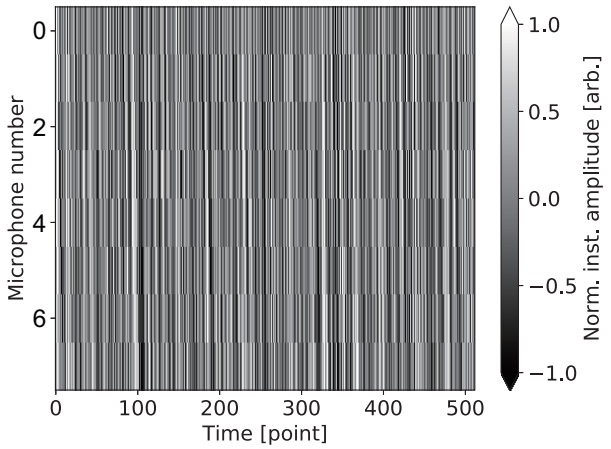


Fig. 2. Differential image obtained by eight sensor microphones when white noise arrives from 90°. The length of a temporal segment is 512 points.

A differential image with temporal points of N and SMs of M is given as a function $d(n,m)$ where n and m are the discrete time and microphone index. Its 2D spectrum $D(k_{\mathrm{t}},k_{\mathrm{s}})$ is obtained using the following equation:

$$\begin{aligned} D(k_{\mathrm{t}},k_{\mathrm{s}}) &= \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} d(n,m)W_N^{nk_{\mathrm{t}}}W_M^{mk_{\mathrm{s}}} \\ &= \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\left\{u(n-\tau_m)-u(n)\right\}W_N^{nk_{\mathrm{t}}}W_M^{mk_{\mathrm{s}}} \\ &= \frac{1}{M}\sum_{m=0}^{M-1}\left\{U(k_{\mathrm{t}})W_N^{\tau_m f_s k_{\mathrm{t}}}-U(k_{\mathrm{t}})\right\}W_M^{mk_{\mathrm{s}}} \end{aligned} \tag{2}$$

where k_{t} and k_{s} are the indices of the temporal and spatial frequency bins, respectively. Moreover, $W_N = e^{-j2\pi/N}$ and $W_M = e^{-j2\pi/M}$ denote the rotators for the temporal and spatial axes, respectively.

The second term of the right side of Eq. (2) represents the *noise spectrum at the RM*, $U_{\mathrm{RM}}(k_{\mathrm{t}},k_{\mathrm{s}})$ that was subtracted for deriving the differential spectrum. The term is extracted and modified to the following form:

$$\begin{aligned} U_{\mathrm{RM}}(k_{\mathrm{t}},k_{\mathrm{s}}) &= -\frac{U(k_{\mathrm{t}})}{M}\sum_{m=0}^{M-1}W_M^{mk_{\mathrm{s}}} \\ &= \begin{cases} -U(k_{\mathrm{t}}) & (k_{\mathrm{s}}=0) \\ 0 & (k_{\mathrm{s}}\neq 0). \end{cases} \end{aligned} \tag{3}$$

This equation implies that $U_{\mathrm{RM}}(k_{\mathrm{t}},k_{\mathrm{s}})$ is perfectly localized as the spatial DC components ($k_{\mathrm{s}}=0$). Although some studies have considered the 2D spectrum for noise suppression [9], [10], [11], [12], [13], they have not used the following property: the aimed components localize as the spatial DC components, i.e., the essential aim of this study is to use this property.

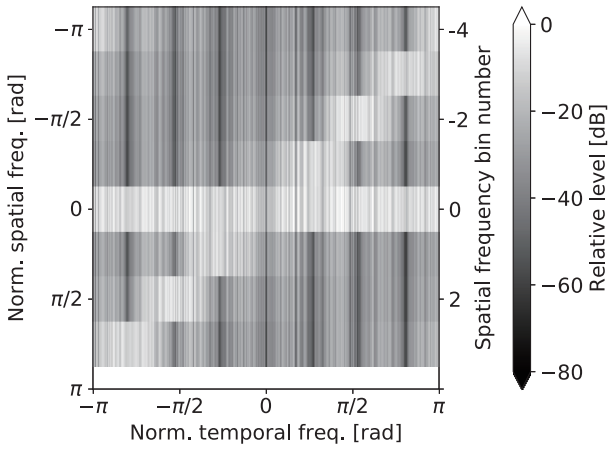


Fig. 3. Two dimensional amplitude spectrum of the differential image in Fig. 2.

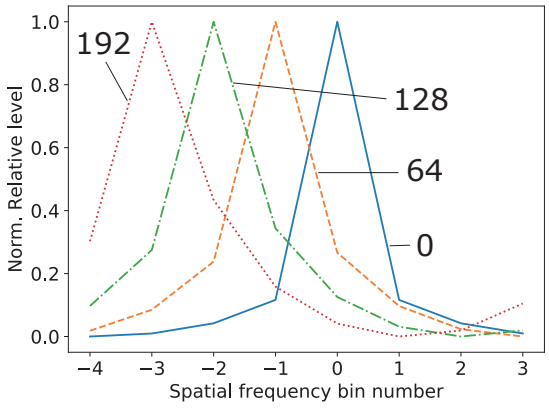


Fig. 4. Examples of amplitude spatial spectral patterns taken from Fig. 3. The parameters are the temporal frequency bin numbers.

The first term of the right side of Eq. (2) exhibits the *noise spectrum regarding the SMs*, $U_{\mathrm{SM}}(k_{\mathrm{t}}, k_{\mathrm{s}})$.

$$U_{\mathrm{SM}}(k_{\mathrm{t}}, k_{\mathrm{s}}) = \frac{U(k_{\mathrm{t}})}{M} \sum_{m=0}^{M-1} W_N^{\tau_m f_s k_{\mathrm{t}}} W_M^{m k_{\mathrm{s}}} \quad (4)$$

When the spatial #0 bin ($k_s = 0$) is examined, Eq. (4) is modified to the following equation:

$$U_{\mathrm{SM}}(k_{\mathrm{t}}, 0) = \frac{U(k_{\mathrm{t}})}{M} \sum_{m=0}^{M-1} W_N^{\tau_m f_s k_{\mathrm{t}}}. \quad (5)$$

This equals to the spectrum that is obtained when the delay-and-sum (DAS) beamformer [1] is applied to the noises observed at SMs.

These imply that the two spectra of the *noise at RM which was subtracted* ($-U(k_{\mathrm{t}})$) and *the output of the DAS beamformer with SMs* are superimposed on the spatial #0 bin to form $D(k_{\mathrm{t}}, 0)$. Thus, $-U(k_{\mathrm{t}})$ can be extracted by the following two steps: first, estimate the *noise components at SM on #0 bin*, $U_{\mathrm{SM}}(k_{\mathrm{t}}, 0)$ using the components at bins other than #0, $U_{\mathrm{SM}}(k_{\mathrm{t}}, k_{\mathrm{s}})$ with $k_{\mathrm{s}} \neq 0$; and second, subtract the estimated values from the observed values at #0 bin. Finally, the target signal $s(n)$ is determined by subtracting the waveform $u(n)$ that is obtained by applying the inverse FFT (IFFT) of the estimated noise $U(k_{\mathrm{t}})$, from the output of the RM.

Note, if $-U(k_{\mathrm{t}})$ is replaced with $S(k_{\mathrm{t}})$, this idea can be regarded to be equivalent to the processing for 2D spectrum when a non-differential-type array is used [7].

B. Amplitude estimation

Based on the aforementioned properties, the amplitude and phase of a noise are estimated. Figure 3 presents the 2D amplitude spectrum $|D(k_{\mathrm{t}}, k_{\mathrm{s}})|$ of the differential image shown in Fig. 2. When k_{t} is fixed to a specific value k_0, a spectral pattern of $|D(k_0, k_{\mathrm{s}})|$ is referred to as a spatial amplitude spectral pattern. Four examples of spatial amplitude spectral patterns of only $|U_{\mathrm{SM}}(k_{\mathrm{t}}, k_{\mathrm{s}})|$ are shown in Fig. 4 which were taken from Fig. 3. Every pattern has a single peak, and it shifts along the abscissa as the temporal frequency k_{t} increases. Thus, a light area along the diagonal in Fig. 3 can be observed.

The amplitude of the *noise observed at the RM*, $|U(k_{\mathrm{t}})|$, is estimated as follows. First, the values of a spectral pattern, except for #0 bin, are fed into an NN and the NN recalls the value at #0 bin. This will be done for every temporal frequency bin and the spatial DC components $|U_{\mathrm{SM}}(k_{\mathrm{t}}, 0)|$ are estimated as a result. Subsequently, these values are subtracted from $|D(k_{\mathrm{t}}, 0)|$ to obtain $|U(k_{\mathrm{t}})|$.

C. Phase estimation

In the authors' previous study [7], spectral subtraction was conducted based on the amplitude spectrum only, i.e., the phases of noise components were not estimated but the true values were used. In this study, the phase estimation is essential because the two noise components of the *noise at the RM* and *noise regarding the SMs* are included in the #0 bin spectrum.

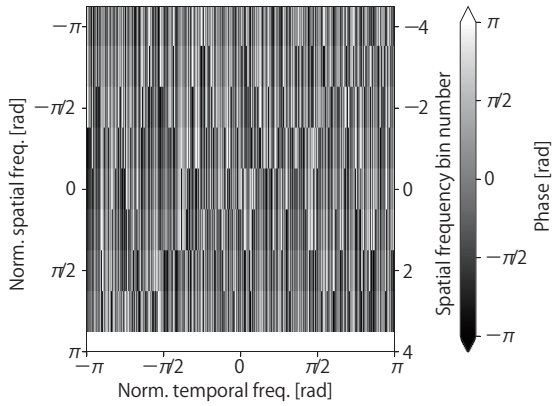
Figure 5(a) exhibits the 2D phase spectrum $\mathrm{Arg}\,[D(k_{\mathrm{t}}, k_{\mathrm{s}})]$ of the differential image shown in Fig. 2. The phase spectrum presents complex changes, making it difficult to estimate the phase at the spatial #0 bin. However, if the same phase data are presented as the relative phases to #1 bin, a few systematic changes can be observed, as shown in Figure 5(b).

Four examples of phase spatial–spectral patterns observed in Fig. 5(b) are shown in Fig. 6. Based on these systematic changes, the phases at the spatial #0 bin are to be estimated by preparing an independent NN.

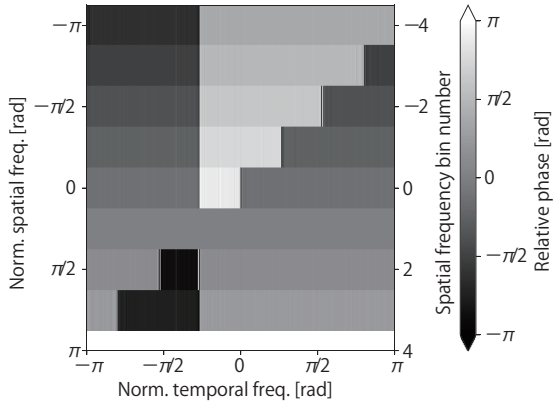
IV. Estimation of Noise Spectrum using NNs

A. Setup of estimation

Computational simulation experiments were conducted by assuming a system shown in Fig. 1: nine microphones were arranged in 2-cm intervals. A temporal segment consisted of 512 points with the sampling frequency of 16 kHz. Because the



(a) Observed phases.



(b) Representation as the relative phase to #1 bin.

Fig. 5. Two dimensional phase spectrum of the differential image in Fig. 2.

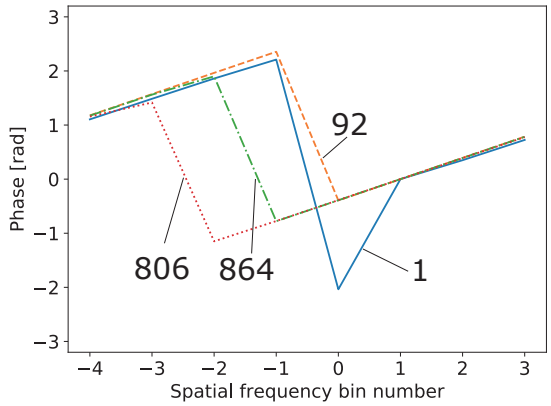


Fig. 6. Examples of phase spatial patterns taken from Fig. 5(b). The parameters are the temporal frequency bin numbers.

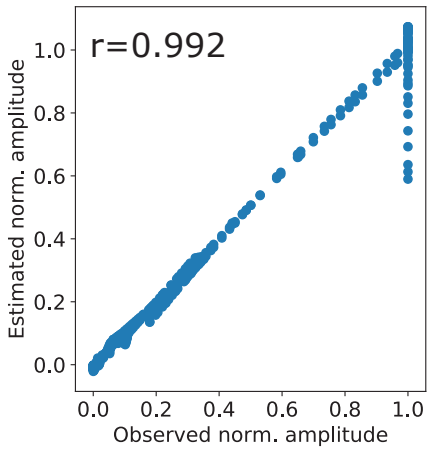


Fig. 7. Results of the closed test: scatter plot of amplitude estimation.

estimation processing was conducted for every 32-ms segment independently, it is referred to as instantaneous estimation [7], [15]. Temporal segments were created using the 512-point Hanning window with a 256-point shift, while no windowing was applied to the spatial axis before conducting the 2D FFT.

Two NNs were prepared for estimating the amplitude and phase of $U_{\mathrm{SM}}(k_{\mathrm{t}}, 0)$ independently at every temporal frequency bin in a segment. The numbers of units for each NN were 7-15-1 for the input-hidden-output layers. The NN for amplitude estimation has two hidden layers while that for phase estimation has one. The activation functions for the hidden layers were the rectified linear unit (ReLU).

The training data of the NNs were 512 spatial–spectral patterns taken from the 2D spectrum of a differential image when white noise arrived from 90°. Each temporal segment was a 512-point waveform; thus, the 2D spectrum consisted of 512 temporal frequency bins.

Chainer V4 [16] was used for constructing the NNs with the optimization algorithm of Adam. Training sessions of 1000 epochs with the batch size of 32 were conducted separately for the two NNs.

B. Learning to estimate the amplitude

First, every amplitude spatial–spectral pattern was represented as relative levels where the maximum value was assigned to 0 dB. Subsequently, the levels were normalized: the minimum and maximum levels were assigned to 0 and 1, respectively. The NN was trained so that the normalized amplitude at the spatial #0 bin was to be recalled using the amplitude data of the remaining seven spatial bins.

As a result of learning, the closed test results are shown in Fig. 7. The overall accuracy is sufficient but there are discrepancies when the observed value is 1. This occurs when a spatial spectrum pattern has its peak at the spatial #0 bin, such as the pattern of the temporal #0 bin in Fig. 4. These patterns are observed when the temporal frequency is low or the DOA of noise is small. It may be difficult for the NN to recall its peak value using small data at the hems.

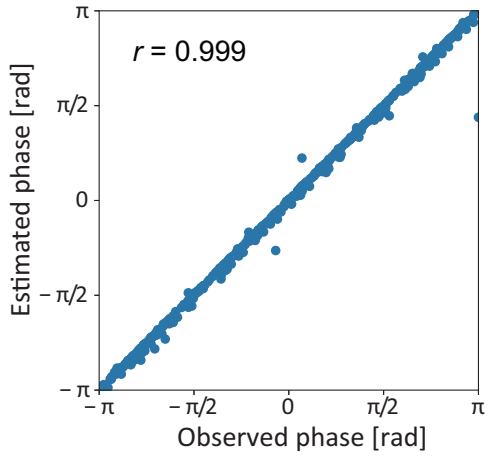


Fig. 8. Results of the closed test: scatter plot of phase estimation.

C. Learning to estimate the phase

Based on the phase characteristics shown in Fig. 6, the NN was trained so that the phase at the spatial #0 bin was estimated using the remaining six out of the seven bins: the value of #1 bin was excluded due to its value always equal to 0 because it was the reference of the relative phase. In addition to the relative phase data, information indicating whether the DOA of noise was positive or negative was used. This information could be easily obtained by the fact that the peak existed in the positive or negative spatial bin. When the peak existed at the negative or positive bin, the value of -1 or 1 was given to one of the input cells of the NN. The NN was trained using these 6-phase data and one DOA information to recall the relative phase value at the spatial #0 bin.

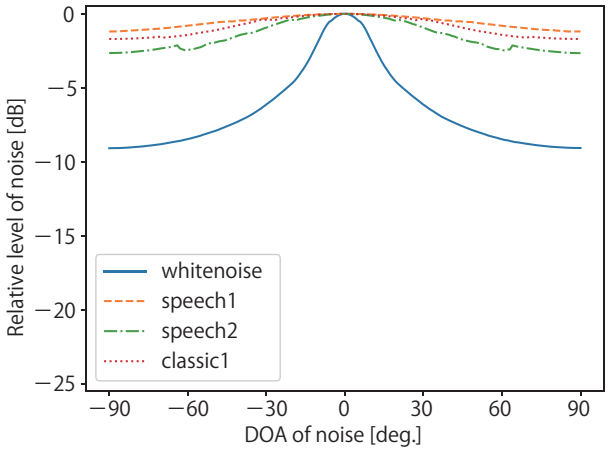
As a result of learning, the closed test results are shown in Fig. 8. The overall accuracy is excellent, except for some outliers which are the data at very low temporal frequencies ($<$ 100 Hz), or high frequencies near the Nyquist frequency.

V. Evaluation of the Performance of the System

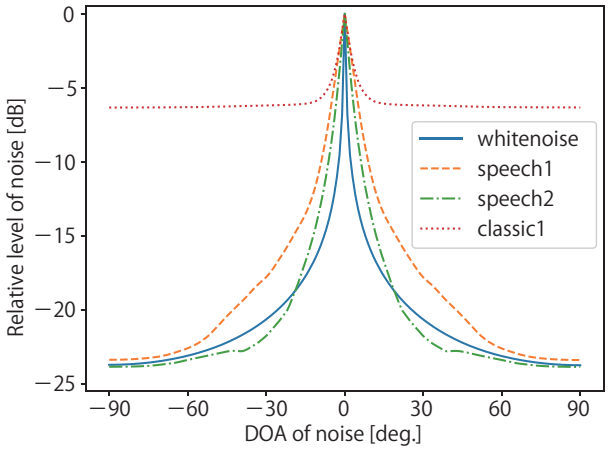
A. Treatment of ill estimations

The estimation of amplitude is relatively difficult at low frequencies where the estimation error is significant, as shown in Fig. 7. In such a case, spectral subtraction of the estimated noise may evoke additional noise. To avoid this, an ill estimation condition was defined and coped with the errors as follows: The noise at the spatial #0 bin $U_{\mathrm{SM}}(k_{\mathrm{t}}, 0)$ was estimated based on the values $U_{\mathrm{SM}}(k_{\mathrm{t}}, k_{\mathrm{s}})$ where $k_{\mathrm{s}} \neq 0$. Here, the estimated values are referred to as $E_{\mathrm{SM}}(k_{\mathrm{t}}, 0)$. Ill estimations occur when the difference between the true value $U_{\mathrm{SM}}(k_{\mathrm{t}}, 0)$ and an estimated value $E_{\mathrm{SM}}(k_{\mathrm{t}}, 0)$ is large. The noise signal is especially enhanced when the following condition holds:

$$|U_{\mathrm{SM}}(k_{\mathrm{t}}, 0) - E_{\mathrm{SM}}(k_{\mathrm{t}}, 0)| \quad > \quad |-U(k_{\mathrm{t}})|. \tag{6}$$



(a) Amplitude estimation only.



(b) Amplitude as well as phase estimation.

Fig. 9. Noise suppression performance by the proposal system.

However, $U_{\mathrm{SM}}(k_{\mathrm{t}}, 0)$ and $-U(k_{\mathrm{t}})$ are not observed; thus, it is impossible to determine whether an ill estimation occurred based on the observed value of $D(k_{\mathrm{t}}, 0)$ and the estimated value of $E_{\mathrm{SM}}(k_{\mathrm{t}}, 0)$. An ill estimation was determined to occur when the estimated amplitude was 10 times larger than the observed value ($10 \times |D(k_{\mathrm{t}}, 0)| \leq |E_{\mathrm{SM}}(k_{\mathrm{t}}, 0)|$). If this condition is satisfied, the estimated value is considered equal to zero ($|E_{\mathrm{SM}}(k_{\mathrm{t}}, 0)| = 0$). This indicates that the system is equivalent to the conventional DAS array when an ill estimation occurs.

B. Evaluation of noise suppression

Noise suppression performances of the proposed system were tested for four types of noise: white noise, words of "Sapporo" (Speech 1), and "Asahikawa" (Speech 2) taken from a database [17], and a piece of music (classic1: classical music No. 1 [18]).

Figure 9(a) presents the results when only amplitudes were

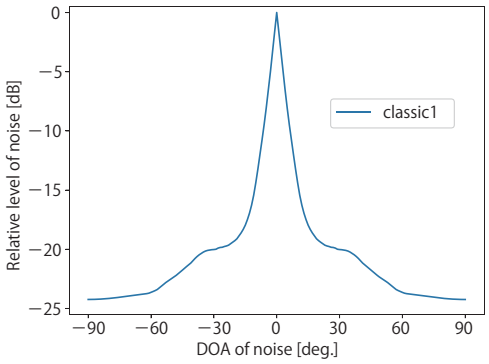


Fig. 10. Noise suppression for the musical noise with bandpass filtering of 100 to 7500 Hz.

estimated, i.e., no estimation on the phase was conducted and the $D(k_t, 0)$ phase was used for the IFFT process. It is uncertain whether a sufficient amount of noise was suppressed.

Figure 9(b) exhibits the results when the phase and amplitudes were estimated. Significantly sharp directivities are achieved for white noise and speech signals, and the maximum suppressions reach approximately -24 dB. The maximum suppression is larger than the results of the previous study [8] in which the values were approximately -21 and -18 dB for the noise used in the learning and a speech signal, respectively. Moreover, if these SMs are used as the conventional DAS beamformer or the minimum variance (MV) beamformer [1], the maximum noise reduction could reach at most -10 or -20 dB, respectively [7]. These imply that the proposed system is effective.

However, the maximum suppression is restricted to approximately -6 dB for the musical noise (classic 1). This is because lower frequency components are dominant for this musical noise. As mentioned above, the proposed system has the tendency for a difficult estimation of the amplitudes of lower frequency components. When the musical noise was preprocessed with a bandpass filter with the passband of 100 to 7500 Hz, the maximum suppression reached approximately -24 dB, as shown in Fig. 10. Note, no bandpass filtering was applied for white noise in Fig 9. This is because white noise has plenty of high frequency components so that well suppression was achieved even when low frequency components existed.

These results suggest that the proposed method is better than the previous method [8] in which processing is conducted in the time domain. Moreover, the present method overcame the issue in the previous method of being effective only for the noise that had been used in training the NN. However, the present method needs further studies for improvement in the estimation of lower frequencies.

VI. Conclusion

This study proposed a noise suppression system composed of a differential-type microphone array and two NNs. In the NNs, 2D amplitude and phase spectra are addressed to estimate the property of a noise. As a result, the system achieved a sharp directivity for noises that have not been used for training the NNs. The authors succeeded to overcome the issues in their previous method.

The present system estimates the amplitude and phase of a noise component using two independent NNs. Further studies will be required to use a complex-number NN for better performance.

Acknowledgment

This project utilized sound data from the Tohoku University–Matsushita–Isolated Word (TMW) database and the Real–World Computing (RWC) music database.

References

[1] M. R. Bai, J-G. Ih, and J. Benesty, *Acoustic Array Systems: Theory, Implementation, and Application*, John Wiley & Sons, Singapore, 2013.
[2] H. Kobatake, W. Morita and Y. Yano, "Super directive sensor array with neural network structure," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP 92)*, vol. 2, pp. 321–324, 1992.
[3] W. Morita and H. Kobatake, "Super directive microphone array system with nonlinear multi-layer structure," *J. Acoust. Soc. Jpn. (J)*, vol. 49, pp. 28–33, 1993.
[4] A. Iseki, Y. Kinoshita, and K. Ozawa, "Optimization of neural-network-based superdirective microphone-array system using a genetic algorithm," *Acoust. Sci. & Tech.*, vol. 36, pp. 326–332, 2015.
[5] A. Iseki, K. Ozawa, and Y. Kinoshita, "Neural-network-based microphone-array system trained with temporal-spatial patterns of multiple sinusoidal signals," *Acoust. Sci. & Tech.*, vol. 38, pp. 63–70, 2017.
[6] M. Mizumachi, M. Origuchi and Y. Nishijima, "Non-linear broadband beamformer with deep neural network," *Proc. 24th Int. Cong. on Sound and Vibration (ICSV24)*, Paper ID: 873, 6 pages, 2017.
[7] K. Ozawa, M. Morise, and S. Sakamoto, "Sound source separation by instantaneous-estimation-based spectral subtraction," *Proc. 5th Int. Conf. on Systems and Informatics (ICSAI 2018)*, pp. 870–875, 2018.
[8] K. Ozawa, K. Shiozawa, and T. Ise, "Sound source separation using spatio-temporal sound pressure distribution images and machine learning," *Proc. of 2019 Amity Int. Conf. on Artificial Intelligence (AICAI 2019)*, pp. 54–60 (2019).
[9] D. A. Gray, "Frequency wavenumber beamforming by use of the two dimensional Fourier transform," *Technical Report WSRL-0162-TR, DoD*, pp. 1–22, 1980.
[10] K. Nishikawa, H. Ohno, X. Tang, T. Kanamori, and H. Naono, "A design method of 2D FIR fan filters for wideband beam forming by means of 2D Fourier series approximation," *Electronics and Communications in Japan, Part 3*, vol. 85, pp. 38–49, 2002.
[11] F. Pinto and M. Vetterli, "Wave field coding in the spacetime frequency domain," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, pp. 365–368, 2008.
[12] F. Pinto and M. Vetterli, "Space-time-frequency processing of acoustic wave fields: theory, algorithms, and applications," *IEEE Trans. Signal Process.*, vol. 58, pp. 4608–4620, 2010.
[13] F. Pinto, M. Kolundžija, and M. Vetterli, "Digital acoustics: processing wave fields in space and time using DSP tools," *APSIPA Trans. Signal and Information Process.*, vol. 3, pp. 1–21, 2014.
[14] M. Ito, K. Ozawa, M. Morise, G. Shimizu, and S. Sakamoto, "Sound source separation using image signal processing based on sparsity of sound field," *J. Acost. Soc. Amer.*, vol. 140, p. 3058, 2016.
[15] K. Ozawa, M. Morise, S. Sakamoto, and K. Watanabe, "Sound source separation by spectral subtraction based on instantaneous estimation of noise spectrum," *Proc. 6th Int. Conf. on Systems and Informatics (ICSAI 2019), pp. 885–890, 2019.*
[16] Chainer: A Powerful, Flexible, and Intuitive Framework for Neural Networks, https://chainer.org/ (2018. 12. 21).
[17] Tohoku University–Matsushita–Isolated Word (TMW) database, http://research.nii.ac.jp/src/en/TMW.html (2020. 6. 18).
[18] The RWC (Real–World Computing) Music Database, Classical music No. 1 to 6, https://staff.aist.go.jp/m.goto/RWC-MDB/ (2020. 6. 18).