DOA ESTIMATION OF SPEECH SIGNAL WITH A SMALL NUMBER OF MICROPHONE ARRAY IN REAL ACOUSTICAL ENVIRONMENT

Yusuke HIOKA[†] *and Nozomu HAMADA*[‡]

[†‡] Signal Processing Lab., School of Integrated Design Engineering, Keio University
Hiyoshi 3-14-1, Kohoku-ku, Yokohama, Kanagawa, 223-8522 Japan
[†]hioka@hamada.sd.keio.ac.jp [‡]hamada@sd.keio.ac.jp

ABSTRACT

In this paper, we propose a DOA(Direction of Arrival) estimation of speech signal under reverberant influence. We previously proposed a method of estimating DOA for speech by exploiting the harmonic structure existing in voiced sound [1]. However this method is not always robust to reverberations that usually exist in practical indoor environments. Our new proposal aims to suppress the reverberation effects by applying an idea of the spatial smoothing technique to our conventional method. To prove the efficiency of the method, we show some results of the experiments in two conference rooms.

1. INTRODUCTION

As a core technology in speech human-machine interfaces, speech recognition system requires to receive speech signal as enhanced as possible. Improving the quality of received speech signal using a microphone array, DOA of target speech is indispensable information. Among several methods for the speech DOA estimation subject[2][3], MUSIC (MUltiple SIgnal Classification)[4] with Coherent Signal Subspace (CSS)[2] is known as an effective method with high spatial resolution. However, it requires rough DOA estimation *a priori*, and this pre-estimation accuracy usually affects the final estimation result significantly. From a practical point of view, the array scale is another subject to be considered. Generally, the performance of an array processing for estimating DOA, as well as for rejecting interferences, is improved by increasing both the number of sensors and the array aperture size. However, they are often restricted due to the limited physical size of the apparatus on which the array is equipped.

For these subjects mentioned above, we previously proposed a DOA estimation method for speech using only two microphones without a pre-estimation process[1]. The estimation accuracy of the method degrades in a real acoustic environment due to reverberation. The spatial smoothing [5] is a well-known measure to reduce the influence of the reverberation at the DOA estimation. It usually requires many numbers of sensors to form the subarrays. In this paper, we adopt not only spatial but also spectral smoothing process, based on the spatial smoothing technique, to the frequency array data introduced in our previous method[1].

This paper is organized as follows. In the Sec.2, we settle our problem and the proposed method is described. Experimental results are shown in Sec.3 and some concluding remarks are stated in Sec.4.

2. PROPOSED METHOD

2.1. Frequency array data in a reverberant condition

Fig.1 shows the flow diagram of the proposed method. At first, we derive the frequency array data[1] of the received signal. Usually, the impulse response in a room consists of the following three components, direct sound, early reflection and reverberation. The early reflection mainly originates from the first order reflection, has directionality and is highly correlated with the direct signal. In contrast, the reverberation is the compound of higher order reflections that are sufficiently attenuated and distorted; therefore, it can be conceived as spatially white noise. As illustrated in Fig.2, we use one speaker and one early reflection model. Thus, the q-th and $q+1$-th input signals in the Fourier domain are given by

$$\begin{aligned} X_q(\omega) &= S(\omega)e^{-j\omega(q-1)\tau} \\ &\quad +\alpha S(\omega)e^{-j\omega(q-1)\tau'}e^{-j\omega t} + N_q(\omega) \quad (1) \\ X_{q+1}(\omega) &= S(\omega)e^{-j\omega q\tau} \\ &\quad +\alpha S(\omega)e^{-j\omega q\tau'}e^{-j\omega t} + N_{q+1}(\omega), \quad (2) \end{aligned}$$

where τ and τ' are the time delay differences of two sensors relating to the direction θ and θ' respectively, defined as $\tau(\theta) = \frac{d\sin\theta}{v}$, $\alpha, t \in \mathcal{R}$ are the attenuation rate and arrival delay of early reflection to the direct sound, and $N_q(n)$ is Fourier Transform of the spatially white noise at q-th sensor including the reverberation and sensor noise. The phase reference is taken at the 1st sensor. The frequency array data is derived by extracting the harmonic elements of the cross

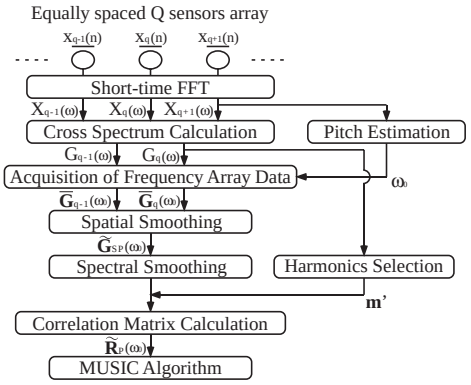

Fig. 1. Proposed Method

spectrum $G_q(\omega)$.

$$\begin{aligned}
&\overline{\mathbf{G}}_q(\omega_0) \\
&= \left[\, G_q(a\omega_0)\, G_q(b\omega_0)\, \cdots \right]^T && (3)\\
&= \mathbf{P}(\omega_0)(\mathbf{d} + \alpha^2 \mathbf{d}' + \alpha \mathbf{z}), && (4)
\end{aligned}$$

where,

$$\begin{aligned}
G_q(\omega) &= E\left[X_q(\omega)^* X_{q+1}(\omega)\right] \\
&= P(\omega)\left\{ e^{-j\omega\tau} + \alpha^2 e^{-j\omega\tau'} \right. \\
&\quad + \alpha e^{-j\omega t} e^{-j\omega\{q\tau' - (q-1)\tau\}} \\
&\quad \left. + \alpha e^{j\omega t} e^{-j\omega\{q\tau - (q-1)\tau'\}} \right\} \\
P(\omega) &= E\left[|S(\omega)|^2\right] \\
\mathbf{P}(\omega) &= \mathrm{diag}\left(\left[\, P(a\omega) \quad P(b\omega) \quad \cdots \right]\right) \\
\mathbf{d} &= \left[\, e^{-j a\omega_0 \tau} \quad e^{-j b\omega_0 \tau} \quad \cdots \,\right]^T \\
\mathbf{d}' &= \left[\, e^{-j a\omega_0 \tau'} \quad e^{-j b\omega_0 \tau'} \quad \cdots \,\right]^T \\
\mathbf{z} &= \left[\, z_a \quad z_b \quad \cdots \,\right]^T \\
z_i &= e^{-j\omega_0 i t} e^{-j\omega_0 i\{q\tau' - (q-1)\tau\}} \\
&\quad + e^{j\omega_0 i t} e^{-j\omega_0 i\{q\tau - (q-1)\tau'\}} \\
&\quad (i = a, b, \cdots \in \mathbf{m})
\end{aligned}$$

ω_0 is the harmonic frequency estimated as shown in [1], and $\mathbf{m}$ is the set of selected $\hat{M}$ harmonics' orders.

As compared to the ideal anechoic case treated in [1], the frequency array data in Eq.(4) contains both the auto and cross correlation terms of the reflecting sound. To suppress these terms, we introduce the spatial and spectral smoothing process.

2.2. Spatial smoothing process

The spatial smoothing method[5] brings suppressing effect to the cross-correlation terms by taking the average of the subarray covariance matrices. Because the first lower diagonal elements are equivalent to $G_q(\omega)$, we take the average of frequency array data to reduce the cross-correlation term as given by Eq.(5).

$$\begin{aligned}
\widetilde{\mathbf{G}}_{\mathrm{SP}}(\omega_0) &= \sum_{q=1}^{Q-1} \delta_q^{\mathrm{SP}} \overline{\mathbf{G}}_q(\omega_0) && (5)\\
&= \mathbf{P}(\omega_0)(\mathbf{d} + \alpha^2 \mathbf{d}' + \alpha \Xi_{\mathrm{SP}} \mathbf{z}), && (6)\\
\Xi &= \mathrm{diag}\left(\left[\, \xi_{\mathrm{SP}}(a\omega_0) \;\; \xi_{\mathrm{SP}}(b\omega_0) \;\; \cdots \right]\right),
\end{aligned}$$

where $\xi_{\mathrm{SP}}(\omega)$ indicates the suppression rate at frequency ω determined by the weight δ_q^{SP}.

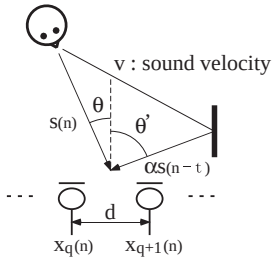

Fig. 2. Modeling of input signal (One direct sound and early reflection)

2.3. Spectral smoothing process

Although the spatial smoothing process reduces the effect of reflection and reverberation, the reduction is still insufficient because we desire to use only a few microphones. The high cross correlation of these components degrades the estimation accuracy as well in the following calculation process of covariance matrix $\widetilde{\mathbf{G}}_{\mathrm{SP}}$. Thus to suppress these cross correlation terms, we introduce the spectral smoothing to the covariance calculation.

As we can find in Fig.3(a), the power spectrum of voiced sound is locally distributed around the harmonic frequencies $m\omega_0 \;\; (m = 1, 2, \cdots)$. Because the DFT analyzes a signal with equal frequency resolution $\triangle\omega$, each DFT bin is conceived as one of a number of equally spaced sensors for the frequency array data. Following this idea, we propose a spectral smoothing process that takes the mean covariance matrices $\widehat{\widetilde{\mathbf{R}}}$ generated from the frequency subarray as illustrated in Fig.3(b).

$$\begin{aligned}
\widetilde{\mathbf{R}} &= \sum_{r=-R}^{R} \delta_r^{\mathrm{FR}} \widetilde{\mathbf{G}}_{\mathrm{SP}}^{(r)}(\omega_0) \widetilde{\mathbf{G}}_{\mathrm{SP}}^{(r)H}(\omega_0) \\
&= \mathbf{P}^2(\omega_0)\{\mathbf{d}\mathbf{d}^H + \alpha^4 \mathbf{d}'\mathbf{d}'^H \\
&\quad + \alpha^2 \Xi_{\mathrm{SP}}^2 \mathbf{z}\mathbf{z}'^H + \Xi_{\mathrm{FR}} Z_{\mathrm{FR}}\}, && (7)
\end{aligned}$$

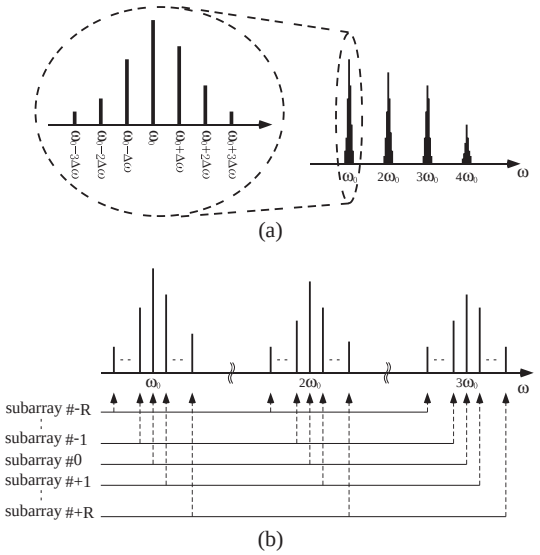

Fig. 3. Spectral Smoothing Process(a)Spectrum distribution of voiced sound (b)Frequency subarray

where,

$$\widetilde{\mathbf{G}}_{\text{SP}}^{(r)}(\omega) \equiv \begin{bmatrix} \sum_{q=1}^{Q-1} \delta_q^{\text{SP}} G_q(a\omega + r\triangle\omega) \\ \sum_{q=1}^{Q-1} \delta_q^{\text{SP}} G_q(b\omega + r\triangle\omega) \\ \vdots \end{bmatrix}$$

$$Z_{\text{FR}} = \left\{\alpha^2(\mathbf{d}\mathbf{d}'^H + \mathbf{d}'\mathbf{d}^H) + \alpha\Xi_{\text{SP}}(\mathbf{d}\mathbf{z}^H + \mathbf{z}\mathbf{d}^H) + \alpha^3\Xi_{\text{SP}}(\mathbf{d}'\mathbf{z}^H + \mathbf{z}\mathbf{d}'^H)\right\},$$

Now we find that the reflection and reverberation terms are highly suppressed in Eq.(7).

2.4. DOA estimation based on MUSIC

Finally, for the DOA estimation, the MUSIC method[4] is applied to the normalized covariance matrix $\widehat{\mathbf{R}}_{\text{NOR}}$ because the power levels of harmonics are different in the frequency array data. The procedure is summarized as follows.

$$[\widetilde{\mathbf{R}}_{\text{NOR}}]_{i,j} = [\widetilde{\mathbf{R}}]_{i,j}/|[\widetilde{\mathbf{R}}]_{i,j}| \quad (8)$$

$$P_{\text{MUSIC}}(\phi) = \frac{1}{\mathbf{e}(\phi)^H \sum_{m=2}^{\widehat{M}} \mathbf{v}_m \mathbf{v}_m^H \mathbf{e}(\phi)} \quad (9)$$

$$\bar{\theta} = \arg\max_{\phi} \left[P_{\text{MUSIC}}(\phi)\right] \quad (10)$$

$$\text{subject to} \quad \mathbf{e}(\phi) = \left[e^{-ja\omega_0\tau(\phi)} \quad e^{-jb\omega_0\tau(\phi)} \quad \dots\right]^T$$

3. EXPERIMENTAL VERIFICATION

3.1. Experiment Condition

We performed experiments in two conference rooms, whose geometrical parameters are shown in Fig.4. For the comparing conventional methods, we also applied our previous method[1], beamforming method[4] and the MUSIC with CSS on the harmonics to the same data. We used the voiced sounds (/a/,/e/,/i/,/o/,/u/) of 10 testees (5 each for male and female) as a source signal and we made 5 trials for each set of data. In Tab.1, we denote the parameters in the proposed method except for the values that follow the same condition appeared in [1].

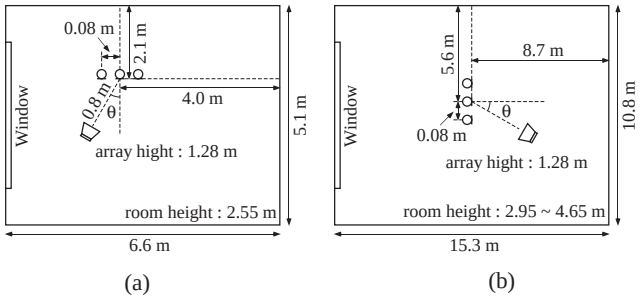

Fig. 4. Geometrical condition for experiment(a)room A : small (b)room B : large

Table 1. Parameters for experiment

Sampling Frequency	16[kHz]
Q	3
R	3
δ_q^{SP}	$\frac{1}{Q}$
δ_r^{FR}	$\frac{1}{2R+1}$

3.2. Experiment Results

3.2.1. Adoption Example

As an example of the estimation, we compare the results of the previous[1] and the proposed methods for a vowel /o/ spoken by a female from the direction of $\theta = 30°$. Fig.5(a) and (b) show the phase of the frequency array data $\overline{\mathbf{G}}_q(\omega_0)$ and $\widetilde{\mathbf{G}}_{\text{FR}}(\omega_0)$ defined by Eq.(11), respectively.

$$\widetilde{\mathbf{G}}_{\text{FR}}(\omega_0) = \sum_{r=-R}^{R} \delta_r^{\text{FR}} \widetilde{\mathbf{G}}_{\text{SP}}^{(r)}(\omega_0) \quad (11)$$

The dotted line shows the phase theoretically derived by the DOA information. This result shows that the spatial and spectral smoothers improve the estimation accuracy of the phase value.

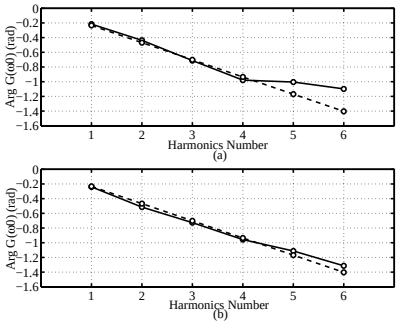

Fig. 5. Phase of the frequency array data (a)Previous (b)Proposed

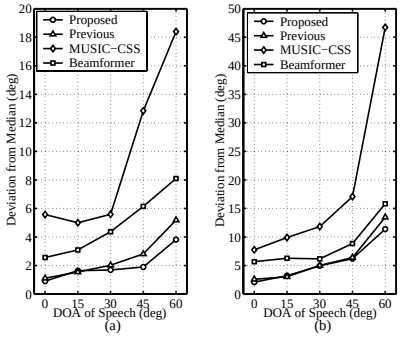

Fig. 6. Deviation from Median in two different rooms (a)room A (b)room B

3.2.2. Quantitative Evaluation

As a qualitative evaluation, we introduce the following criterion called "Deviation from Median(DM) σ," defined by

$$\sigma = \sqrt{\frac{1}{NI}\sum_{i=1}^{I}\sum_{j=1}^{N}(\bar{\theta}_{i,j}-\mu)^2}, \qquad (12)$$

where,

$$\mu = \mathbf{Med}_i\left[\frac{1}{N}\sum_{j=1}^{N}\bar{\theta}_{i,j}\right], \qquad (13)$$

$\mathbf{Med}_i$ means the median for i, $\bar{\theta}_{i,j}$ is the estimation result of a data i at j-th trial, and N and I are the number of trials and data sets, respectively.

In the Fig.6, we show the estimation results for two different rooms. We can find that the proposed method gives best performance. With respect to the room size, the result at the larger room is better than that of at the small room. This is because the attenuation rate α of the early reflection signal increases to 1 in the small room. Thus, the result shows that the proposed method is still obstructed by the early reflection.

4. CONCLUDING REMARKS

In this paper, we propose a method to improve our previous DOA estimation method in real acoustical environments. To reduce the influence of reverberating sound, we proposed a procedure based on spatial smoothing. From some experimental results, we can confirm the improvement brought about by this method. In future consideration, further improvement should be considered at the condition where a strong early reflection exists, such as in a small room or at wall-side.

5. ACKNOWLEDEMENT

This work is supported in part by a Grant in Aid for the 21st century Center Of Excellence for Optical and Electronic Device Technology for Access Network from the Ministry of Education, Culture, Sport, Science, and Technology in Japan.

6. REFERENCES

[1] Y. Hioka, Y. Koizumi and N. Hamada, "Improvement of DOA Estimation Using Virtually Generated Multichannel Data from Two-Channel Microphone Array", *Journal of Signal Processing*, Vol. 7, No. 1, pp.105–109, 2003.

[2] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol.33, No.4, pp.823-831, 1985.

[3] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location, *IEEE Trans. on Speech and Audio Processing*, Vol.5, No.3, pp.288-292, 1997.

[4] D.H. Johnson and D.E. Dedgeon, "Array Signal Processing," PTRP Prentice Hall, 1993.

[5] T.J Shan, M. Wax and T.Kailath, "On Spatial Smoothing for Direction-of-Arrival Estimation of Coherent Signals," IEEE Trans. ASSP, Vol.33, No.4, pp.806–811, 1985.

[6] T. Kikuchi, T. Yamaoka, and N. Hamada, "Microphone Array System with DOA Estimation by using Harmonic Structure of Speech Signals," IEICE Technical Report, DSP98-164, pp.23–28, January 1999.(in Japanese)