Designing the Wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra

Nobutaka Ito, Nobutaka Ono, Emmanuel Vincent, Shigeki Sagayama

DESIGNING THE WIENER POST-FILTER FOR DIFFUSE NOISE SUPPRESSION USING IMAGINARY PARTS OF INTER-CHANNEL CROSS-SPECTRA

Nobutaka Ito[†*], *Nobutaka Ono*[†], *Emmanuel Vincent*[*], *and Shigeki Sagayama*[†]

[†] Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
[*] METISS group, IRISA-INRIA
Campus de Beaulieu, 35402 Rennes Cedex, France

ABSTRACT

This paper describes a new design of the Wiener post-filter for diffuse noise suppression. The Wiener post-filter is well-known as an effective post-processing of the minimum variance distortionless response beamformer, and its output is the optimal estimate of the target signal in the sense of the minimum mean square error. It is essential to accurately estimate the target power spectrum from the observed signals contaminated by noise when designing the Wiener post-filter. In our method, it is estimated from the imaginary parts of the inter-channel observation cross-spectra, under the assumption that the inter-channel noise cross-spectra are real-valued. The post-filter is designed using the estimate and this design is shown to be effective even for a small-sized array through experiments using simulated and real environmental noise.

***Index Terms*—** Diffuse noise, microphone arrays, noise suppression, post-filtering, speech enhancement.

1. INTRODUCTION

Much research has been devoted to microphone array signal processing for enhancing the sounds from a desired direction. Especially, noise suppression with a small-sized array is an important issue, because the array size is often limited in various applications such as automatic speech recognition, hearing aids, mobile communications, and recording. The fundamental delay-and-sum beamformer requires a large array in order to achieve sharp directivity. Adaptive beamformers such as the Minimum Variance Distortionless Response (MVDR) beamformer [1] effectively suppress localized noise arriving from few directions, regardless of the array size. However, they do not sufficiently suppress diffuse noise encountered, *e.g.* at cocktail parties, in reverberant rooms, or in vehicles, because it arrives from many directions. In comparison, post-filtering, *i.e.* post-processing of the output of a beamformer with a time-frequency mask, is suitable for suppression of diffuse noise [2–11]. It has been shown by Simmer *et al.* [5] and Van Trees [12] that the Minimum Mean Square Error (MMSE) estimate of the target signal is obtained by the MVDR beamformer followed by a time-frequency mask called the Wiener post-filter [5, 6, 10, 11]. In order to design the Wiener post-filter, it is essential to accurately estimate the target power spectrum or equivalently the target autocorrelation function from the observed signals containing noise.

Zelinski [2] proposed estimating the target autocorrelation function from the inter-channel observation cross-correlation functions, which are noise-free under the assumption that noise components in different channels are uncorrelated. Since the assumption is valid only when the distances between microphones are large enough compared to the wavelength, the method does not work well for a small-sized array or at low frequencies. McCowan *et al.* [6] proposed estimating the target power spectrum assuming that the inter-channel noise coherences—normalized inter-channel noise cross-spectra—are known, which is the case in some ideal noise fields such as spherically isotropic noise fields [13]. In real environments, however, the accurate values of the inter-channel noise coherences are not necessarily available because they can vary significantly depending on conditions such as the arrangement of the noise sources and the room shape.

Aiming to suppress diffuse noise effectively with a small-sized array, we proposed a blind noise decorrelation approach, in which we diagonalize the spatial covariance matrix of isotropic noise without knowing its value, utilizing symmetrical arrays we call crystal arrays [11]. The target power spectrum is then estimated from the noise-free off-diagonal elements of the spatial observation covariance matrix, and the Wiener post-filter is designed using this estimate. Thanks to noise decorrelation, this method can accurately estimate the target power spectrum to be used in post-filter design even with a small-sized array or at low frequencies, where noise is highly correlated between channels. However, the method is limited in the sense that it is applicable only to crystal arrays.

In this paper, we present a novel design of the Wiener post-filter, which works well with an arbitrarily arranged as well as small-sized array. The method is based on the estimation of the target power spectrum from the imaginary parts of the inter-channel observation cross-spectra, under the assumption that the inter-channel noise cross-spectra are real-valued. We show that the assumption is satisfied to a certain extent by real environmental noise. Compared with our previous method [11], applicability to an arbitrary array widens the range of application: for instance, we can place more microphones in a limited area to improve the noise suppression performance. The rest of the paper is organized as follows. In Section 2, we describe the proposed method for designing the Wiener post-filter. In Section 3, we show results of experiments using simulated and real environmental noise to demonstrate the effectiveness of the method, and some conclusions are drawn in Section 4.

2. PROPOSED METHOD

2.1. Notation

We use the following notation throughout the paper. The superscripts *, $^{\mathsf{T}}$, and $^{\mathsf{H}}$ denote complex conjugation, transposition, and Hermitian transposition, respectively. Signals are represented in the time-frequency domain with $\tau \in \mathbb{Z}$ and $\omega \in \mathbb{R}$ representing the frame index and the angular frequency, respectively. The cross-spectrum of scalar signals $\alpha(\tau, \omega)$ and $\beta(\tau, \omega)$ is denoted by

$$\phi_{\alpha\beta}(\tau, \omega) \triangleq \mathcal{E}[\alpha(\tau, \omega)\beta^*(\tau, \omega)], \tag{1}$$

and the covariance matrix of a zero-mean random vector $\boldsymbol{\gamma}(\tau,\omega)$ by

$$\boldsymbol{\Phi}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\tau,\omega) \triangleq \mathcal{E}[\boldsymbol{\gamma}(\tau,\omega)\boldsymbol{\gamma}^{\mathsf{H}}(\tau,\omega)], \tag{2}$$

where $\mathcal{E}[\cdot]$ denotes expectation.

2.2. Observation Model

In this paper, we consider a situation where a microphone array receives a target signal arriving from a known direction in the presence of diffuse noise. Let $s(\tau,\omega)$ be the target signal component at a reference point (*e.g.* the array centroid), and $x_m(\tau,\omega)$ and $v_m(\tau,\omega)$ be the observed signal and the diffuse noise component at the m-th microphone, respectively. Under the assumption that the target source is in the far field of the array, the transfer function from $s(\tau,\omega)$ to $x_m(\tau,\omega)$ is modeled as $d_m(\omega) \triangleq \mathrm{e}^{-\mathrm{j}\omega\delta_m}$, where δ_m denotes the delay of arrival of the target signal from the reference point to the m-th microphone. The delay δ_m can be calculated using the Direction Of Arrival (DOA) of the target signal and the microphone positions, which are both assumed to be known in this paper. Consequently, our observation model is given by

$$x_m(\tau,\omega) = s(\tau,\omega)d_m(\omega) + v_m(\tau,\omega). \tag{3}$$

This equation can be rewritten in a vector form as

$$\boldsymbol{x}(\tau,\omega) = s(\tau,\omega)\boldsymbol{d}(\omega) + \boldsymbol{v}(\tau,\omega), \tag{4}$$

where the vectors are defined by

$$\boldsymbol{x}(\tau,\omega) \triangleq \begin{bmatrix} x_1(\tau,\omega) & x_2(\tau,\omega) & \dots & x_M(\tau,\omega) \end{bmatrix}^{\mathsf{T}}, \tag{5}$$

$$\boldsymbol{d}(\omega) \triangleq \begin{bmatrix} d_1(\omega) & d_2(\omega) & \dots & d_M(\omega) \end{bmatrix}^{\mathsf{T}}, \tag{6}$$

$$\boldsymbol{v}(\tau,\omega) \triangleq \begin{bmatrix} v_1(\tau,\omega) & v_2(\tau,\omega) & \dots & v_M(\tau,\omega) \end{bmatrix}^{\mathsf{T}}, \tag{7}$$

with M denoting the number of microphones in the array. It is assumed that $s(\tau,\omega)$ and $\boldsymbol{v}(\tau,\omega)$ are uncorrelated zero-mean random variables.

2.3. Wiener Post-Filter

Of the linear estimators of the target signal of the form

$$\hat{s}(\tau,\omega) \triangleq \boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{x}(\tau,\omega), \tag{8}$$

the one minimizing the mean square error $\mathcal{E}[|\hat{s}(\tau,\omega) - s(\tau,\omega)|^2]$ is given by the following equation [5, 12]:

$$\hat{s}_{\mathrm{MMSE}}(\tau,\omega) \triangleq \phi_{ss}(\tau,\omega)\boldsymbol{d}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{x}\boldsymbol{x}}^{-1}(\tau,\omega)\boldsymbol{x}(\tau,\omega). \tag{9}$$

It can be shown that (9) is closely related to the output of the MVDR beamformer given by

$$y(\tau,\omega) \triangleq \frac{\boldsymbol{d}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{x}\boldsymbol{x}}^{-1}(\tau,\omega)\boldsymbol{x}(\tau,\omega)}{\boldsymbol{d}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{x}\boldsymbol{x}}^{-1}(\tau,\omega)\boldsymbol{d}(\omega)}. \tag{10}$$

From this equation, the power spectrum of $y(\tau,\omega)$ is given by

$$\phi_{yy}(\tau,\omega) = \frac{1}{\boldsymbol{d}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{x}\boldsymbol{x}}^{-1}(\tau,\omega)\boldsymbol{d}(\omega)}. \tag{11}$$

Therefore, (9) is rewritten as follows [5, 12]:

$$\hat{s}_{\mathrm{MMSE}}(\tau,\omega) = \underbrace{\frac{\phi_{ss}(\tau,\omega)}{\phi_{yy}(\tau,\omega)}}_{\triangleq\, p(\tau,\omega)} \cdot \underbrace{\frac{\boldsymbol{d}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{x}\boldsymbol{x}}^{-1}(\tau,\omega)\boldsymbol{x}(\tau,\omega)}{\boldsymbol{d}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{x}\boldsymbol{x}}^{-1}(\tau,\omega)\boldsymbol{d}(\omega)}}_{=\, y(\tau,\omega)}. \tag{12}$$

This equation means that the MMSE estimator $\hat{s}_{\mathrm{MMSE}}(\tau,\omega)$ is obtained by post-processing the MVDR beamformer's output $y(\tau,\omega)$ with the time-frequency mask $p(\tau,\omega)$ called the Wiener post-filter. In order to design $p(\tau,\omega)$, it is crucial to accurately estimate the target power spectrum $\phi_{ss}(\tau,\omega)$ from the noisy observed signals.

2.4. Property of Inter-channel Cross-spectra of Isotropic Noise

Focusing on the spatial characteristics of diffuse noise, we assume that the inter-channel noise cross-spectrum is determined by the distance between the corresponding microphones [11, 14, 15]:

$$r_{mn} = r_{kl} \;\Rightarrow\; \phi_{v_m v_n}(\tau,\omega) = \phi_{v_k v_l}(\tau,\omega), \tag{13}$$

with r_{mn} denoting the distance between the m-th and n-th microphones.

It can be easily shown that the inter-channel cross-spectra of such noise are real-valued. To show this, note first that the following equation holds from (1):

$$\phi_{v_n v_m}(\tau,\omega) = \phi^*_{v_m v_n}(\tau,\omega). \tag{14}$$

Besides, from (13), we have

$$\phi_{v_n v_m}(\tau,\omega) = \phi_{v_m v_n}(\tau,\omega), \tag{15}$$

because $r_{nm} = r_{mn}$. Therefore,

$$\phi_{v_m v_n}(\tau,\omega) \in \mathbb{R}. \tag{16}$$

The assumption (16) is satisfied by uncorrelated noise assumed by Zelinski [2] and spherically isotropic noise fields assumed by McCowan *et al.* [6], but it is true for a larger class of diffuse noise.

2.5. Post-filter Design Using Imaginary Parts of Cross-Spectra

The assumption (16) implies that the imaginary parts of the inter-channel observation cross-spectra are noise-free. The inter-channel observation cross-spectrum $\phi_{x_m x_n}(\tau,\omega)$ is expressed as follows:

$$\phi_{x_m x_n}(\tau,\omega) = \phi_{ss}(\tau,\omega)d_m(\omega)d_n^*(\omega) + \phi_{v_m v_n}(\tau,\omega) \tag{17}$$

$$= \phi_{ss}(\tau,\omega)e^{-\mathrm{j}\omega(\delta_m - \delta_n)} + \phi_{v_m v_n}(\tau,\omega), \tag{18}$$

which is a consequence of (1) and the uncorrelatedness of the target signal and noise. Because of (16), we can eliminate the noise term $\phi_{v_m v_n}(\tau,\omega)$ in (18) by taking the imaginary parts of both sides:

$$\Im[\phi_{x_m x_n}](\tau,\omega) = -\phi_{ss}(\tau,\omega)\sin[\omega(\delta_m - \delta_n)], \tag{19}$$

where $\Im[\cdot]$ denotes the imaginary part. From this equation, we estimate $\phi_{ss}(\tau,\omega)$ by the least squares method as follows:

$$\hat{\phi}_{ss}(\tau,\omega) = -\frac{\sum_{m<n}\sin[\omega(\delta_m - \delta_n)]\Im[\phi_{x_m x_n}](\tau,\omega)}{\sum_{m<n}\sin^2[\omega(\delta_m - \delta_n)]}. \tag{20}$$

On the other hand, $\phi_{yy}(\tau,\omega)$ in the denominator of $p(\tau,\omega)$ is estimated by Zelinski's estimator [2]

$$\hat{\phi}_{yy}(\tau,\omega) \triangleq \frac{1}{M}\sum_{m=1}^{M}\phi_{x_m x_m}(\tau,\omega). \tag{21}$$

Although it is possible to estimate $\phi_{yy}(\tau,\omega)$ directly using $y(\tau,\omega)$, we use this equation because we observed that it tended to result in better performance in practice.

Consequently, our design of the Wiener post-filter is given by the following equation:

$$\hat{p}(\tau,\omega) \triangleq \frac{\hat{\phi}_{ss}(\tau,\omega)}{\hat{\phi}_{yy}(\tau,\omega)}. \tag{22}$$

Since $p(\tau,\omega)$ is in the range 0 to 1 by definition, we perform the following simple post-processing:

$$\begin{cases} \hat{p}(\tau,\omega) \leftarrow 0, & \text{if } \hat{p}(\tau,\omega) < 0, \\ \hat{p}(\tau,\omega) \leftarrow 1, & \text{if } \hat{p}(\tau,\omega) > 1. \end{cases} \tag{23}$$

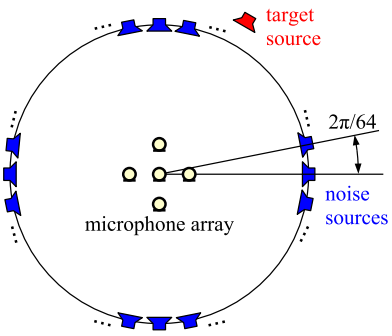


Fig. 1. Configuration of the sources and the microphones.

3. EXPERIMENTS

We conducted two experiments to confirm the effectiveness of the proposed method with simulated and real-world noise. We used the Signal-to-Noise Ratio (SNR) as the criterion for evaluating noise suppression and target distortion. In this paper, it is defined as the energy of the component parallel to the target signal divided by that of the orthogonal component, where the signals are considered N-dimensional vectors (N: the number of samples). The orthogonal decomposition allows an arbitrary overall gain of the system, which is not a serious problem in speech enhancement.

3.1. Experiment Using Simulated Noise

We generated the observed signals in the following way. Fig. 1 illustrates the configuration of the sources and the microphones. We simulated a cocktail-party situation, where the target speech arrived from a known direction and distinct interfering speech signals arrived from 64 equally spaced directions in the horizontal plane. We assumed plane wave propagation in an anechoic environment. The speech data of the target and interfering signals were taken from the ATR Japanese speech database [16]. The array was a cruciform array with 5 microphones, and its diameter was 10 cm unless otherwise stated. The SNR at the central microphone was adjusted to 0 dB. The data length was 4 s, and the sampling frequency was 16 kHz.

The performance of the proposed post-filter was compared to that of Zelinski's post-filter [2]. Each post-filter was preceded by the MVDR beamformer. The frame length and the frame shift for STFT were 512 samples and 16 samples respectively, and the Hamming window was used. We calculated $\boldsymbol{\Phi}_{\boldsymbol{xx}}$ for the beamformer by averaging $\boldsymbol{x}(\tau,\omega)\boldsymbol{x}^{\mathsf{H}}(\tau,\omega)$ temporally over all frames. On the other hand, $\phi_{x_m x_n}(\tau,\omega)$ for the proposed post-filter (20) and Zelinski's post-filter was calculated by averaging $x_m(\tau,\omega)x_n^*(\tau,\omega)$ temporally every 32 frames, where we can reasonably assume signal stationarity.

First, we compare the accuracy of the target power spectrum estimation by the proposed estimator (20) and Zelinski's estimator [2]. In Fig. 2, a scatter plot of the actual value (horizontal axis) and the estimated value (vertical axis) of $\phi_{ss}(\tau,\omega)$ is shown. The actual values were calculated using the clean target signal—which is of course unavailable in practice—by averaging $|s(\tau,\omega)|^2$ temporally every 32 frames. Each point in Fig. 2 corresponds to a time-frequency point. The points corresponding to the proposed method were much more concentrated around the 45° line than those of Zelinski's method, which means that the estimation by the proposed method was much more accurate.

To examine the ability of the proposed method to suppress diffuse noise with a small-sized array, we plotted in Fig. 3 the output SNR as a function of the array diameter. As we can see from the figure, as the diameter decreased, the output SNR of Zelinski's method gradually degraded to that of the beamformer, while the proposed method gave a high SNR even when the diameter was small.

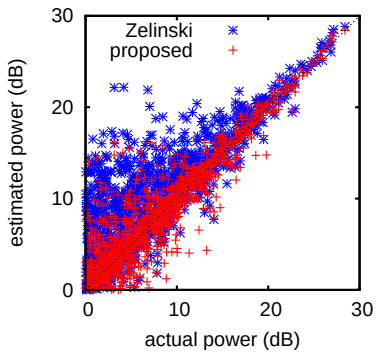


Fig. 2. Scatter plot of the actual value of $\phi_{ss}(\tau,\omega)$ (horizontal axis) and its estimate (vertical axis) for the experiment using simulated noise.

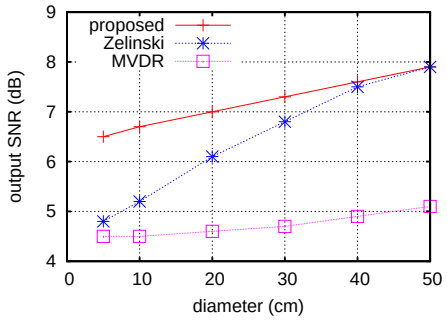


Fig. 3. Output SNR as a function of the array diameter for the experiment using simulated noise.

In Figure 4, an example of spectrograms is shown. Since it is based under the assumption of uncorrelated noise, Zelinski's method did not suppress noise sufficiently at low frequencies, where the noise signals at the microphones are highly correlated. In contrast, the proposed method suppressed noise effectively at all frequencies.

3.2. Experiment Using Real Environmental Noise

We also conducted an experiment using real environmental noise. We fabricated a square array with a diameter of 5 cm and recorded environmental noise at several places (station square, in station, and in train) around a station in Tokyo. We used electret-type microphones (SONY ECM-C10) and a multi-channel input board with microphone amplifiers (Tokyo Electron Device TD-BD-8CSUSB). The target speech was added to the noise recording afterward under the assumption of plane wave propagation. The SNR at the first microphone was adjusted to 0 dB. The data length was 4 s, and the sampling frequency was 16 kHz. The other conditions were the same as those in Section 3.1.

To verify the real-valued model of the inter-channel noise cross-spectra, we plotted in Fig. 5 histograms of the phase of the inter-channel noise cross-spectrum. The noise cross-spectrum was calculated for a pair of adjacent microphones in the array. The noise environment was the station square. Figs. 5 (a) and 5 (b) are the histograms at 2.5 kHz and 5.0 kHz, each of which was made by counting time-frequency slots in each frequency bin. In both figures, the phase concentrated around 0 or $\pm\pi$, showing the validity of the model.

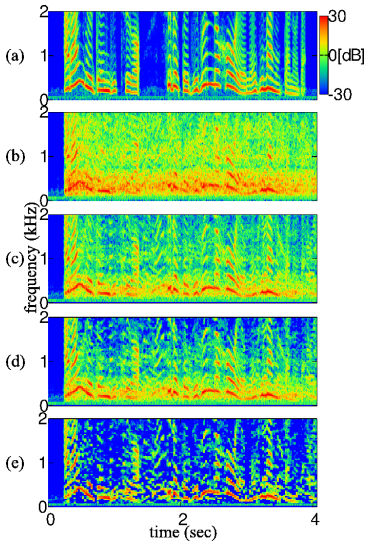


Fig. 4. Performance comparison using spectrograms for the experiment using simulated noise. (a) Target speech; (b) observed signal (SNR: 0 dB); (c) MVDR beamformer (SNR: 6.4 dB); (d) Zelinski's method (SNR: 6.6 dB); (e) proposed method (SNR: 7.8 dB).

Table 1. Output SNR (dB) for real environmental noise.

environment	MVDR	Zelinski	proposed
station square	10.8	11.2	12.4
in station	9.7	10.3	13.2
in train	10.9	11.3	11.5

In Table 1, the output SNRs of the methods for the three environments are shown. The proposed method gave the highest SNR among all methods for all environments.

4. CONCLUSION

This paper described a new design of the Wiener post-filter for diffuse noise suppression. It is based on the estimation of the target power spectrum from the imaginary parts of the inter-channel observation cross-spectra, under the assumption that the inter-channel noise cross-spectra are real-valued. Experiments using simulated and real environmental noise have shown that the proposed method is effective even for a small-sized array.

The future work includes source localization and blind source separation in diffuse noise environments based on the imaginary parts of the inter-channel observation cross-spectra.

This work was supported by Grant-in-Aid for Young Scientists (B) 21760309.

5. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin: Springer-Verlag, 2001.

[2] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. ICASSP '88*, New York, Apr. 1988, pp. 2578–2581.

[3] K. U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," in *Second Cost 229 Workshop on Adaptive Algorithms in Communications*, Bordeaux, Oct. 1992, pp. 185–194.

[4] S. Fischer and K. U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Commun.*, vol. 20, no. 3–4, pp. 215–227, Dec. 1996.

[5] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin: Springer-Verlag, 2001, ch. 3, pp. 39–60.

[6] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[7] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1149–1160, May 2004.

[8] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[9] J. Li and M. Akagi, "A noise reduction system based on hybrid noise estimation technique and post-filtering in arbitrary noise environments," *Speech Commun.*, vol. 48, no. 2, pp. 111–126, Feb. 2006.

[10] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Commun.*, vol. 49, no. 7–8, pp. 657–666, July–Aug. 2007.

[11] N. Ito, N. Ono, and S. Sagayama, "A blind noise decorrelation approach with crystal arrays on designing post-filters for diffuse noise suppression," in *Proc. ICASSP 2008*, Las Vegas, USA, Apr. 2008, pp. 317–320.

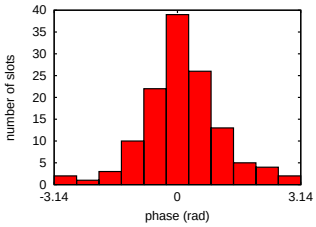
[12] H. L. V. Trees, *Optimum Array Processing*. New York: John Wiley & Sons, 2002.

[13] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and J. M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, Nov. 1955.

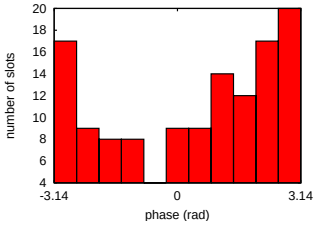
[14] H. Shimizu, N. Ono, K. Matsumoto, and S. Sagayama, "Isotropic noise suppression in the power spectrum domain by symmetric microphone arrays," in *Proc. WASPAA*, New Paltz, NY, Oct. 2007, pp. 54–57.

[15] N. Ono, N. Ito, and S. Sagayama, "Five classes of crystal arrays for blind decorrelation of diffuse noise," in *Proc. SAM*, Darmstadt, Germany, July 2008, pp. 151–154.

[16] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," vol. 9, no. 4, pp. 357–363, Aug. 1990.



(a) 2.5 kHz



(b) 5.0 kHz

Fig. 5. Histograms of the phase of the inter-channel cross-spectrum at (a) 2.5 kHz and (b) 5.0 kHz for real environmental noise recorded in a station square.