A TRACKING ALGORITHM OF SPEAKER DIRECTION USING MICROPHONES LOCATED AT VERTICES OF EQUILATERAL-TRIANGLE

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
phone: +81 45 563 1141, fax: +81 45 566 1720, email: hioka@hamada.sd.keio.ac.jp

ABSTRACT

In this paper, we propose a tracking algorithm of speaker direction using microphones located at vertices of an equilateral triangle. The method realizes tracking by minimizing a performance index that consists of the cross spectra at three different microphone pairs in the triangular array. For guaranteeing global convergence to the correct direction with accuracy, we use the steepest descent method with the weighted cross spectra at harmonic frequencies of voiced sound. During the adaptation, the weights are altered depending on the convergence state. Through some computer simulation and experiments in a real acoustic environment, we show the effectiveness of the proposed method.

1. INTRODUCTION

Under the rapid increase of the data transmission bit rate recently, the teleconference or remote lecturing system are getting familiar even with individual users. The speaker direction tracking is one of the essential technology at these systems not only for focusing the desired speech signal but also for steering the camera to point at the speaker, and several tracking algorithms using microphone array have been reported[1]–[5]. Some of them rely on adaptive beamforming, such as LCMV[1] and GSC[2] that captures the desired speech while suppressing the directional interferences adaptively, so that the speaker direction is determined from the beampattern of the given array weights. However, the method requires the beampattern calculation at every weight update, therefore, it heavily increases the computational cost. On the other hand, Kawakami *et al.*[3] proposed a method that realizes tracking by minimizing the output power of null steering fixed beamformer, and Suyama *et al.*[4] extended this method to double talk situation by introducing the data classification on the time-frequency domain. These methods directly update the speaker's direction, without calculating beampattern, but it does not track the abrupt movement due to that the evaluation function has some local minima. Zhang *et al.*[5] proposed another tracking method by combining focusing and TDE(Time Delay Estimation), however, this scheme is also applicable to the speaker with smooth movement. In the case of teleconference or remote lecturing, the current speaker position often alternates to another one, therefore these conventional methods are not appropriate for such application. From the practical point of view in addition, the tracking accuracy should be spatially uniform for omni-direction and the small number of microphones and array aperture is preferable.

In this paper, we propose a new tracking algorithm of speaker direction using three microphones located at vertices of an equilateral-triangle (we call it "equilateral-triangular microphone array"). By the same microphone configuration system, we previously proposed a DOA(Direction Of Arrival) estimation method[6]. In this method, we make integrated use of three different microphone pairs extracted from the array to realize uniform spatial resolution. However, the estimation algorithm is difficult to extend to the adaptive scheme due to its huge computational cost.

The main proposals of the new method are summarized as follows.

- A novel tracking mechanism realized by the integrated use of three cross spectra from the equilateral-triangular microphone array.
- Utilize of cross spectra at harmonic frequencies as the evaluation function with variable weights depending on the convergence state.

The former is to realize the tracking system for omni-direction with uniform resolution. The method localizes the speaker by minimizing the performance index containing the phase components of the cross spectra. The second idea enhances the tracking accuracy. Because much of the speech signal power concentrates at specific harmonic frequencies, the SNR at the harmonics are rather high and it results in high accuracy in estimation. Due to the varying characteristic of evaluation function along the frequency band, we alter the weights depending on the convergence state and SNR at each harmonic component to exploit these features.

This paper is organized as follows. In the following Sec.2, we mention the problem formulation with the equilateral-triangular microphone array, and the proposed method is described in Sec.3. Simulation and experimental results are shown in Sec.4 to confirm the efficiency of the proposed method, and some concluding remarks are stated in Sec.5.

2. PROBLEM FORMULATION

In this study, we use the equilateral-triangular microphone array as shown in Fig.1. A speaker in the direction θ utters a speech signal $s(n)$. The microphones receive the signal $x(n)$, $y(n)$ and $z(n)$ given by Eq.(1), respectively, that consists of $\tau_{\mathtt{x}}(\mathtt{x}=x,y,z)$ delayed speech and additive sensor noise signals $n_{\mathtt{x}}(n)(\mathtt{x}=x,y,z)$ that can be modeled as spatially uncorrelated. Here, $\tau_{\mathtt{x}}$ is signal arrival delay at microphone $\mathtt{x}$ with respect to the reference point located at the array origin o.

$$\mathtt{x}(n) = s(n-\tau_{\mathtt{x}}) + n_{\mathtt{x}}(n) \tag{1}$$

From such configuration, we can take three pairs of microphones that have equal distance D between microphones and each pair faces to different direction of every $\frac{\pi}{3}$[rad]. As we stated in [6], the cross spectrum of each microphone pair contains the speaker direction information in its phase term. So the aim of the proposed method is to realize the tracking by using the cross spectra derived from three different microphone pairs.

Here we assume next a) and b) for the input signal without loss of generality.

a) Only one speech signal is received.
At the situation of teleconference, it is usual to assume that more than one speaker do not speak simultaneously.

b) The location of the speaker is restricted on the array plane.

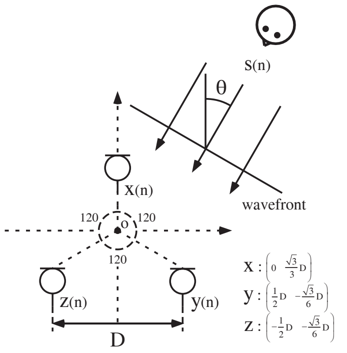
S(n)
θ
wavefront
x(n)
120
O
120
120
z(n)
y(n)
D
x : $\begin{pmatrix} 0 & \frac{\sqrt{3}}{3}D \end{pmatrix}$
y : $\begin{pmatrix} \frac{1}{2}D & -\frac{\sqrt{3}}{6}D \end{pmatrix}$
z : $\begin{pmatrix} -\frac{1}{2}D & -\frac{\sqrt{3}}{6}D \end{pmatrix}$

Figure 1: Model of input signal to the equilateral-triangular microphone array

3. PROPOSED METHOD

3.1 Model of input signal

The short-time Fourier transforms of each microphone input signals $x(n)$, $y(n)$ and $z(n)$ in Fig.1 are given by

$$\left\{ \begin{array}{l} X(\omega) = S(\omega)e^{-j\omega\tau_x} + N_x(\omega) \\ Y(\omega) = S(\omega)e^{-j\omega\tau_y} + N_y(\omega) \\ Z(\omega) = S(\omega)e^{-j\omega\tau_z} + N_z(\omega) \end{array} \right. , \qquad (2)$$

where $S(\omega)$ and $N_{\mathrm{x}}(\omega)$ are the Fourier transform of the speech $s(n)$ and noise $n_{\mathrm{x}}(n)(\mathrm{x} = x, y, z)$, respectively. Here we can define the cross spectra of three microphone pairs given by

$$\left\{ \begin{array}{l} G_{xy}^{(\omega)}(\theta) = E[X^*(\omega)Y(\omega)] = P_S(\omega)e^{-j\omega\tau_{xy}(\theta)} \\ G_{yz}^{(\omega)}(\theta) = E[Y^*(\omega)Z(\omega)] = P_S(\omega)e^{-j\omega\tau_{yz}(\theta)} \\ G_{zx}^{(\omega)}(\theta) = E[Z^*(\omega)X(\omega)] = P_S(\omega)e^{-j\omega\tau_{zx}(\theta)} \end{array} \right. , \qquad (3)$$

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. The delay constants in Eq.(3) are the function of θ given by

$$\tau_{xy}(\theta) = D\sin(\theta + \tfrac{2}{3}\pi)/c \qquad (4)$$

$$\tau_{yz}(\theta) = D\sin(\theta)/c \qquad (5)$$

$$\tau_{zx}(\theta) = D\sin(\theta - \tfrac{2}{3}\pi)/c \qquad (6)$$

where c denotes the sound velocity.

Because the power of speech signal 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 in the following process, we utilize the cross spectra at the harmonic frequencies ω_m selected by the higher SNR η_m than a threshold T[6]. Here ω_m is smaller than $\omega_{max} = \frac{\pi c}{D}$ determined by the spatial sampling theory.

3.2 Cross spectra integration

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

$$\tau_{x2y}(\phi) \equiv \tau_{yz}(\phi) - \tau_{xy}(\phi) = \sqrt{3}D\sin(\phi - \tfrac{\pi}{6})/c \qquad (7)$$

$$\tau_{z2y}(\phi) \equiv \tau_{yz}(\phi) - \tau_{zx}(\phi) = \sqrt{3}D\sin(\phi + \tfrac{\pi}{6})/c \qquad (8)$$

Then, we define the following *phase rotation factors* composed of the above phase compensating components with respect to the signal from direction ϕ.

$$G_{x2y}^{(\omega_m)}(\phi) \equiv e^{-j\omega\tau_{x2y}(\phi)} \qquad (9)$$

$$G_{z2y}^{(\omega_m)}(\phi) \equiv e^{-j\omega\tau_{z2y}(\phi)} \qquad (10)$$

Using these *phase rotation factors*, we define the following *integrated cross spectrum*.

$$G_{\phi,\theta}^{(\omega_m)} = G_{x2y}^{(\omega_m)}(\phi)G_{xy}^{(\omega_m)}(\theta) + G_{yz}^{(\omega_m)}(\theta) + G_{z2y}^{(\omega_m)}(\phi)G_{zx}^{(\omega_m)}(\theta) \quad (11)$$

Now for the $G_{\phi,\theta}^{(\omega_m)}$, following theorem[1] is satisfied.

[Theorem]

In general,

$$|G_{\phi,\theta}^{(\omega_m)}| \leq 3 \qquad (12)$$

and the equality is satisfied if and only if $\phi = \theta$.

From this theorem, we define the following non-negative performance index who takes its global minimum at $\phi = \theta$.

$$Q_{\phi,\theta}^{(\omega_m)} = 9 - \left|G_{\phi,\theta}^{(\omega_m)}\right|^2 \geq 0 \qquad (13)$$

Thus, the tracking problem results in searching ϕ that satisfies $Q_{\phi,\theta}^{(\omega_m)} = 0$.

3.3 Weighted steepest descent method using harmonics

For minimizing the nonlinear function $Q_{\phi,\theta}^{(\omega_m)}$, we derive the following steepest descent method given by

$$\phi_{i+1} = \phi_i - \frac{\rho}{M(i)} \sum_m \mu_m(i) \nu_m \frac{\partial Q_{\phi,\theta}^{(\omega_m)}}{\partial \phi}, \qquad (14)$$

where i and ρ are the number of iteration and the stepsize parameter, respectively, and $\mu_m(i)$ is the weight for frequency ω_m at i-th iteration. $M(i)$ denotes the number of weights μ_m whose value is not zero. Because $\partial Q/\partial\phi$ is proportional to ω, we put the normalizing factor ν_m defined as $\nu_m = \omega_{max}/\omega_m$.

Fig.2 shows the feature of Q at different frequency bands. We find several local minima at the higher band, but through some simulations, we have verified that the global convergence is guaranteed at the lower band $\omega_m \leq \frac{\omega_{max}}{4}$ without loss of generality. In contrast, the function around the correct DOA steeply decreases in the higher band, so that the convergence speed would be faster. We start to update using the performance index at the low frequency band, and use the higher band at the later part using the weight $\mu_m(i)$ that is switched by the convergence degree. Fig.3 shows the procedure flow to determine $\mu_m(i)$. The initial weight $\mu_m(0)$ is given as following

$$\mu_m(0) = \left\{ \begin{array}{ll} 1 & \omega_m \leq \omega_{max}/4 \\ 0 & \text{otherwise} \end{array} \right. , \qquad (15)$$

then μ_m for the next higher band is set to 1 if the following conditions are simultaneously satisfied.

$$\nu_m \left| \frac{\partial Q_{\phi_i,\theta}^{(\omega_m)}}{\partial \phi} \right| < T_{dQ} \qquad (16)$$

$$\frac{\partial^2 Q_{\phi_i,\theta}^{(\omega_m)}}{\partial \phi^2} > 0 \qquad (17)$$

While all the weights have been changed to 1, it means ϕ is sufficiently close to the convergent point, so we change the weights to the SNR at each harmonic to improve the accuracy of final estimation result.

$$\mu_m(i) = \frac{\eta_m}{\sum_m \eta_m} \qquad (18)$$

[1] The proof of this theorem is given in [6]

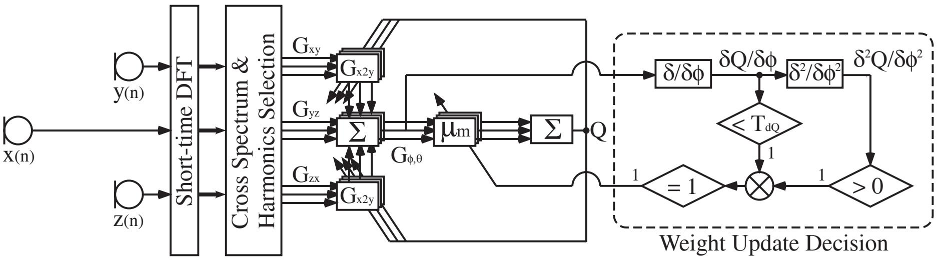

Figure 3: Flow diagram of the proposed method

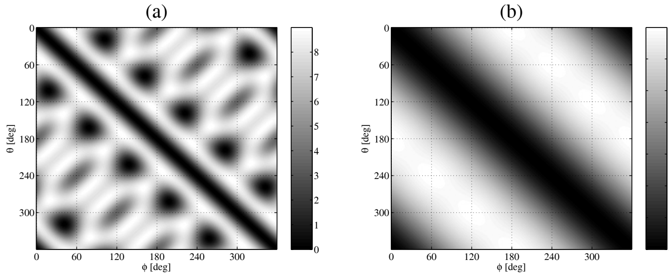

Figure 2: Performance index $Q_{\phi,\theta}^{(\omega_m)}$ at different frequencies (a):ω_{max} (b):$\omega_{max}/4$

Table 1: Parameters for simulation

Input SNR	20dB
Sampling Frequency	16000Hz
Wave Velocity c	340m/s
Microphone Distance D	0.08m
T	15dB
T_{dQ}	1
ρ	1
Window	Hamming
FFT point	2048
Frame Length	1024
Frame Overlap	768
Data Length	112ms

4. RESULTS OF SIMULATION AND EXPERIMENT

For the computer simulation and experiment in a real acoustic environment, 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. The array signal for computer simulation is virtually generated by delaying the signal with an appropriate delay samples according to θ and sum up with additive white noise as the sensor noise. As the conventional method for comparison, we adopt the method using null beamformer[3]. The parameters used in the simulation and experiment are shown in Tab.1. For quantitative evaluation, we use the deviation of estimation error (DEE) defined in [6].

4.1 Evaluation with computer simulation

Fig.4 is the DEEs at different speaker direction after 500 times iteration. We can confirm that the proposed method keeps the high and spatially uniform estimation accuracy at every speaker direction. Fig.5 shows the DEEs of the estimation after learning iteration of different times. The parameter always converges after the iteration of more than 200 times. For the calculation of DEEs, we took 10 trials for every data in these simulations.

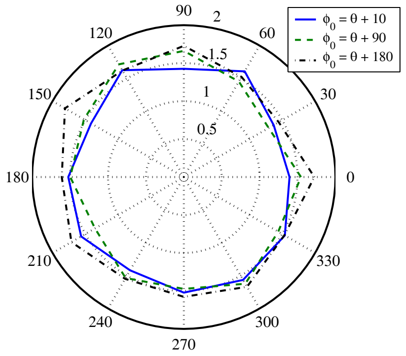

Figure 4: DEEs at the computer simulation

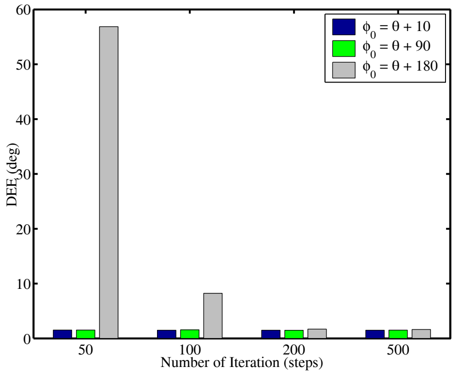

Figure 5: Relation between accuracy and number of iteration

4.2 Experiments at real acoustic environment

To verify that the proposed method is effective even at a real acoustic environment, we performed some experiments in a conference room (W×D×H : 18 × 15 × 4[m]). The speech data and parameters are the same as in the preceding computer simulation except for the input SNR lying around 18dB and the threshold T is 10dB. Now we have some experiments with the same conditions that adopted in the computer simulation with 5 trials for each data. At first, the convergence state of both the conventional and proposed method are shown in Fig.6 and Fig.7, respectively. Here the speaker is located at $\theta = 0$[deg] and the initial learning parameter ϕ_0 is settled at different directions. From the results, the proposed method guarantees the convergence at global minimum wherever the initial parameter is settled against the conventional method that converges at the nearest local minimum. The DEEs at experiment as shown in Fig.8 reveals that the proposed method is sufficiently effective even at the real acoustic environment.

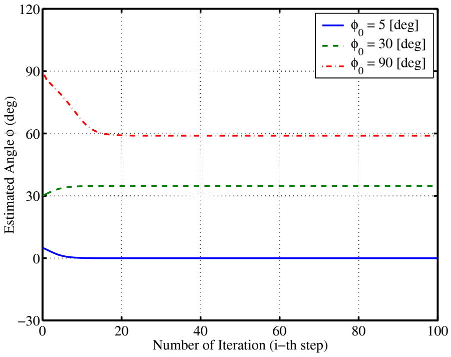

Figure 6: Learning state of the conventional method at experiment

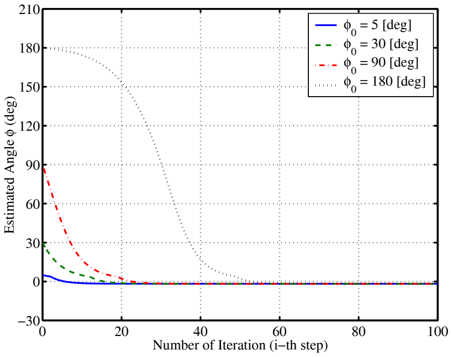

Figure 7: Learning state of the proposed method at experiment

4.3 Tracking examples for moving speakers

In this subsection, we show some tracking results for moving speakers. Fig.9 shows the simulation result of tracking 5 speakers located at different directions as given in Tab.2. Here we performed the proposed method with 200 learning iteration on each set of 4 frames, and the initial parameter ϕ_0 is settled at 60[deg]. From the result, the method promptly tracks the speaker's direction even he/she moves abruptly.

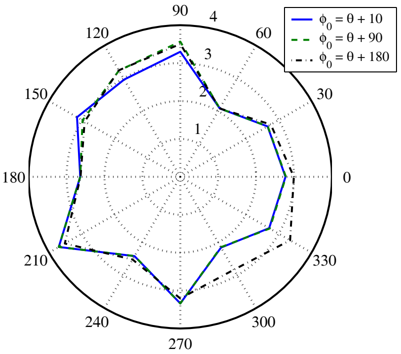

Figure 8: DEEs at the experiment

Table 2: Position of speakers

	Direction	Period (ms / frame set*)	Phoneme
Speaker 1	0	0 – 256 / 1 – 13	Female /a/
Speaker 2	180	266 – 512 / 14 – 29	Male /i/
Speaker 3	90	513 – 896 / 30 – 53	Female /u/
Speaker 4	240	897 – 1152 / 54 – 69	Male /e/
Speaker 5	300	1153 – 1664 / 70 – 98	Female /o/

* Number of frame set

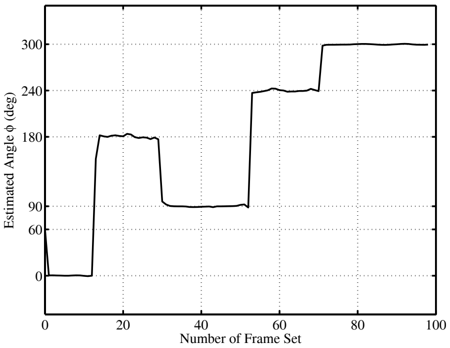

Figure 9: Tracking result of moving speakers

5. CONCLUSION

In this paper, we have proposed a new algorithm for speaker direction tracking. In the method, we find the speaker direction by minimizing the performance index derived by the three pairs of microphones in the equilateral-triangular microphone array, using the weighted steepest descent algorithm. Some computer simulation and experiment results show the effect of the method that it keeps uniform accuracy for omni-direction and it does not lose track of the speaker even if he/she moves abruptly. For a future subject, the tracking problem for 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.

REFERENCES

[1] G. Nokas and E. Dermatas, "Speaker Tracking for Hands-Free Continuous Speech Recognition in Noise Based on a Spectrum-Entropy Beamforming Method," in IEICE Trans. on Inf. & Syst., Vol.E86-D, No.4, pp.755–758, Apr. 2003.

[2] Y. Nagata and M. Abe, "Two-Channel Adaptive Microphone Array with Target Tracking," in IEICE Trans. in Fundamentals, Vol.J82-A, No.6, pp.860–866, Jun. 1999. (in Japanese)

[3] H. Kawakami, M. Abe, and M. Kawamata, "A Two-Channel Microphone Array with Adaptive Target Tracking Using Frequency Domain Generalized Sidelobe Cancellers," in IEEE Int. Symp. on Intelligent Sign. Process. & Commun. Systems, pp.291–296, Nov. 2002.

[4] K. Suyama and T. Tasaki, "A Study on Target Talker Tracking via Two Microphones," in Proc. 18th IEICE DSP Symposium, A5-6, Nov. 2003. (in Japanese)

[5] M. Zhang and M. H. Er, "An alternative Algorithm for Estimating and Tracking Talker Location by Microphone Arrays," in J. Audio Eng. Soc., Vol.44, No.9, pp.729–736, Sep. 1996.

[6] Y. Hioka and N. Hamada, "DOA Estimation of Speech Signal Using Microphones Located at Vertices of Equilateral Triangle," in IEICE Trans. on Fundamentals, Vol.87-A, No.3, Mar. 2004. (Accepted for Publication)