

DOA Estimation of Speech Signal using Equilateral-Triangular Microphone Array

Yusuke Hioka, Nozomu Hamada

School of Integrated Design Engineering
Keio University, Japan

hioka@hamada.sd.keio.ac.jp

hamada@sd.keio.ac.jp

Abstract

In this contribution, we propose a DOA (Direction Of Arrival) estimation method of speech signal whose angular resolution is almost uniform with respect to DOA. Our previous DOA estimation method[1] achieves high precision with only two microphones, however its resolution degrades as the propagating direction apart from the array broadside. In the proposed method, the equilateral-triangular microphone array is adopted, and the subspace analysis is applied. The efficiency of the proposed method is shown both from the simulation and experimental results.

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 subject[2]–[4], MUSIC(MUltiple SIngle Classification)[5] with Coherent Signal Subspace(CSS)[3] is known as an effective method with high spatial resolution. However, it requires rough DOA estimation *a priori*. Usually this pre-estimation accuracy highly affects the final estimation result. 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, it is often restricted in practical use 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 pre-estimation process[1]. In the method, however, the estimation resolution degrades as the propagating direction apart from the array broadside. This is an inherent problem in the linear array. Obviously, linear array is the only arrangement that a two-sensor microphone array can take. In this study, we settle our research purpose at preventing this degradation and realizing spatial uniform resolution.

Our main proposals in this research are summarized as the following two ideas.

- Adoption of the equilateral-triangular microphone array configuration
- Utilize the subspace structure alteration of the integrated array data

The former idea aims to make uniform resolution with respect to DOA. In an equilateral-triangular sensor arrangement,

we can extract three microphone pairs and each of them constructs a linear 2-sensors microphone array to which our previous method[1] is applied. Because the broadside of each microphone pair faces to different angle individually, we can expect the resolution improvement by integrating the array data at these three pairs. The second idea is introduced to achieve the integrated use of the three array data without pre-estimation. For such array data integration of non-linear array arrangements, a method analogous to the CSS called Array Interpolation[6] is well known, however, this method requires the DOA pre-estimation as well. At the estimation stage in our proposed method, we do not require any *a priori* DOA knowledge by utilizing the subspace analysis of the integrated array data.

This paper is organized as follows. In the following Sec.2, we briefly review our previous method to explain our problem settings, and the details of the proposed method are described in Sec.3. Simulation and experimental results are shown in Sec.4, and some concluding remarks are stated in Sec.5.

2. Virtually generated multichannel data

First, let us consider the two channel signals obtained by two 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 source signal, τ is the time delay between two microphones, and $n_1(n)$ and $n_2(n)$ are mutually uncorrelated noise signals. Thus, the Fourier transform 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

$$G_{12}(\omega) = X_1(\omega)^* X_2(\omega) = P(\omega)e^{-j\omega\tau} \quad (5)$$

respectively, where $P(\omega)$ is the power spectral density of $s(n)$. When we set $\omega = \omega_m$, where ω_m is the m -th higher harmonics of the fundamental frequency ω_0 ,

$$\omega_m = m\omega_0 \quad (6)$$

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, which is 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 obtained from the virtual multichannel signals, which are narrow-band signals acquired by an equally spaced linear array. For the

determination of frequency ω_0 and its harmonics, we use the harmonic structure of voiced sound. That is, we set ω_0 as the fundamental frequency of voiced sound in speech, therefore, its harmonics are used for generating multichannel signals. We define the following frequency array data $\mathbf{G}(\omega_0)$ for the received signal.

$$\mathbf{G}(\omega_0) = [G_{12}(a\omega_0) G_{12}(b\omega_0) \cdots]^T \quad (7)$$

$(a, b, \cdots \in \mathbf{m})$

Because the power spectrum distribution depends on speaker and phoneme, here we select the harmonics that contains the speech components in high SNR. In Eq.(7), \mathbf{m} is the set of the \hat{M} harmonics order selected by the magnitude-squared coherence function[1][7], and the fundamental frequency ω_0 is estimated by logarithmic harmonic product spectrum[8]. Then, the MUSIC algorithm is applied to $\mathbf{G}(\omega_0)$ for DOA estimation. Because the power of a voiced sound is localized in its harmonic frequencies, the SNR of the extracted data is rather high, and as a result, it contributes to improving the estimation accuracy of the MUSIC. Here we note that the magnitude of each components of $\mathbf{G}(\omega_0)$ are normalized at the harmonics selection.

3. Proposed method

In the proposed method, there are two main ideas as stated in Sec.1. The first idea is to receive the target signal by an equilateral-triangular microphone array. Because each microphone pair faces to different angle in this configuration, the frequency array data of each pair has its high spatial resolution at each facing direction. So we estimate the DOA by integrated use of these three array data for improving spatial resolution globally. Furthermore, this integration has another advantage that it can suppress the influence of noises including room reverberation. The second idea is to use the subspace structure of the integrated frequency array data. We explain its detail in Sec.3.3.

3.1. Model of input signal

Fig.1 shows the array input signal model received by the equilateral-triangular microphone array. Here we assume for the input signal as follows.

- Only one speech signal is received
- The elevation of the DOA is settled on the array plane

The Fourier transforms of each microphone input signals $x(n), y(n)$ and $z(n)$ are given by

$$\begin{cases} X(\omega) = S(\omega)e^{-j\omega\tau_x} + N_x(\omega) \\ Y(\omega) = S(\omega)e^{-j\omega\tau_y} + N_y(\omega) \\ Z(\omega) = S(\omega)e^{-j\omega\tau_z} + N_z(\omega) \end{cases} \quad (8)$$

where τ_x denotes the signal arrival delay at microphone \mathbf{x} with respect to the reference point located at the array origin. Here we can define the cross spectra of three microphone combinational pairs as shown in Eq.(9).

$$\begin{cases} G_{xy}(\omega) = X^*(\omega)Y(\omega) = P(\omega)e^{-j\omega\tau_{xy}} \\ G_{yz}(\omega) = Y^*(\omega)Z(\omega) = P(\omega)e^{-j\omega\tau_{yz}} \\ G_{zx}(\omega) = Z^*(\omega)X(\omega) = P(\omega)e^{-j\omega\tau_{zx}} \end{cases} \quad (9)$$

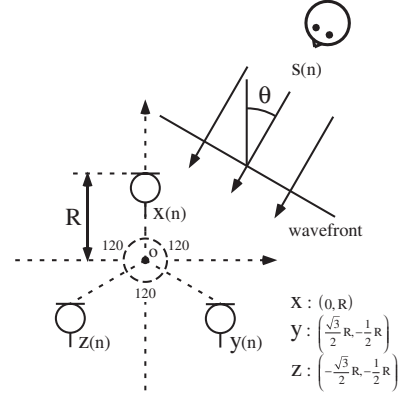


Figure 1: Model of input signal

The delay variables in Eq.(9) are the function of DOA θ given by

$$\tau_{xy}(\theta) = \sqrt{3}R \sin(\theta + \frac{2}{3}\pi)/v \quad (10)$$

$$\tau_{yz}(\theta) = \sqrt{3}R \sin(\theta)/v \quad (11)$$

$$\tau_{zx}(\theta) = \sqrt{3}R \sin(\theta - \frac{2}{3}\pi)/v \quad (12)$$

where v denotes the sound velocity. From Eq.(9)–Eq.(12), we form the frequency array data $\mathbf{G}_{xy}(\omega_0, \theta)$, $\mathbf{G}_{yz}(\omega_0, \theta)$ and $\mathbf{G}_{zx}(\omega_0, \theta)$ respectively, by extracting the \hat{M} harmonic components and normalize the magnitude of each element. For simplicity, we omit the ω_0 in the following part of this paper.

3.2. Integration of frequency array data

Now let us consider the difference of delay term (which determines the phase value) between two frequency array data for a signal propagating from direction ϕ .

$$\tau_{x2y}(\phi) \equiv \tau_{yz}(\phi) - \tau_{xy}(\phi) = 3R \sin(\phi - \frac{\pi}{6})/v \quad (13)$$

$$\tau_{z2y}(\phi) \equiv \tau_{yz}(\phi) - \tau_{zx}(\phi) = 3R \sin(\phi + \frac{\pi}{6})/v \quad (14)$$

For the interpolation of these delay differences, we define the following $\hat{M} \times \hat{M}$ matrices called *rotation matrices* that consist of the phase rotating components relating to the signal from direction ϕ .

$$\mathbf{G}_{x2y}(\phi) \equiv \text{diag} [e^{-ja\omega_0\tau_{x2y}(\phi)} e^{-jb\omega_0\tau_{x2y}(\phi)} \dots] \quad (15)$$

$$\mathbf{G}_{z2y}(\phi) \equiv \text{diag} [e^{-ja\omega_0\tau_{z2y}(\phi)} e^{-jb\omega_0\tau_{z2y}(\phi)} \dots] \quad (16)$$

Multiplying each frequency array data by these *rotation matrices*, and summing up the rotated array data together as shown in Eq.(17).

$$\mathbf{G}_m(\phi, \theta) = \{\mathbf{G}_{x2y}(\phi)\mathbf{G}_{xy}(\theta) + \mathbf{G}_{yz}(\theta) + \mathbf{G}_{z2y}(\phi)\mathbf{G}_{zx}(\theta)\}/3 \quad (17)$$

When the variable ϕ equals to θ , the phases of each terms in the right side of Eq.(17) are equal.

3.3. Subspace analysis

Here let us note the delay term of each rotated frequency array data in Eq.(17).

$$\begin{aligned} \mathbf{G}_{x2y}(\phi)\mathbf{G}_{xy}(\theta) &\rightarrow \tau_{x2y}(\phi) + \tau_{xy}(\theta) \equiv \check{\tau}_{xy}(\phi, \theta) \\ \mathbf{G}_{yz}(\theta) &\rightarrow \tau_{yz}(\theta) \equiv \check{\tau}_{yz}(\phi, \theta) \\ \mathbf{G}_{z2y}(\phi)\mathbf{G}_{zx}(\theta) &\rightarrow \tau_{z2y}(\phi) + \tau_{zx}(\theta) \equiv \check{\tau}_{zx}(\phi, \theta) \end{aligned} \quad (18)$$

The sign " \rightarrow " denotes to extract the delay term of a frequency array data. Now for any direction $\phi \in [-\pi, \pi]$, the following lemma is satisfied. Its proof is described in the appendix A.

[Lemma]

Delay terms of all the three rotated frequency array data in Eq.(17) are equal if and only if ϕ is equal to θ .

$$\phi = \theta \iff \check{\tau}_{xy} = \check{\tau}_{yz} = \check{\tau}_{zx} \quad (19)$$

From this lemma, we can replace our DOA estimation problem by searching a *rotation matrix*, which equalizes the delay terms of all three frequency array data. To solve this subject, we introduce the theorem below.(see appendix B for its proof)

[Theorem]

The integrated frequency array data $\mathbf{G}_m(\phi, \theta)$ is equal to a steering vector $\mathbf{s}(\phi)$ defined by

$$\mathbf{s}(\phi) = [e^{-ja\omega_0\tau_{yz}(\phi)} \quad e^{-jb\omega_0\tau_{yz}(\phi)} \quad \dots]^T \quad (20)$$

if and only if the all interpolated delay terms are equal. That is,

$$\check{\tau}_{xy} = \check{\tau}_{yz} = \check{\tau}_{zx} \iff \mathbf{G}_m(\phi, \theta) = \mathbf{s}(\phi) \quad (21)$$

This theorem leads the DOA problem to search the parameter ϕ satisfying the equality $\mathbf{G}_m(\phi, \theta) = \mathbf{s}(\phi)$. In order to determine ϕ that satisfies this condition, we use the subspace structure of covariance matrix $\mathbf{R}_m(\phi)$ derived as

$$\mathbf{R}_m(\phi) = E [\mathbf{G}_m \mathbf{G}_m^H] \quad (22)$$

Because $\mathbf{R}_m(\phi)$ is an Hermitian matrix, each eigenvector \mathbf{v}_i of $\mathbf{R}_m(\phi)$ is mutually orthogonal.

$$\mathbf{v}_i^H \mathbf{v}_j = \delta_{ij} \quad (23)$$

We also have,

$$\mathbf{R}_m(\phi) = \sum_{i=1}^{\hat{M}} \lambda_i \mathbf{v}_i \mathbf{v}_i^H \quad (24)$$

From the well-known theorem of the array covariance matrix[5], the eigenvector relating to the largest eigenvalue λ_1 is equal to the vector \mathbf{G}_m in case of rank-1 model. We can estimate DOA $\hat{\theta}$ by the following null search strategy.

$$P(\phi) = \frac{1}{\sum_{i=2}^{\hat{M}} \mathbf{s}^H(\phi) \mathbf{v}_i \mathbf{v}_i^H \mathbf{s}(\phi)} \quad (25)$$

$$\hat{\theta} = \arg \max_{\phi} |P(\phi)| \quad (26)$$

Fig.2 shows the flow diagram of the proposed method.

4. Simulation and experiment results

This section shows some results of both computer simulation and experiments in real acoustic environment to evaluate the proposed method.

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 for each sensor according to θ and sum up with additive white noise as the sensor noise. As the conventional methods for comparison, we

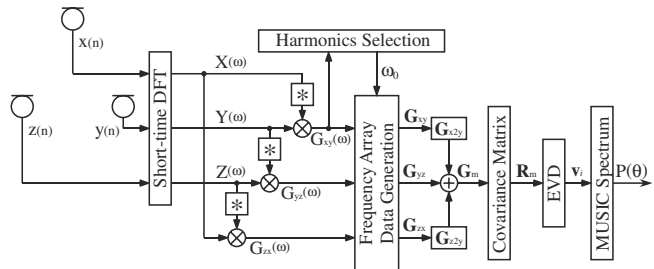


Figure 2: Flow diagram of the proposed method

Table 1: Parameters for simulation

Input SNR	20dB
Sampling Frequency	16000Hz
Array Aperture R	$\frac{0.08}{\sqrt{3}}$ m
Threshold T [1]	15dB
Window	Hamming
Frame Length	600
Frame Overlap	300
Data Length	625ms

adopt our previous method with linearly located 2, 3 and 4 microphones to perform impartial comparison at the number of sensors and frequency array data respectively. Furthermore, we also compare with common DOA estimation method that is MUSIC with CSS. The pre-estimation is covered with the beamforming method, and we add estimation error factor following Gaussian distribution due to reflect the effect of its low spatial resolution. All the same parameters shown in Tab.1 are adopted to every method, and for the conventional methods, we use only the same harmonics selected in the proposed method.

Fig.3 shows the $P(\phi)$ (called "spectrum") of the proposed method. The spectrum given by the proposed method shows sharpness at the estimated angle as high as that of the conventional methods. Now for the evaluation of spatial resolution, Fig.4 shows the deviation of estimation error. From this result, we can distinctly recognize that the proposed method keeps its high spatial resolution to every direction, and its accuracy is almost same level as that of the MUSIC-CSS with the correct pre-estimated DOA.

4.2. Experiments 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 SNR lying around 9dB and the threshold T is settled at 10dB, and here we also made 5 trials for each data. Fig.5 shows the results of the experiment. Due to the source and microphone position error, the true direction θ is not possibly given in the experiment. So we regard the median value of the estimation results as the true direction and evaluate the error deviation from this value. This result shows that the proposed method provides the best resolution for the DOA estimation, and furthermore, the result of the proposed method is even better than that of MUSIC-CSS which is crucially degraded by the pre-estimation error.

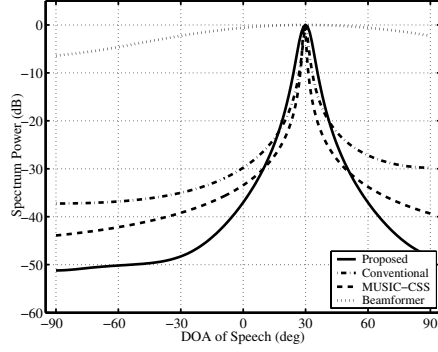


Figure 3: Estimated Spectra (female /a/, $\theta = 30^\circ$)

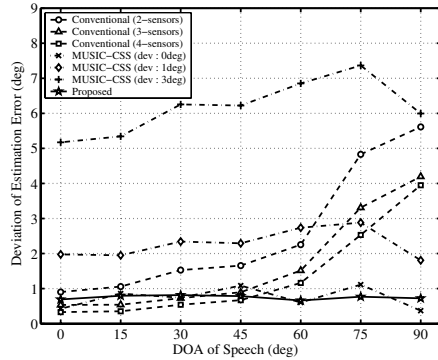


Figure 4: Deviation of Estimation Error

5. Conclusions

In this contribution, we proposed a DOA estimation of a speech signal with high spatial resolution. The proposed method uses an equilateral-triangular microphone array to improve the estimation resolution, and subspace analysis of integrated frequency array data to avoid the pre-estimation stage. Through both computer simulation and experiment in a real acoustic environment, we have confirmed the enhancement by the proposed method. For future subject, the estimation error and possibility for the elevation direction estimation should be considered. Another future subject is to estimate directions of more than one speaker automatically.

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. The participation expense for this conference is also supported partially by NTT DoCoMo Co..

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] G. Su and M. Morf, "Signal subspace approach for

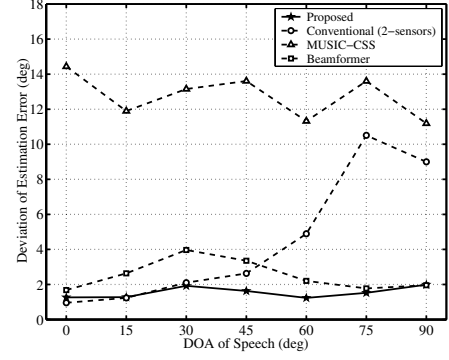


Figure 5: Results of experiment at real acoustic environment

multiple wide-band emitter location," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol.31, No.12, pp.1502-1522, 1983.

- [3] 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.
- [4] 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.
- [5] D.H. Johnson and D.E. Dudgeon, "Array Signal Processing," PTRP Prentice Hall, 1993.
- [6] B. Friedlander and A. Weiss, "Direction Finding Using Spatial Smoothing with Interpolated Arrays," *IEEE Trans. AES*, 28, pp.574–587, Apr. 1992.
- [7] H. Kanai "Spectrum Analysis of Sound and Vibration," CORONA Pub. Co. LTD., 1999.(in Japanese)
- [8] M.R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Am.* Vol.43:829-834, 1968.

Appendix A

- $\check{\gamma}_{xy}(\phi, \theta) = \check{\gamma}_{yz}(\phi, \theta) \Rightarrow \phi = \theta$ and $\phi = -\frac{2}{3}\pi - \theta$
 $\check{\gamma}_{zx}(-\frac{2}{3}\pi - \theta, \theta) \neq \check{\gamma}_{xy}(-\frac{2}{3}\pi - \theta, \theta) = \check{\gamma}_{yz}(-\frac{2}{3}\pi - \theta, \theta)$
- $\check{\gamma}_{yz}(\phi, \theta) = \check{\gamma}_{zx}(\phi, \theta) \Rightarrow \phi = \theta$ and $\phi = \frac{2}{3}\pi - \theta$
 $\check{\gamma}_{xy}(\frac{2}{3}\pi - \theta, \theta) \neq \check{\gamma}_{yz}(\frac{2}{3}\pi - \theta, \theta) = \check{\gamma}_{zx}(\frac{2}{3}\pi - \theta, \theta)$
- $\check{\gamma}_{zx}(\phi, \theta) = \check{\gamma}_{xy}(\phi, \theta) \Rightarrow \phi = \theta$ and $\phi = -\theta$
 $\check{\gamma}_{yz}(-\theta, \theta) \neq \check{\gamma}_{zx}(-\theta, \theta) = \check{\gamma}_{xy}(-\theta, \theta)$

Appendix B

The magnitude of k -th element in \mathbf{G}_m is less than 1 not as far as all the interpolated delay terms are equal.

$$\begin{aligned}
 & |[\mathbf{G}_m]_k| \\
 &= |e^{jk\omega_0 \check{\tau}_{xy}} + e^{jk\omega_0 \check{\tau}_{yz}} + e^{jk\omega_0 \check{\tau}_{zx}}|/3 \\
 &\leq \{|e^{jk\omega_0 \check{\tau}_{xy}}| + |e^{jk\omega_0 \check{\tau}_{yz}}| + |e^{jk\omega_0 \check{\tau}_{zx}}|\}/3 = 1
 \end{aligned}$$

The equality is satisfied only if the three complex values are equal.