SEPARATE ESTIMATION OF AZIMUTH AND ELEVATION DOA USING MICROPHONES LOCATED AT APICES OF REGULAR TETRAHEDRON

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

ABSTRACT

In this paper, we propose a DOA (Direction Of Arrival) estimation method of speech signal for both azimuth and elevation directions. Our previous DOA estimation method achieves high precision and uniform spatial resolution by integrating the *frequency array data*[1] generated from microphone pairs in an equilateral-triangular microphone array[2]. In the method presented here, we extend the method using four microphones located at the apices of regular tetrahedron to enable the method to estimate the elevation angle from the array plane as well. Furthermore, we introduce an idea for separate estimation of azimuth and elevation to reduce the computational loads.

1. INTRODUCTION

As a core technology in speech human-machine interfaces, speech recognition requires the received speech signal to be of sufficiently high quality. To improve the quality of received speech signal using microphone array, DOA of target speech is indispensable information. Among several methods for the speech DOA estimation [3]–[5], MUSIC(MUltiple SIgnal Classification)[6] with CSS(Coherent Signal Subspace)[4] is known as an effective method with high spatial resolution. However, it requires rough DOA pre-estimation, and the final estimation result is highly sensitive to the accuracy of the pre-estimation. Additionally, array scale is another subject to be considered from the practical point of view. Generally, the performance of an array processing for estimating DOA, as well as rejecting interferences, is improved by increasing both the number of sensors and the array aperture size. However, they are often restricted in practical use due to the limited physical size of the apparatus on which the array is equipped.

For taking account of these subjects mentioned above, we previously proposed a DOA estimation method for speech signal using only a pair of microphones. In the method, we generate the *frequency array data*[1] by extracting the harmonic components in a voiced speech signal to achieve highly accurate estimation without pre-estimation process. Added to this, we introduce the *rotation matrices* in the study [2] to integrate three *frequency array data*, generated from three microphones located at the vertices of equilateral triangle, for realizing uniform azimuth resolution. In this paper, we aim at estimating both azimuth and elevation using these methods. The main proposals in this research are summarized as the following two ideas.

- Utilize four microphones located at the apices of regular tetrahedron
- Separate estimation procedure of both azimuth and elevation

The former idea aims at realizing the discriminability about the elevation angle. In addition to the three pairs of microphones consisting the planar equilateral triangular configuration, we can extract another three pairs of microphones located vertically to the planar triangular array. Using these additional pairs, we can measure elevation angle around the planar triangular array because each microphone pair has the spatial discriminability along the array axis and the best spatial resolution is obtained at its broadside direction[2]. The second idea is introduced to reduce the calculation load. In the methods [1] and [2], as well as in many other DOA estimation methods[6], we evaluate the performance function at all possible directions and search its maximum. For the problem considered here, we have to search the solution within two-dimensional parameter plane. We propose a new strategy to estimate both azimuth and elevation DOAs by separating the six pairs into two groups in respect of their geometrical advantages in DOA estimation.

This paper is organized as follows. In the following Sec.2, we briefly review the *frequency array data*[1] to explain the problem settings, and the details of the proposed method including *rotation matrices* are described in Sec.3. Simulation and experimental results are shown in Sec.4, and some concluding remarks are stated in Sec.5.

2. FREQUENCY ARRAY DATA FROM A PAIR OF MICROPHONES

Let us consider the two channel signals $\{x_1(n), x_2(n)\}$ in Fig.1, obtained by a pair of microphones, represented by

$$x_1(n) = s(n) + n_1(n) \quad (1)$$
$$x_2(n) = s(n-\tau) + n_2(n), \quad (2)$$

where $s(n)$ is a voiced speech signal, τ is the time delay between two microphones which is a function of the source signal's DOA θ, and $n_1(n)$ and $n_2(n)$ are mutually uncorrelated white noise signals. Thus, the Fourier transforms of $x_1(n)$ and $x_2(n)$, and their cross spectrum, are represented by

$$X_1(\omega) = S(\omega) + N_1(\omega) \quad (3)$$
$$X_2(\omega) = S(\omega)e^{-j\omega\tau} + N_2(\omega) \quad (4)$$

and

$$G_{12}(\omega) = E\left[X_1^*(\omega)X_2(\omega)\right] = P_S(\omega)e^{-j\omega\tau} \quad (5)$$

respectively, where $P_S(\omega)$ and the expectation $E[\cdot]$ denote the power spectral density of $s(n)$ and the average of DFT at several frames respectively, and $*$ means the complex conjugate. When

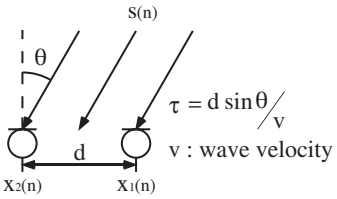

Fig. 1. Microphone pair model to derive a *frequency array data*

we set $\omega = \omega_m$, where ω_m is the m-th higher harmonics of the fundamental frequency ω_0 of $s(n)$, i.e.

$$\omega_m = m\omega_0, \tag{6}$$

the phase term in $G_{12}(\omega_m)$ is replaced by $e^{-j\omega_0 m\tau}$. This phase term is interpreted as a time delay, m times τ, of a narrow-band signal whose central frequency is ω_0. This interpretation leads us to the idea that the $G_{12}(\omega_m)$ might be the virtual multichannel array signals, which are narrow-band signals acquired by an equally spaced linear multiple sensors. For determining the frequency ω_0 and its harmonics, we use the harmonic structure of voiced sound $s(n)$. That is, we set ω_0 as the fundamental frequency of voiced sound in speech. Because the power of a voiced sound is localized in its harmonic frequencies, the SNRs at these frequencies are rather high, and as a result, harmonic elements contribute to improving the estimation accuracy. Thus, we define the following *frequency array data* $\mathbf{G}(\omega_0)$ for a pair of microphone signals.

$$\mathbf{G}(\omega_0) = \left[\frac{G_{12}(\alpha\omega_0)}{|G_{12}(\alpha\omega_0)|} \quad \frac{G_{12}(\beta\omega_0)}{|G_{12}(\beta\omega_0)|} \quad \cdots \right]^T \tag{7}$$
$$(\alpha, \beta, \cdots \in \mathbf{m})$$

Since the power spectrum distribution depends on speaker and phoneme, here we select the $\hat{M}$ harmonics that contains the voiced speech components in higher SNR condition. In Eq.(7), $\mathbf{m}$ is a set of the $\hat{M}$ harmonics order selected by thresholding the magnitude-squared coherence function [2][7]. The fundamental frequency ω_0 is estimated by evaluating logarithmic harmonic product spectrum [2][8].

3. PROPOSED METHOD

3.1. Problem settings

In the proposed method, we receive the target signal by 4 microphones located at the apices of regular tetrahedron as shown in Fig.2. A speaker in the direction {azimuth,elevation} $= \{\theta, \psi\}$ utters a voiced speech signal $s(n)$, and the microphones receive the signal $a(n)$, $b(n)$, $c(n)$ and $h(n)$ respectively, with additive sensor noise signals $n_a(n)$, $n_b(n)$, $n_c(n)$ and $n_h(n)$ that can be modeled as spatially uncorrelated. From such configuration, we have six pairs of microphones whose distances between microphones are equal and each of them derives the *frequency array data*[1]. Here we assume for the input signal that only one voiced speech signal is received as well as in the previous works[1][2].

3.2. Separate DOA estimation using classified microphone pairs

A linear array, including a microphone pair as the simplest case, has spatial discriminability at the direction along with the aperture

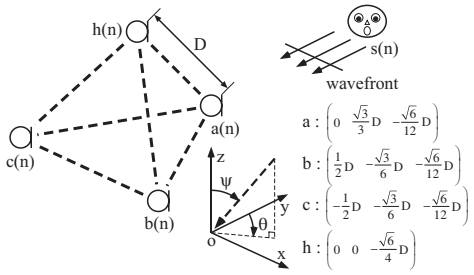

Fig. 2. Model of input signal

line, and the highest resolution of it is obtained to facing (broadside) direction. Now in the case of the regular tetrahedral arrangement, we can separate the six microphone pairs into two groups from such spatial resolution point of view as shown in Fig.3. The first group, we call *Triangle*, consists of three pairs parallel to the x-y plane, i.e. $[ab]$, $[bc]$ and $[ca]$, where $[ij]$ means the pair of microphones i and j. The pairs in *Triangle* form the equilateral-triangular array used in [2], therefore they have good discriminability to the azimuth direction. The second group, we call *Radiation*, is composed of the rest of the pairs, i.e. $[ha]$, $[hb]$ and $[hc]$. Because their apertures are nearly vertical to the x-y plane, they should cover elevation angle estimation. We make proper use of these two groups that *Triangle* for azimuth estimation and *Radiation* for elevation estimation. The respective estimation processes are performed separately as shown in Fig.4.

3.3. Initial azimuth estimation using *Triangle* -Step 1-

Although the previous method assumes the speaker's position on the x-y plane, the effect of deviation from the plane to the azimuth estimation result is small as far as the deviation is within a few tens degrees[2]. From this fact, we preliminarily estimate the azimuth using *Triangle* by [2]. In this method, we integrate the frequency array data generated from three microphone pairs in *Triangle* by the use of *rotation matrices*. Then we estimate the azimuth DOA by analyzing the subspace structure of the *integrated frequency array data*. This initial azimuth angle estimate, we denote $\overline{\theta'}$, is used in the next elevation estimation process.

3.4. Elevation estimation using *Radiation* -Step 2-

Now let us consider the difference of delay terms (which determine the phase values) between two frequency array data for a signal propagating from direction (ϕ, γ).

$$\tau_{hb2a}(\phi,\gamma) \equiv \tau_{ha} - \tau_{hb} = D\sin(\phi - \tfrac{\pi}{3})\sin\gamma/v \tag{8}$$
$$\tau_{hc2a}(\phi,\gamma) \equiv \tau_{ha} - \tau_{hc} = D\sin(\phi - \tfrac{2\pi}{3})\sin\gamma/v, \tag{9}$$

where τ_{ij} means the time delay difference between microphones i and j. Then we define the following *rotation matrices*,

$$\mathbf{G}_{hb2a}(\phi,\gamma) \equiv \mathrm{diag}\left[e^{-j\alpha\omega_0\tau_{hb2a}} \quad e^{-j\beta\omega_0\tau_{hb2a}} \quad \cdots \right] \tag{10}$$
$$\mathbf{G}_{hc2a}(\phi,\gamma) \equiv \mathrm{diag}\left[e^{-j\alpha\omega_0\tau_{hc2a}} \quad e^{-j\beta\omega_0\tau_{hc2a}} \quad \cdots \right], \tag{11}$$

to generate the following *integrated frequency array data*.

$$\mathbf{G}_R(\theta,\psi,\phi,\gamma) \equiv \{\mathbf{G}_{ha} + \mathbf{G}_{hb2a}\mathbf{G}_{hb} + \mathbf{G}_{hc2a}\mathbf{G}_{hc}\}/3 \tag{12}$$

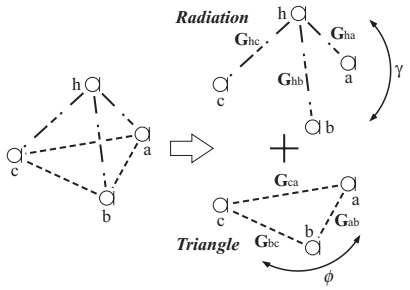

Fig. 3. Separate DOA estimation using different set of microphone pairs

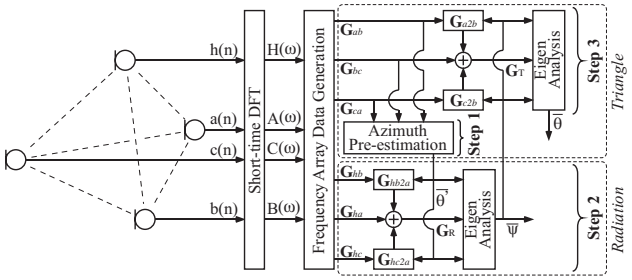

Fig. 4. Data flow diagram of the proposed method

It is noted that the phases of each term in the right side of Eq.(12) are equal if and only if $\phi = \theta$ and $\gamma = \psi$. From this fact, our problem results in searching ϕ and γ that satisfies $\mathbf{G}_R = \mathrm{s}_{ha}$, where s_{ha} is the steering vector that contains the time delay between microphones h and a defined by

$$\mathrm{s}_{ha}(\phi,\gamma) = \left[\ e^{-j\alpha\omega_0\tau_{ha}(\phi,\gamma)} \quad e^{-j\beta\omega_0\tau_{ha}(\phi,\gamma)} \quad \cdots\ \right]^T. \quad (13)$$

We analyze the subspace structure of $\mathbf{G}_R$ to solve this problem, namely we perform eigendecomposition to the covariance matrix $\mathbf{R}_R = \mathbf{G}_R\mathbf{G}_R^H$, and find the angle by the following maximum search.

$$\overline{\psi} = \arg\max_{\gamma} |P(\phi,\gamma)|_{\phi\in\Theta,\gamma\in\Psi}, \quad (14)$$

where,

$$P(\phi,\gamma) = \frac{1}{\sum_{i=2}^{\hat{M}} \mathbf{s}_{ha}^H \mathbf{v}_i \mathbf{v}_i^H \mathbf{s}_{ha}}. \quad (15)$$

Because we have interest only in the elevation here, the search area of ϕ is restricted to $\Theta = \overline{\theta'}$.

3.5. Azimuth determination using *Triangle* -Step 3-

As the final step, we estimate the azimuth using *Triangle* again. The estimation method is almost same as that in Step 1, but this time we use the estimated elevation $\overline{\psi}$ of Eq.(14). Thus the *rotation matrices* and the *integrated frequency array data* used in this step are defined as following.

$$\mathbf{G}_{a2b}(\phi) \equiv \mathrm{diag}\left[\, e^{-j\alpha\omega_0\tau_{a2b}} \quad e^{-j\beta\omega_0\tau_{a2b}} \quad \cdots \right] \quad (16)$$

$$\mathbf{G}_{c2b}(\phi) \equiv \mathrm{diag}\left[\, e^{-j\alpha\omega_0\tau_{c2b}} \quad e^{-j\beta\omega_0\tau_{c2b}} \quad \cdots \right] \quad (17)$$

$$\mathbf{G}_T(\theta,\phi) \equiv \{\mathbf{G}_{a2b}\mathbf{G}_{ab} + \mathbf{G}_{bc} + \mathbf{G}_{c2b}\mathbf{G}_{ca}\}/3, \quad (18)$$

Table 1. *Parameters for simulation*

Input SNR	20dB
Sampling Frequency	16000Hz
Array Aperture D	8cm
Threshold T [1]	15dB
Window	Hamming
Frame Length	600
Frame Overlap	300
Data Length	625ms

where

$$\tau_{a2b}(\phi) \equiv \tau_{bc} - \tau_{ab} = \sqrt{3}D\sin(\phi - \tfrac{\pi}{6})\sin\overline{\psi}/v \quad (19)$$

$$\tau_{c2b}(\phi) \equiv \tau_{bc} - \tau_{ca} = \sqrt{3}D\sin(\phi + \tfrac{\pi}{6})\sin\overline{\psi}/v. \quad (20)$$

The eigenvector analysis is performed using the steering vector given by

$$\mathrm{s}_{bc}(\phi) = \left[\ e^{-j\alpha\omega_0\tau_{bc}(\phi,\gamma)} \quad e^{-j\beta\omega_0\tau_{bc}(\phi,\gamma)} \quad \cdots\ \right]^T\Big|_{\gamma=\overline{\psi}}, \quad (21)$$

and the estimated azimuth $\overline{\theta}$ is derived by the maximum search given by the same form as Eq.(14) and Eq.(15) with respect to ϕ within $\Theta = [-\pi, \pi]$.

For further improvement of the accuracy, we can repeat from Step.2 to Step.3 as far as the increase of computation cost is allowed.

4. SIMULATION AND EXPERIMENTAL RESULTS

4.1. Evaluation with computer simulation

For the computer simulation, we use the real 5 phoneme data (/a/, /e/,/i/,/o/,/u/) uttered by 10 subjects(5 each for male and female) as the source signal and had 5 trials for every data. The microphone array input signal is virtually generated by delaying the signal with an appropriate samples according to θ and sum up with additive white noise as the sensor noise. As the conventional methods for comparison, we adopt MUSIC with CSS. For the pre-estimated DOA information, we add estimation error factor following Gaussian distribution due to reflect the pre-estimation inaccuracy. All the same parameters shown in Tab.1 are adopted to every method, and for the conventional method, we use only the same harmonics selected in the proposed method.

Fig.5 and Fig.6 show the deviation of final estimation error (DEE). From these results, we can recognize that the proposed method keeps its high accuracy at every direction, and it is almost same level as that of the MUSIC-CSS with accurately pre-estimated DOA.

4.2. Evaluation at real acoustic environment

To verify that the proposed method is effective even at real acoustic environment, we performed some experiments at a large conference room ($W \times D \times H : 18 \times 15 \times 4$[m]). The speech data and parameters are same as in the computer simulation except for the threshold T settled at 10dB, and here we also made 5 trials for each data. Fig.7 and Fig.8 show the results of the experiment.

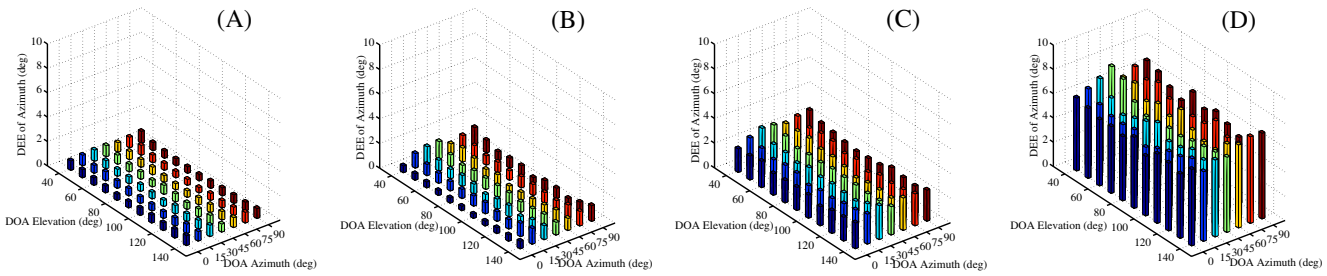

Fig. 5. DEE of azimuth at ideally anechoic case (A)Proposed (B)–(D)MUSIC-CSS ((B)dev=0[deg] (C)dev=1[deg] (D)dev=3[deg])

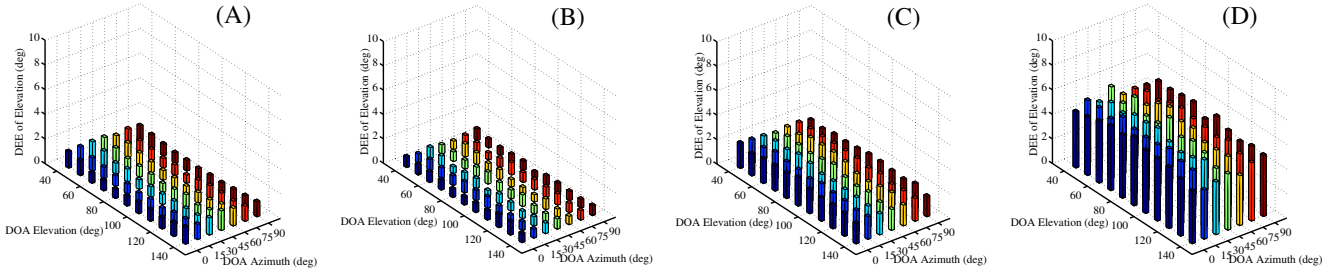

Fig. 6. DEE of elevation at ideally anechoic case (A)Proposed (B)–(D)MUSIC-CSS ((B)dev=0[deg] (C)dev=1[deg] (D)dev=3[deg])

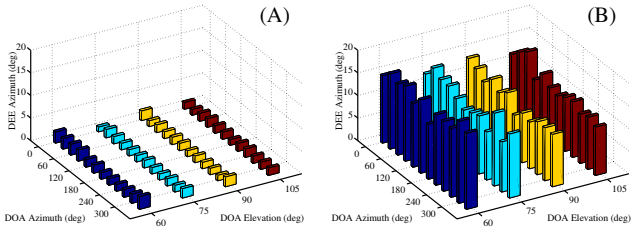

Fig. 7. DEE of azimuth at experiment (A)Proposed (B)MUSIC-CSS

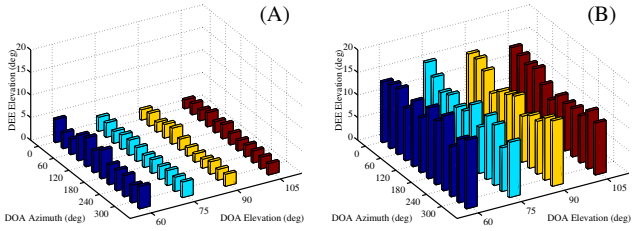

Fig. 8. DEE of elevation at experiment (A)Proposed (B)MUSIC-CSS

5. CONCLUSIONS

In this paper, we have proposed DOA estimation method of a speech signal using microphones located at apices of equilateral pyramid. The proposed method estimates both azimuth and elevation by separating the microphone pairs into two groups, and it achieves accurate estimation with less computational load. Through both computer simulation and experiment in a real acoustic environment, we have confirmed the performance of the proposed method. For a future subject, the estimation of more than one simultaneous speaker should be considered.

6. ACKNOWLEDGEMENT

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.

7. 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] Y. Hioka and N. Hamada, "DOA Estimation of Speech Signal using Microphones Located at Vertices of Triangle," Technical report of IEICE, EA2003-44, pp.9–16, 2003.

[3] G. Su and M. Morf, "Signal subspace approach for multiple wide-band emitter location," *IEEE Trans. on ASSP*, Vol.31, No.12, pp.1502-1522, 1983.

[4] 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 ASSP*, Vol.33, No.4, pp.823-831, 1985.

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

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

[7] H. Kanai, Spectrum Analysis of Sound and Vibration, CORONA Pub., 1999.(in Japanese)

[8] M.R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," J. Acoust. Soc. Am. vol.43,pp.829-834, 1968.