ESTIMATING DOA OF MULTIPLE SPEECH SIGNALS BY IMPROVED HISTOGRAM MAPPING METHOD

[1]*Masao Matsuo,* [2]*Yusuke Hioka, and* [3]*Nozomu Hamada*

[1]masao@hamada.sd.keio.ac.jp
[123]Hamada Laboratory, System Design Engineering, Keio University
Yokohamashi Kouhokuku Hiyoshi 3-14-1, Kanagawa, Japan

ABSTRACT

Multiple speech source localization has been an important research topic over the recent years. Histogram mapping method proposed by Huang *et al.*[1] has two major advantages over the other conventional studies, that it does not require the preliminary DOA estimates, and the estimation is achieved with low computational complexity. In this study, we attempt to improve the performance of Huang's method by introducing two new procedures. The first improvement is a novel method in which the DOA is obtained from coordinate transformation. The second procedure is the deletion of useless narrow-band components. Simulation results show that the estimation error is reduced to about half of the original histogram mapping method when two speech sources are active simultaneously.

1. INTRODUCTION

Speech source localization is used for numbers of applications, such as teleconference systems and automatic voice recognition systems. In teleconference systems, the direction of arrival (DOA) is tracked in order to steer the video camera toward the active speaker. In automatic voice recognition systems, audio beam is directed toward the speaker's direction in order to separate the desired speech signal from the environmental noise.

The most basic approach to finding the direction of sound source is to estimate the time difference of arrival (TDA) between the sensors of a microphone array. Knapp [2] introduced Generalized Cross Correlation (GCC) method in which the TDA is estimated from cross-correlation of the input signals. The general weighting function introduced in this method works as a prefilter to facilitate the estimation of TDA. However, GCC method becomes problematic when multiple sources are active simultaneously.

The well-known, long-standing method for multiple sound source localization is the delay-and-sum beamformer method [3]. In this method, the DOA is estimated by shifting the main beam of the beamformer and detecting the peaks of the output power. However, the resolution capability of this method is limited, which makes it impractical for real-life applications.

Wang and Kaveh [4] proposed coherent signal-subspace processing for multiple wide-band source localization. This method is based on signal-subspace processing which was originally developed for narrow-band DOA estimation. However, this method requires a preliminary DOA estimates in the neighborhoods of the true DOAs, and the accuracy of the final DOA estimates is significantly dependent on the accuracy of preliminary estimates.

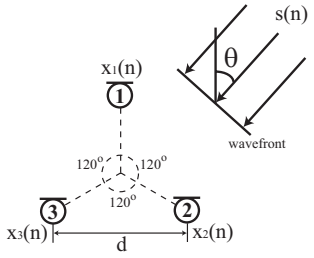

Figure 1: *Triangular microphone array.*

The method proposed in this paper is based on histogram mapping method [1], which is reported to be an effective multiple speech source localization method. This method has two major advantages over the other conventional studies, that it does not require the preliminary DOA estimates, and the estimation is achieved with low computational complexity. The following two procedures are adopted in the proposed method in order to improve the performance of histogram mapping method.

1. DOA estimation via coordinate transformation
2. Deletion of useless narrow-band components

As a result, the accuracy of estimating DOA and detecting the number of speakers are both improved.

2. HISTOGRAM MAPPING METHOD

The histogram mapping method uses the triangular microphone array illustrated in Figure 1. There are three microphones located at the vertices of equilateral-triangle whose inter-microphone distance is d, and they receive the speech signal $s(n)$ propagating from the direction θ. In the experiment of [1], d is settled at $0.135\mathrm{m}$.

2.1. Decomposing the Input Speech Signal

In histogram mapping method, the multiple speech signals input to the microphone array are decomposed into its frequency components via short-time Fourier transform, and TDA is estimated for each frequency component. Since spectral differences

between the individual human voices are known to exist, the following hypothesis is adopted.

> Single frequency cell of the input signal in any time frame contains spectral component of only *one* speech signal.

Therefore, decomposing the input signal into its frequency components via STFT transforms a wide-band, multiple source problem into a set of narrow-band, single source problems. Quantitative evaluation of this hypothesis is given in [5] and [6].
The STFT of the input signal at i-th microphone can be modeled as,

$$X_i(n,\omega) = S(n,\omega)e^{-j\omega\tau_i} \tag{1}$$

where n represents the time frame, ω represents the angular frequency, $S(n,\omega)$ represents the STFT of an arriving signal, and τ_i represents the signal delay with respect to the center of the array. Within the triangular array, there can be considered three pairs of microphones each facing a different direction by 120°. TDAs are estimated between each microphone pair, using the phase angle differences between the input signals.

$$\hat{\tau}_{12}(n,\omega) = \frac{1}{\omega}\angle\frac{X_1(n,\omega)}{X_2(n,\omega)} \tag{2}$$

$$\hat{\tau}_{23}(n,\omega) = \frac{1}{\omega}\angle\frac{X_2(n,\omega)}{X_3(n,\omega)} \tag{3}$$

$$\hat{\tau}_{31}(n,\omega) = \frac{1}{\omega}\angle\frac{X_3(n,\omega)}{X_1(n,\omega)} \tag{4}$$

2.2. Creating TDA Histograms

The above $\hat{\tau}_{12}(n,\omega)$, $\hat{\tau}_{23}(n,\omega)$, and $\hat{\tau}_{31}(n,\omega)$ are mapped into three separate histograms with TDA along the horizontal axis. The three histograms are then averaged to form a single TDA histogram and smoothed in order to remove insignificant small peaks. The number of peaks in the resultant histogram can be considered the number of active speakers at the n-th time frame, and the center of the i-th peak $\hat{\tau}_i(n)$ is the estimated TDA corresponding to the i-th speech signal. Finally, the azimuth of the i-th speaker $\hat{\theta}_i(n)$ is calculated from $\hat{\tau}_i(n)$ by the following equation

$$\hat{\theta}_i(n) = \arcsin\left(\frac{c}{d}\hat{\tau}_i(n)\right) \tag{5}$$

where c represents the propagation velocity.

3. PROPOSED METHOD

In the proposed method, the three TDAs obtained by the triangular microphone array are collectively used in a vector form to improve the accuracy of DOA estimation. In addition, a procedure to discriminate and delete useless frequency components is introduced.

3.1. Vectorial Analysis of TDAs

The three TDAs estimated by equations (2)∼(4) are integrated in a vector form as follows.

$$\hat{\boldsymbol{\tau}}(n,\omega) = \left[\hat{\tau}_{12}(n,\omega), \hat{\tau}_{23}(n,\omega), \hat{\tau}_{31}(n,\omega)\right]^T \tag{6}$$

We refer to $\hat{\boldsymbol{\tau}}(n,\omega)$ as the *TDA vector*. Meanwhile, assuming that the location of the sources are restricted on the array plane, the theoretical relationship between the azimuth of a single speech source θ and the three TDAs are described by the following equations.

$$\tau_{12}(\theta) = \tfrac{d}{c}\sin(\theta + \tfrac{2}{3}\pi) \tag{7}$$

$$\tau_{23}(\theta) = \tfrac{d}{c}\sin(\theta) \tag{8}$$

$$\tau_{31}(\theta) = \tfrac{d}{c}\sin(\theta - \tfrac{2}{3}\pi) \tag{9}$$

Thus, the theoretical TDA vector can also be given as a function of θ.

$$\boldsymbol{\tau}(\theta) = \left[\tau_{12}(\theta), \tau_{23}(\theta), \tau_{31}(\theta)\right]^T \tag{10}$$

There is a one-to-one relationship between θ and $\boldsymbol{\tau}(\theta)$. In other words, the DOA of the speech signal uniquely determines the TDA vector.

3.2. DOA Estimation via Coordinate Transformation

By using the one-to-one relationship between θ and $\boldsymbol{\tau}(\theta)$, it is possible to determine the DOA from the TDA vector. In this subsection, we will explain a procedure to obtain θ by transforming the coordinates of $\boldsymbol{\tau}(\theta)$. The original coordinate $[\tau_{12}, \tau_{23}, \tau_{31}]$ will be transformed into a new coordinate composed of the plane containing the locus of $\boldsymbol{\tau}(\theta)$, and the axis orthogonal to this plane.
First, we will define the transformation matrix $\boldsymbol{T}$. To make the x-axis of the new coordinate corresponds to the direction of $\boldsymbol{\tau}(\pi/2)$, the first unit basis vector $\boldsymbol{e}_1$ of the transformation matrix $\boldsymbol{T}$ is defined as the normalized vector of $\boldsymbol{\tau}(\pi/2)$.

$$\begin{aligned}\boldsymbol{e}_1 &= \tfrac{1}{\|\boldsymbol{\tau}(\pi/2)\|}\boldsymbol{\tau}(\pi/2) \\ &= \left[\sqrt{\tfrac{2}{3}}\sin(-\tfrac{1}{6}\pi), \sqrt{\tfrac{2}{3}}, \sqrt{\tfrac{2}{3}}\sin(\tfrac{7}{6}\pi)\right]^T\end{aligned} \tag{11}$$

Likewise, to make the y-axis of the new coordinate correspond to the direction of $\boldsymbol{\tau}(0)$, the second unit basis vector $\boldsymbol{e}_2$ of the transformation matrix $\boldsymbol{T}$ is defined as the normalized vector of $\boldsymbol{\tau}(0)$.

$$\begin{aligned}\boldsymbol{e}_2 &= \tfrac{1}{\|\boldsymbol{\tau}(0)\|}\boldsymbol{\tau}(0) \\ &= \left[\sqrt{\tfrac{2}{3}}\sin(\tfrac{2}{3}\pi), 0, \sqrt{\tfrac{2}{3}}\sin(-\tfrac{2}{3}\pi)\right]^T\end{aligned} \tag{12}$$

Lastly, the third unit basis vector $\boldsymbol{e}_3$ of $\boldsymbol{T}$ is defined as the normalized vector orthogonal to both $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$.

$$\boldsymbol{e}_3 = \left[\sqrt{\tfrac{1}{3}}, \sqrt{\tfrac{1}{3}}, \sqrt{\tfrac{1}{3}}\right]^T \tag{13}$$

Thus, the transformation matrix $\boldsymbol{T}$ is defined as

$$\boldsymbol{T} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3] \tag{14}$$

We will obtain the new coordinate $[x, y, z]$ by transforming the current coordinate $[\tau_{12}, \tau_{23}, \tau_{31}]$ as follows.

$$\begin{aligned}\boldsymbol{\tau}'(\theta) &= [x,y,z]^T = \boldsymbol{T}^T\boldsymbol{\tau}(\theta) \\ &= \begin{bmatrix} \sqrt{\frac{2}{3}}\sin(-\frac{1}{6}\pi) & \sqrt{\frac{2}{3}} & \sqrt{\frac{2}{3}}\sin(\frac{7}{6}\pi) \\ \sqrt{\frac{2}{3}}\sin(\frac{2}{3}\pi) & 0 & \sqrt{\frac{2}{3}}\sin(-\frac{2}{3}\pi) \\ \sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} \end{bmatrix} \begin{bmatrix} \frac{d}{c}\sin(\theta+\frac{2}{3}\pi) \\ \frac{d}{c}\sin(\theta) \\ \frac{d}{c}\sin(\theta-\frac{2}{3}\pi) \end{bmatrix} \\ &= \begin{bmatrix} 1.5\sqrt{\frac{2}{3}}\frac{d}{c}\sin\theta \\ 1.5\sqrt{\frac{2}{3}}\frac{d}{c}\cos\theta \\ 0 \end{bmatrix}\end{aligned} \tag{15}$$

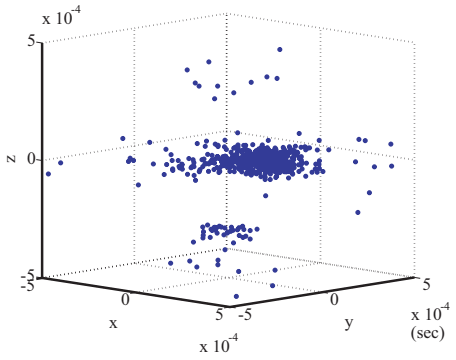

Figure 2: $\hat{\boldsymbol{\tau}}(\theta)$ *of actual speech signal.*

Then, θ is obtained by

$$\theta = \arctan \frac{x}{y}. \tag{16}$$

In the same way, the DOA *estimate* of each narrow-band component of the input array signal can be calculated by transforming the coordinate of the estimated TDA vector $\hat{\boldsymbol{\tau}}(n,\omega)$,

$$\begin{aligned} \hat{\boldsymbol{\tau}}'(n,\omega) &= [x(n,\omega), y(n,\omega), z(n,\omega)]^T \\ &= \boldsymbol{T}^T \hat{\boldsymbol{\tau}}(n,\omega) \end{aligned} \tag{17}$$

and by applying the following operation.

$$\hat{\theta}(n,\omega) = \arctan \frac{x(n,\omega)}{y(n,\omega)} \tag{18}$$

3.3. The Effect of Elevation Angle

In real environments, the assumption that the sound sources are located on the array plane, may not be satisfied. If we consider the elevation angle ϕ with respect to the array plane, the each component of the TDA vector is multiplied by the same factor $\cos\phi$. Multiplying a vector by a constant will only change the norm of the vector. Therefore, DOA estimation via coordinate transformation explained in the preceding section is also applicable to sources that are apart from the array plane.

3.4. Deletion of Useless Narrow-Band Components

In this section, we will introduce a step to discriminate frequency components of the input signal that are useless for DOA estimation. Figure 2 shows the plot of $\hat{\boldsymbol{\tau}}'(n,\omega)$ obtained from the real speech signal input to the microphone array. In the graph, the majority of $\hat{\boldsymbol{\tau}}'(n,\omega)$ is located on the xy plane as they should be theoretically. However, some are apart from the xy plane. Those outlier data do not satisfy the geometrical property of an accurate TDA vector. In other words, they are useless for giving a correct DOA estimation. Therefore, we delete frequency components that satisfy the following threshold function.

$$|z(n,\omega)| > \mathrm{TH} \tag{19}$$

4. RESULTS AND PERFORMANCE EVALUATION

4.1. Computer Simulation

The array input signal is generated by delaying the source signal with appropriate samples according to the source direction θ. For the source signals, we use the speech samples recorded in a database CD-ROM provided by Acoustic Society of Japan [7]. The signal is also added with Gaussian white noise to simulate the sensor noise. The parameters are listed in Table 1.

Table 1: Simulation parameters

Input SNR	14dB
Sampling frequency	16000Hz
Wave velocity c	340m/s
Microphone distance d	0.08m
Window	Hamming
STFT point	4096
Frame length	1024 samples
Frame overlap	512 samples
Length of snapshot	4096 samples (256msec)
TH	10^{-19} sec

4.1.1. Evaluation of Resolution Capability

First, we evaluate the performance of the proposed method in a "double-talk" situation. Figure 3 shows result of a simulated experiment when two different speech sources are active simultaneously. The first speaker is held fixed at 0°, and the position of the second speaker is changed from 10° to 180° at 10° intervals. The horizontal axis corresponds to the true direction of the second speaker, and the vertical axis denotes the estimated DOA for both speakers. The result is compared to that of the delay-and-sum beamformer [3].

As can be observed from the figure, the proposed method is successfully able to estimate the DOA for both speakers with very low deviation errors. The directions of the two speakers are resolved even when they are only 20° apart.

In contrast, the beamformer gives estimation results with significant amount of bias and deviation errors. Also, the limited resolution capability of the beamformer can be seen; the directions of the two speakers are resolved only when their angular distance exceeds 140°.

4.1.2. Estimation Accuracy for Multiple Active Sources

Figure 4 shows the root mean square error (RMSE) with respect to the number of active speech sources. The result is compared against that of the original histogram mapping method. It is shown that the RMSE of the proposed method is reduced to 51% of the original method when two sources are active, 57% when three sources are active, and 67% when four sources are active.

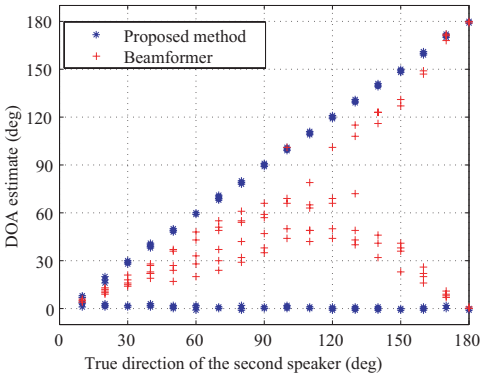

Figure 3: *DOA estimation for two active sources.*

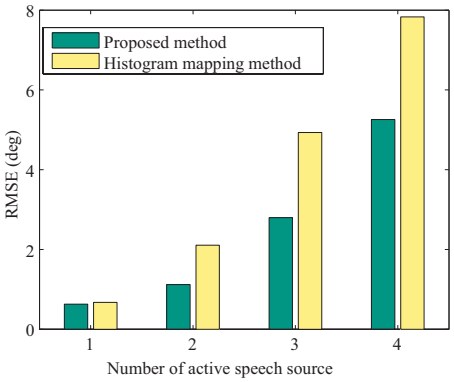

Figure 4: *RMSE of DOA estimation for multiple active sources.*

4.2. Experimental Results

To verify the effectiveness of the proposed method in a real acoustic environment, we performed an experiment in a $7.3 \times 8.2 \times 3.5m^3$ conference room. The source signals are real human voices speaking Japanese sentences of about 6 seconds in lengths. Two people are positioned 1.5m away from the center of the microphone array, at the direction of $\theta_1 = 90^\circ$ and $\theta_2 = 0^\circ$, and spoke toward the array at about the time. Figure 5 (a) shows the waveform of the signal obtained at microphone 1, Fig.5 (b) shows the DOA estimation results for histogram mapping method, and Fig.5 (c) shows the DOA estimation results for the proposed method. Notice from Fig.5 (b) that the histogram mapping method repeatedly fails to detect the second speaker located at $\theta_2 = 0^\circ$. In contrast, the proposed method gives accurate estimation results for both of the speech signals. We can conclude that not only the estimation accuracy, but also the detection capability is improved in the proposed method.

5. CONCLUSION

In this study, we attempted to improve the performance of histogram mapping method proposed by Huang et al. In Huang's method, three TDA histograms obtained by a triangular microphone array are averaged for DOA estimation. In our method, the three TDAs were integrated in a vector form and collectively used for DOA estimation. Moreover, we have introduced a step to discriminate and delete the frequency components that are useless for DOA estimation. Through computer simulation and actual experiment, we have confirmed that the proposed method outperforms the original histogram mapping method both in estimating DOA and detecting the number of speech sources.

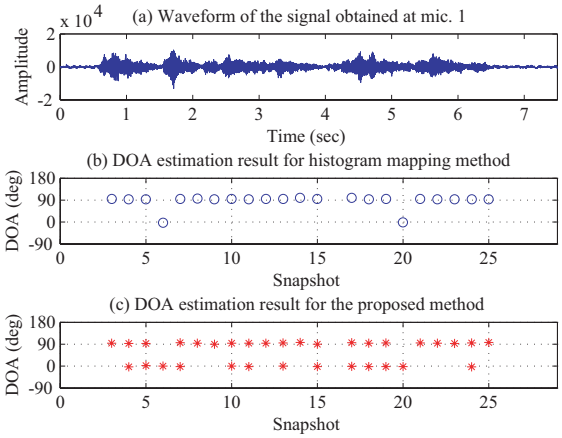

Figure 5: *DOA estimation for two active sources in a real acoustic environment.*

6. REFERENCES

[1] Jie Huang, Noboru Ohnishi, and Noboru Sugie, "A Biomimetic System for Localization and Separation of Multiple Sound Sources" IEEE Trans. on Instrumentation and Meas. Vol. 44, No. 3, pp.733–738, 1995.

[2] Charles H. Knapp, "The Generalized Correlation Method for Estimation of Time Delay"*IEEE Trans. on ASSP*, Vol. ASSO-24, No. 4, pp320–327, 1976.

[3] N. Kikuma, "Beamformer Method," in Adaptive Signal Processing with Array Antenna, pp.178–181, Science and Technology Publishing Company, Inc., 1999. (in Japanese)

[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. ASSP-33, No. 4, pp.823–831, 1985.

[5] Baruch Berdugo, Judith Rosenhouse, Haim Azhari, "Speaker's Direction Finding Using Estimated Time Delays in the Frequency Domain" *Signal Processing 82*, pp.19–30, 2002.

[6] Ozgur Yilmaz, Scott Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking" *IEEE Trans. on Signal Processing*, Vol. 52, Issue 7, pp1830–1847, 2004.

[7] "Continuous Speech Corpus for Research Vol.1–3," Japan Information Processing Development Center, ©Shuichi Itahashi (Edited by the Acoustical Society of Japan) 1991.