ISOTROPIC NOISE SUPPRESSION IN THE POWER SPECTRUM DOMAIN BY SYMMETRIC MICROPHONE ARRAYS

Hikaru Shimizu, Nobutaka Ono, Kyosuke Matsumoto, Shigeki Sagayama

Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
{h-shimizu,onono,matsumoto,sagayama}@hil.t.u-tokyo.ac.jp,

ABSTRACT

In this paper, we propose a new array processing framework for suppressing isotropic noise in the power spectrum domain. As a theoretical basis, we discuss the characteristics of the isotropic noise covariance matrix and show that it can be diagonalized by a definite unitary matrix when the microphone array has a certain symmetry. By the diagonalization, our method gathers noise components to diagonals of the basis-transformed covariance matrix and restores the power spectrum of the target source in a specified direction from non-diagonal components by Maximum Likelihood (ML) method. We performed simulations which show the efficiency of our method for both stationary noise field and nonstationary noise field situations.

1. INTRODUCTION

Noise suppression in speech signals has received a great deal of attention in the acoustic signal processing field because target speech in real environments is often interfered by other speech, various ambient noise, reverberation, and so on. One of the most popular methods is Spectral Subtraction (SS) [1], which estimates the power spectrum of the target speech by subtracting the noise mean power from the observed spectrum. Although it is very simple and effective for stationary noise, it requires a priori knowledge of the noise spectrum, and suffers from artifacts like musical noise for nonstationary noise.

When the noise field has strong directional dependence, array processing methods are effective [2, 3]. Especially, if the noise consists of acoustic waves generated by point sources of which there are less than the number of microphones, the noise vector is constrained to a subspace which dimension is equal to the number of sources, that is, it has a null space with dimension higher than one. Thus, by projecting the observed signal on that null space, the target signal can be separated from noise. It is the principle of the conventional null beamformer. However, noise field in real environments has sometimes diffuse nature rather than the directional dependence, which occurs in crowded space like a cocktail party, stations or airports, or in reverberant environments. Noise subspace then becomes full-rank, and the problem becomes more difficult [4, 5].

In this paper, we describe a new array processing framework for suppressing diffuse noise in the power spectrum domain, which is significant for speech recognition. Due to the nonlinear operation, the array processing in the power spectrum domain has some potential to overcome the limitation of the linear array processing [6, 7]. In our study, we statistically model the diffuse noise as an isotropic field and discuss the characteristics of the noise covariance matrix, which consists of the noise power spectra and cross spectra. Under several assumptions, we show that it has a specific structure depending only on the arrangement of the microphones, and that the covariance matrices corresponding to several symmetry arrangements are orthogonalized by constant unitary matrices, independently of the noise power spectrum. Based on this idea, our method restores the target power spectrum from the non-diagonal components of the observed covariance matrix through basis transformation by the unitary matrix. We give a theoretical overview of the method and present experimental results which show its efficiency.

2. MODELING OF ISOTROPIC NOISE FIELDS

Let $\boldsymbol{O}(\omega)$ be the observed vector in the frequency domain for M microphones. It can written as

$$\boldsymbol{O}(\omega) = F(\omega)\boldsymbol{b}(\omega) + \boldsymbol{N}(\omega), \tag{1}$$

where $F(\omega)$ is a target signal, $\boldsymbol{b}(\omega)$ is a steering vector related to the target direction, and $\boldsymbol{N}(\omega)$ represents noise.

In contrast with the noise field caused by a finite number of noise sources, we here consider an isotropic noise field satisfying:

- the noise power spectrum does not depend on the observation position,
- the noise cross spectrum between two observations does not depend on the angles between their observed points.

For example, observing an isotropic noise field with four microphones arrayed at the vertices of a square and numbered circularly, the covariance matrix can be written as:

$$V_N = \sigma^2 \begin{pmatrix} 1 & \alpha & \beta & \alpha \\ \alpha & 1 & \alpha & \beta \\ \beta & \alpha & 1 & \alpha \\ \alpha & \beta & \alpha & 1 \end{pmatrix}. \tag{2}$$

Since the microphones in the pairs 1-2, 2-3, 3-4 and 4-1 form four sides of a square and are located at the same distance of each other, their cross-spectra are all identical, and denoted as α. In the same way, 1-3 and 2-4 elements of the matrix are also the same, denoted as β. Although V_N, σ, α and β depend on the frequency, ω is omitted in eq. (2) for simplicity.

Giving another interesting example, the regular icosahedron array has twelve microphones and there are three kinds of length between them. When microphone is numbered as shown in Fig. 1, the isotropic noise covariance matrix has the following structure:

$$V_N = \sigma^2 \begin{pmatrix} C(1,\alpha) & C(\beta,\alpha) & C(\alpha,\beta) & C(\gamma,\beta) \\ C(\beta,\alpha) & C(1,\beta) & C(\gamma,\alpha) & C(\alpha,\beta) \\ C(\alpha,\beta) & C(\gamma,\alpha) & C(1,\beta) & C(\beta,\alpha) \\ C(\gamma,\beta) & C(\alpha,\beta) & C(\beta,\alpha) & C(1,\alpha), \end{pmatrix} \tag{3}$$

978-1-4244-1619-6/07/$25.00 ©2007 IEEE

where

$$C(x,y) = \begin{pmatrix} x & y & y \\ y & x & y \\ y & y & x \end{pmatrix}. \quad (4)$$

Note that this structure is determined only by the arrangement of the microphones.

3. SYMMETRY ARRAY TO ORTHOGONALIZE ISOTROPIC NOISE FIELDS

In sensor array processing, suppressing the noise arriving from various surrounding directions by controlling directivity is a difficult task since a large aperture is essential. Thus, to deal with the problem from another aspect, we focus on the power spectrum domain. Assuming that the target signal $F(\omega)$ and the noise $N(\omega)$ in eq. (1) are uncorrelated, the observed covariance matrix $V_O(\omega) = E[O(\omega)O(\omega)^h]$ is obtained by

$$V_O(\omega) = S(\omega)\boldsymbol{b}(\omega)\boldsymbol{b}(\omega)^h + V_N(\omega), \quad (5)$$

where $S(\omega)$ represents the power spectrum of the target signal and the superscript h denotes the Hermitian conjugate. Since $V_N(\omega)$ is Hermitian, it is diagonalized by a unitary matrix $P(\omega)$ as

$$P(\omega)^h V_O(\omega)P(\omega) = S(\omega)(P(\omega)^h\boldsymbol{b}(\omega))(P(\omega)^h\boldsymbol{b}(\omega))^h + D(\omega) \quad (6)$$

where $D(\omega) = P(\omega)^h V_N(\omega)P(\omega)$.

Generally, the determination of a unitary matrix $P(\omega)$ which diagonalizes $V_N(\omega)$ depends on $V_N(\omega)$ itself. It means that we have to observe the noise covariance $V_N(\omega)$ under the active target, which cannot be realized. But when the noise field is isotropic, the covariance matrix has a special structure as shown in eq. (2) or eq. (3), which is determined only by the sensor arrangements. By focusing on this symmetry, we have found that several arrangements have a special covariance matrix which can be diagonalized by a constant unitary matrix, independently of its elements. This fact is summarized in the following theorem.

Theorem 1 *When sensors are disposed at the vertices of a 1) regular polygon, 2) rectangle, 3) regular polyhedron, 4) rectangular solid, or 5) regular polygonal prism. each of the corresponding isotropic noise covariance matrices is diagonalized by a constant unitary matrix independent of the explicit values of its elements.*

Due to space limitation, we cannot describe all of the proof here. But the key point is that a circulant matrix can be diagonalized by the left- and right-multiplication of a DFT matrix Z_n [8] defined by

$$Z_n = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \zeta & \cdots & \zeta^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \zeta^{n-1} & \cdots & \zeta^{(n-1)(n-1)} \end{pmatrix}, \quad (7)$$

$$\zeta = e^{-j2\pi/n}. \quad (8)$$

In the case of regular polygons, circularly numbering microphones yields a circulant noise covariance matrix as shown in eq. (2). Thus, $P = Z_4$ is a unitary matrix to diagonalize eq. (2) independently of the values of α and β. A regular tetrahedron also has a circulant noise covariance matrix.

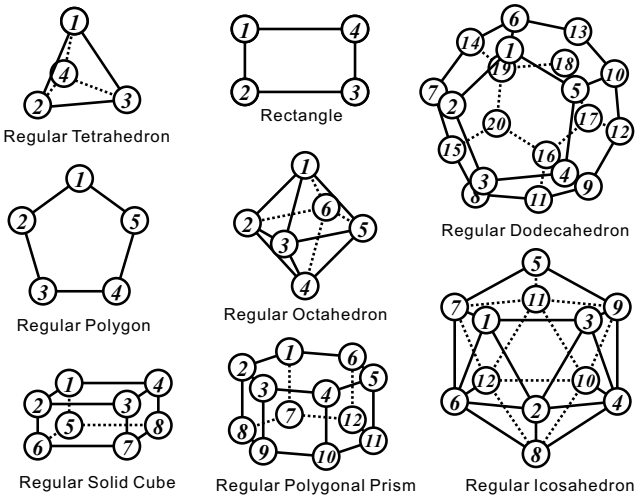

Figure 1: Symmetric arrays to orthogonalize isotropic noise fields. Numbering microphones is an example.

In the case of regular polygonal prisms, a cube, or a regular octahedron, the noise covariance matrices become block-circulant by appropriately numbering as shown in Fig. 1, which are diagonalized by $Z_n \otimes Z_2$, where $\otimes$ represents a direct product.

In other cases, each unitary matrix to diagonalize the noise covariance matrix has a unique form, especially, for a regular dodecahedron and a regular icosahedron arrangement. For eq. (3), we obtain

$$P = \begin{pmatrix} Z_3 & Z_3 & Z_3 & Z_3 \\ Z_3 & -Z_3 & -Z_3R_+ & -Z_3R_- \\ Z_3 & -Z_3 & Z_3R_+ & Z_3R_- \\ Z_3 & Z_3 & -Z_3 & -Z_3 \end{pmatrix}, \quad (9)$$

$$R_+ = \mathrm{diag}\begin{pmatrix} 2+\sqrt{5} \\ \frac{1+\sqrt{5}}{2} \\ \frac{1+\sqrt{5}}{2} \end{pmatrix}, \quad R_- = \mathrm{diag}\begin{pmatrix} 2-\sqrt{5} \\ \frac{1-\sqrt{5}}{2} \\ \frac{1-\sqrt{5}}{2} \end{pmatrix} \quad (10)$$

where diag() represents a diagonal matrix. We can also utilize another form of P with only real-valued elements since P includes conjugate-pair bases.

Although the diagonalization is mathematically interesting, our theorem currently gives only a sufficient condition for the array and the necessity condition remains as an unsolved problem.

4. POWER SPECTRUM ESTIMATION FROM NOISE-FREE CROSS SPECTRUM

Since the isotropic noise in the observation is gathered to the diagonal elements of eq. (6), the next problem is how to estimate the target power spectrum S from the non-diagonal elements, which are ideally noise-free. The non-diagonal (i,j) element of eq. (6) can be written as

$$\Phi_{ij}(\omega) = \tilde{b}_{ij}(\omega)S(\omega) + \varepsilon_{ij}(\omega), \quad (11)$$

where $\Phi_{ij}(\omega)$ and $\tilde{b}_{ij}(\omega)$ are the (i,j) elements of $\tilde{V}_O(\omega) = P^h V_O(\omega)P$ and $(P^h\boldsymbol{b}(\omega))(P^h\boldsymbol{b}(\omega))^h$, respectively, and $\varepsilon_{ij}(\omega)$ represents a modeling error. As the target direction is assumed

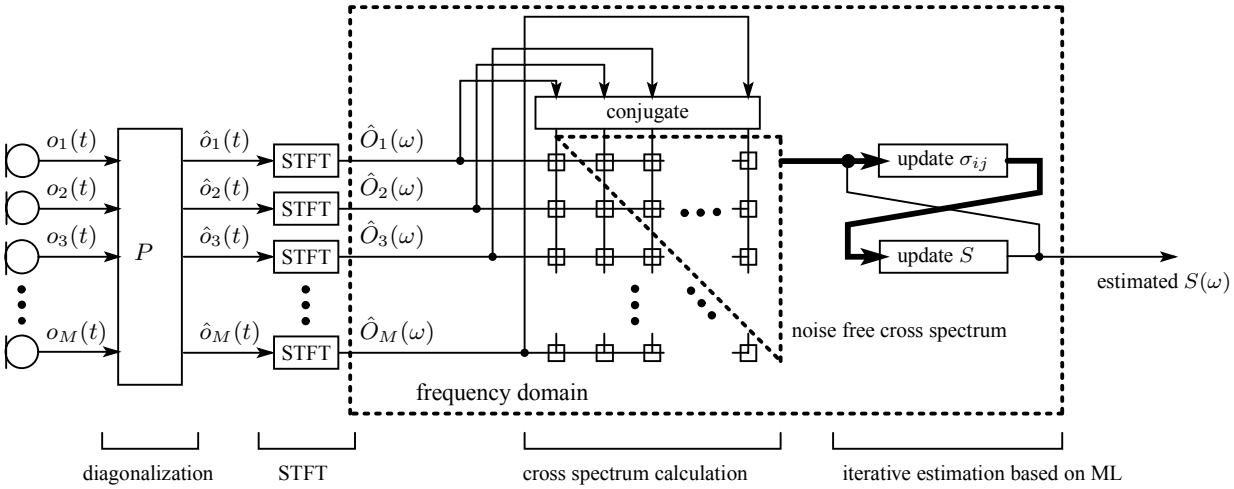

Figure 2: The block diagram of the proposed method

to be known, $\tilde{b}_{ij}(\omega)$ can be represented explicitly. For example, $\tilde{b}_{12}(\omega)$ in a square array is obtained by

$$\tilde{b}_{12}(\omega) = -(j\sin\omega\tau_1 + \sin\omega\tau_2)(\cos\omega\tau_1 + \cos\omega\tau_2), \quad (12)$$

$$\tau_1 = \frac{D\cos\theta}{c}, \quad \tau_2 = \frac{D\sin\theta}{c}, \quad (13)$$

where D is the side length of the square, c is the sound velocity, and θ is the target direction.

If the modeling error $\varepsilon_{ij}(\omega)$ is negligible, $S(\omega)$ is easily obtained from one non-diagonal element by $\hat{S}(\omega) = \Phi_{ij}(\omega)/\tilde{b}_{ij}(\omega)$. But the real environment is not the perfect isotropic field and a finite time observation for calculating expectation also introduces some undesirable errors. So, we need some way to integrate the information from the $M(M-1)/2$ independent non-diagonal elements. One key point is that the variance of $\varepsilon_{ij}(\omega)$ is unknown and should be different for each (i,j) since each $\Phi_{ij}(\omega)$ itself has a different magnitude due to the basis transformation by P. One reasonable way to handle this problem is to apply the Maximum Likelihood (ML) method. Assuming that each $\varepsilon_{ij}(\omega)$ follows a complex Gaussian distribution with zero mean and an unknown variance, the log-likelihood is given by

$$L(\omega) = \sum_{i\neq j}\Big(-\log\sigma_{ij}^2(\omega) - \frac{|\Phi_{ij}(\omega) - \tilde{b}_{ij}(\omega)S(\omega)|^2}{2\sigma_{ij}^2(\omega)}\Big), \quad (14)$$

where some constant term is omitted. By maximizing it in terms of $S(\omega)$ and $\sigma_{ij}(\omega)$ iteratively, we obtain the following update equations:

$$\sigma_{ij}^2(\omega)^{(t+1)} = \frac{1}{2}\left|\Phi_{ij}(\omega) - \tilde{b}_{ij}(\omega)S(\omega)^{(t)}\right|^2$$

$$S(\omega)^{(t+1)} = \frac{\displaystyle\sum_{i\neq j}\frac{\mathrm{Re}[\tilde{b}_{ij}(\omega)^*\Phi_{ij}(\omega)]}{\sigma_{ij}^2(\omega)^{(t+1)}}}{\displaystyle\sum_{i\neq j}\frac{|\tilde{b}_{ij}(\omega)|^2}{\sigma_{ij}^2(\omega)^{(t+1)}}}. \quad (15)$$

5. EXPERIMENTS AND RESULTS

The effectiveness of the proposed method was evaluated based on simulation. First, four microphones were fixed at the vertices of a square of diagonal length 100[mm], and both stationary and nonstationary noise fields were simulated. For the stationary isotropic noise field, 64 white noises of the same power spectrum were added from 64 different directions. Here, the ith white noise was added from an angle of $i\cdot 2\pi/64$[rad]. For the nonstationary isotropic noise field, 50 different speech signals were added from 10 directions. Here, the 5 speech signals of the ith group were added from an angle of $i\cdot 2\pi/10$[rad]. For each noise field, the target signal was added from a given direction. The processing flow is shown in Fig. 2. The 16[kHz] was used as a sampling frequency. In this experiment, we applied the diagonalization by P in the time domain since P no longer depends on ω. The cross spectrum was calculated by the overlap-add method [9]. In our conditions, 32 short-time cross spectra obtained on 8[ms] frames windowed by a Hamming function, with a frame shift of 1[ms] were averaged. Finally, estimation accuracy was evaluated by using the Spectral Distortion measure (SD) [10] given by

$$\mathrm{SD} = 10\sqrt{\frac{1}{C}\sum_{k=1}^{C}(\log_{10}S(\omega_k) - \log_{10}\hat{S}(\omega_k))^2}, \quad (16)$$

where C denotes the total number of sample points, $S(\omega_k)$ denotes the power spectrum in the noiseless case, and $\hat{S}(\omega_k)$ denotes the estimated power spectrum from the noise-mixed observed signal. For the adequateness of evaluation, the power spectrum was also estimated by other conventional methods, namely the Delay-and-Sum (DS) and Spectral Subtraction (SS) methods, using the same four microphones. The DS method calculated the spectrum based on the basic phase compensation, and the SS method was given the mean noise power over the observation time.

Fig. 3 and Fig. 4 show SD changes of the power spectrum under stationary and nonstationary noise field respectively, where the vertical axis represents the input SNR. Fig. 5 shows the estimated power spectra in a certain set of 32 frames. For the stationary noise field condition shown in Fig. 3, the proposed method is superior to other methods in the more than 0dB, even comparing with the spectral subtraction to which the mean noise power was given as a priori knowledge. Furthermore, for the nonstationary noise field condition, spectral subtraction method achieves far worse performance because of the misestimation of noise mean power, while

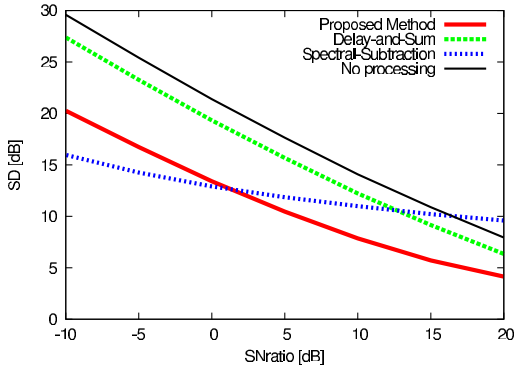

Figure 3: The SD after applying several methods under the stationary isotropic noise case

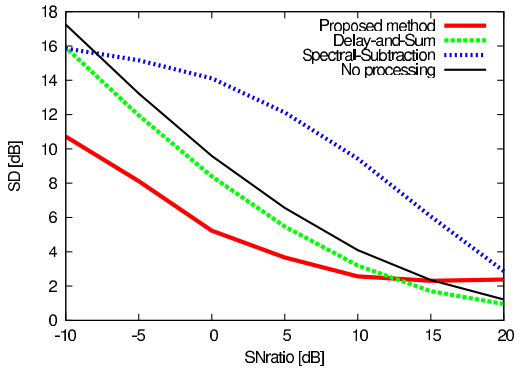

Figure 4: The SD after applying several methods under the nonstationary isotropic noise case

the proposed method is still effective for SNRs of the observed signal lower than about 12.5[dB]. In Fig. 5, only the proposed method restores the precise structure in the power spectrum of the target signal, which would have a beneficial effect on speech recognition rate.

From these experimental results, we conclude that the proposed method is effective not only under stationary sound field, but also under nonstationary sound field, which is thought to be a more realistic environment.

6. CONCLUSION

Our contribution in this paper is summarized as follows.

1. We showed the following sensor arrangements as sufficient conditions for the isotropic noise covariance matrix to be diagonalized by a constant unitary matrix: 1) regular polygon, 2) rectangle, 3) regular polyhedron, 4) rectangular solid, or 5) regular polygonal prism.
2. We proposed a new power spectrum estimation method using noise-free cross spectrum derived from the diagonalization of the covariance matrices.
3. Our method works well with comparison to other conventional methods in isotropic noise environment as shown through simulation.

Although we used white noise as stationary noise field and various ambient speech noises as nonstationary noise field in the simulations, we have not yet simulated a reverberant noise field, which

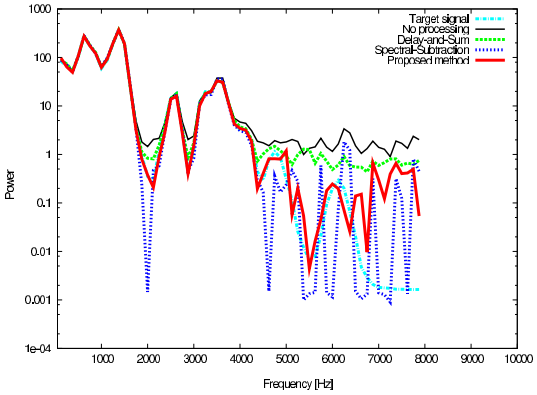

Figure 5: The comparison of estimated power spectra by several methods

often degrades the performance of speech recognition in real environments. Thus, future work includes simulations in a reverberant noise field, and of course, experiments in real environments.

7. REFERENCES

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoustic, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] D. Johnson and D. Dudgeon, *Array Signal Processing*. Prentice Hall, 1993.

[3] M. Brandstein, D. Ward, *Microphone Arrays*. Springer, 2001.

[4] S. E. Nordholm and Y. H. Leung, "Performance Limits of the Broadband Generalized Sidelobe Cancelling Structure in an Isotropic Noise Field," *J. Acoust. Soc. Jpn*, vol. 107, no. 2, Feb. 2000.

[5] I. A. McCowan and H. Bourlard, "Microphone Array Post-Filter Based on Noise Field Coherence," *IEEE Trans. Acoustic, Speech, and Audio Processing*, vol. 11, no. 6, Nov. 2003.

[6] H. Saruwatari, S. Kajita, and K. Takeda, F. Itakura, "Speech Enhancement Using Nonlinear Microphone Array Based on Complementary Beamforming," *IEICE Trans. Fundamentals*, vol. E82-A, no. 8, pp. 1501–1510, Aug. 1999.

[7] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual cross-talk components in blind source separation using time-delayed spectral subtraction," in *Proc. ICASSP*, vol. II, May 2002, pp. 1789–1792.

[8] G. Golub and C. V. Loan, *Matrix Computations*. Johns Hopkins University Press, 1996.

[9] C. K. Yuen, "A Comparison of Five Methods for Computing the Power Spectrum of a Random Process Using Data Segmentation," in *Proc. IEEE*, vol. 65, no. 6, Jun. 1977, pp. 984–986.

[10] F. Nordén and T. Eriksson, "A Speech Spectrum Distortion Measure with Interframe Memory," *IEEE Trans. Acoustic, Speech, and Signal Processing*, vol. 2, pp. 712–720, Jul. 2001.