ICSV14
Cairns • Australia
9-12 July, 2007

DIRECT ALGEBRAIC METHOD FOR SOUND SOURCE LOCALIZATION WITH FINEST RESOLUTION BOTH IN TIME AND FREQUENCY

Shigeru ANDO, Nobutaka ONO, and Takaaki NARA

Department of Information Physics and Computing, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
E-mail: ando@alab.t.u-tokyo.ac.jp

Abstract

We deal with the problem of direction and distance estimates of sound sources in 3-D space and arrive at a new exact and direct algorithm from finite observations both in space and time. We first derive a partial differential equation (PDE) which we call the sound source constraint (SSC). We show that the general solution of the SSC-PDE is a diverging spherical wave from a point source with arbitrary temporal waveform. The SSC enables the observer to determine the source location (distance R and direction $\boldsymbol{n}$) from local measurements of the wavefield. For the measurements of wavefield, we consider weighted temporal integrals of arrayed microphone outputs in a finite duration. We obtain exact formulae for localizing a single source from single weight measurements and multiple sources from the combination of differently weighted measurements. We examine the performance by simulating non-stationary complex multi-source environments.

1. INTRODUCTION

Directional sensing of sounds enables the localization of their sources in space. Sound source localization is one of the most important functions of auditory systems. It can aid in the separation of signals from multiple sources and in their identification. Applications include the localization and tracking of speakers in conference rooms, improved hearing aids having directional sensitivity, and the realization of ears of mobile robots to endow them with localizing, separating, and communicating capability[1] even for moving sources and microphones[2]. Most localization methods depend on two types of physical variables derived from sensor signals: time delay of arrival (TDOA) and direction of arrival (DOA). DOA can be estimated by exploiting the phase difference at sensors and is applicable to narrow band sources. Typical algorithms for multiple sources include ESPRIT[4] and MUSIC[3] based on the subspace decomposition of covariance matrix of received signals. TDOA is applicable to broadband source and relies mostly on on the accurate measurement of time delays between pairs of microphones based on the cross correlation. Consistent triangulation of all delays produce the source distribution[5, 6].

However the problems of existing methods are in several aspects. Firstly they are based on statistical quantities (covariance of arrayed sensors or correlation functions of received signals), hence long observation duration is required for their stability. Otherwise the result becomes noisy with numerous spurious sources. It also worsens the mixture condition of multiple sources both in time and frequency, and makes incapable to use inherent granular structure in these domains of most sound sources like human voice and environmental noise. Another problem is that no direct algorithm is available for conventional types of microphone array. Iterations for global optimization or an entire sweep of sound field are necessary. This requires the use of massive and costly computing power, which is not suitable for implementation in most practical systems and applications.

The purpose of this study is therefore to develop an algorithm with following properties and show its significance in complex, noisy, and reverberant environments being common in realworld-dapplications. The first issue is on the usage of data. The algorithm should not rely on statistical quantities among them but process the waveform data directly. The second issue is on the exactness of algorithm for finite observation. By it, the use of any short observation duration is possible without concerning errors due to theoretical approximation. The third issue is on the directness of the algorithm. An explicit, closed-form algorithm provides the solution directly and efficiently. This excludes massive and costly hardwares, and makes much easier its implementation in tiny sensor nodes. In the next section, we begin with the fundamental equation in the form of PDE, and integrate it to obtain an algebraic algorithm.

2. SOUND SOURCE CONSTRAINT PDE

Let the location vectors of the observation point P and an only one source S be $\boldsymbol{r} = (x, y, z)$ and $\boldsymbol{r}_0 = (x_0, y_0, z_0)$, respectively. Let the sound pressure of the source be $g(t)$. Then, from the general spherically symmetric solution of the wave equation, the pressure on P is expressed as

$$f(\boldsymbol{r}, t) = \frac{1}{|\boldsymbol{r} - \boldsymbol{r}_0|} g(t - \frac{|\boldsymbol{r} - \boldsymbol{r}_0|}{c}), \tag{1}$$

where c is the sound velocity, and

$$|\boldsymbol{r} - \boldsymbol{r}_0| = \{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2\}^{1/2}. \tag{2}$$

Let the unit direction vector from P to S be $\boldsymbol{n}$. The pressure gradient at P is expressed as

$$\nabla f(\boldsymbol{r}, t) = \frac{1}{|\boldsymbol{r} - \boldsymbol{r}_0|^2} g(t - \frac{|\boldsymbol{r} - \boldsymbol{r}_0|}{c})\, \boldsymbol{n} + \frac{1}{c|\boldsymbol{r} - \boldsymbol{r}_0|} \dot{g}(t - \frac{|\boldsymbol{r} - \boldsymbol{r}_0|}{c})\, \boldsymbol{n}, \tag{3}$$

and the temporal gradient is

$$\dot{f}(\boldsymbol{r}, t) = \frac{1}{|\boldsymbol{r} - \boldsymbol{r}_0|} \dot{g}(t - \frac{|\boldsymbol{r} - \boldsymbol{r}_0|}{c}). \tag{4}$$

Therefore, substituting Eq.(1),(4) to Eq.(3) to eliminate the source waveform $g(t)$, we obtain an equation of only the pressure and its gradient at the observation point P as [7, 8]

$$\begin{aligned} \nabla f(\boldsymbol{r}, t) &= \{\frac{1}{|\boldsymbol{r} - \boldsymbol{r}_0|} f(\boldsymbol{r}, t) + \frac{1}{c} \dot{f}(\boldsymbol{r}, t)\}\, \boldsymbol{n} \\ &= \{\frac{1}{R} f(\boldsymbol{r}, t) + \frac{1}{c} \dot{f}(\boldsymbol{r}, t)\}\, \boldsymbol{n}, \end{aligned} \tag{5}$$

where $R \equiv |\boldsymbol{r} - \boldsymbol{r}_0|$ is the distance from P to S. We hereafter call this equation the sound source constraint (SSC) PDE. For the SSC-PDE, we can verify the following property.

Theorem 1 *The general solution of the SSC is a spherical wave with arbitrary temporal waveform, which is emitted from a source* $(R, \boldsymbol{n})$.

This theorem assures that the SSC-PDE provides both necessary and sufficient description of the sound field that enables to determine the sound source. An only freedom left in the SSC-PDE is the waveform. For the desired location information, it constrains it determinately. Therefore the next subject is to obtain an algebraic equation that relates the measurement quantities and the source location. This requires anyways an integration of the SSC-PDE.

3. INTEGRAL SSC-PDE AND ITS SOLUTION

We consider the integration in time axis. The gradient in spatial axis is left as it is. We assume they are provided by one of the gradient measurement techniques [10, 11, 12, 13, 19, 20].

3.1 Integral form of SSC-PDE

Assume that the location of sound source is stationary during a finite observation interval $[-T/2, T/2]$ (T is the observation time e.g. 2ms, 2.9ms, or 32ms in the experiments of section V). The sound waveform is arbitrary including its on/off. Then the SSC-PDE is satisfied anywhere in $[-T/2, T/2]$. In this case, we can invoke the identity relation with an arbitrary weighting function $w(t)$ as

$$\begin{aligned} &\nabla f - (\frac{1}{R} f + \frac{1}{c}\dot{f})\boldsymbol{n} = \vec{\boldsymbol{0}} \quad {}^{\forall}t \in [-\frac{T}{2}, \frac{T}{2}] \\ &\quad \leftrightarrow \int_{-T/2}^{T/2} \{\nabla f - (\frac{1}{R} f + \frac{1}{c}\dot{f})\boldsymbol{n}\} w(t) dt \quad {}^{\forall}w(t), \end{aligned} \tag{6}$$

where the variables $(\boldsymbol{r}, t)$ of f are omitted for brevity. Actually, the weighted integral form in the second line become identical with the SSC-PDE in the first line when we choose $\{w(t)\}$ as a complete set of function in $[-T/2, T/2]$ and solving them simultaneously.

Using this relation, we can integrate the SSC-PDE in time axis as follows. First, we introduce the complex exponential functions $\{e^{-j\omega t}\}$ as the set of weighting functions. Why we choose $e^{-j\omega t}$ is readily clarified in the course of derivation. The measurement quantities are the weighted integrals of the sound field at and near the observation point $\boldsymbol{r}$ as

$$g_\omega(\boldsymbol{r}) \equiv \int_{-T/2}^{T/2} f(\boldsymbol{r}, t) e^{-j\omega t} dt. \tag{7}$$

Integration of the SSC-PDE with the weight function $e^{-j\omega t}$ in $[-T/2, T/2]$ yields

$$\int_{-T/2}^{T/2} \{\nabla f(\boldsymbol{r}, t) - (\frac{1}{R} f(\boldsymbol{r}, t) + \frac{1}{c}\dot{f}(\boldsymbol{r}, t))\boldsymbol{n}\} e^{-j\omega t} dt$$

$$
\begin{aligned}
&= \nabla \int_{-T/2}^{T/2} f(\boldsymbol{r},t)e^{-j\omega t}dt - \frac{\boldsymbol{n}}{R}\int_{-T/2}^{T/2} f(\boldsymbol{r},t)e^{-j\omega t}dt - \frac{\boldsymbol{n}}{c}\int_{-T/2}^{T/2} e^{-j\omega t}\partial_t f(\boldsymbol{r},t)dt \\
&= \nabla g_\omega(\boldsymbol{r}) - \frac{\boldsymbol{n}}{R}g_\omega(\boldsymbol{r}) - \frac{\boldsymbol{n}}{c}\big[f(\boldsymbol{r},t)e^{-j\omega t}\big]_{-T/2}^{T/2} - \frac{j\omega\boldsymbol{n}}{c}\int_{-T/2}^{T/2} f(\boldsymbol{r},t)e^{-j\omega t}dt \\
&= \nabla g_\omega(\boldsymbol{r}) - \frac{\boldsymbol{n}}{R}g_\omega(\boldsymbol{r}) - \frac{\boldsymbol{n}}{c}\big[f(\boldsymbol{r},t)e^{-j\omega t}\big]_{-T/2}^{T/2} - \frac{j\omega\boldsymbol{n}}{c}g_\omega(\boldsymbol{r}) \\
&= \vec{\mathbf{0}},
\end{aligned}
$$

where $\partial_t \equiv \partial/\partial t$ for brevity. It should be noted that the weighted integral with the weight $\partial_t e^{-j\omega t}$ becomes equal to the one with $e^{-j\omega t}$ except for the coefficient $-j\omega$ by choosing the complex exponential functions as the weight. This significantly reduced the number of measurement quantities. Further more, the integral boundary term $\big[f(\boldsymbol{r},t)e^{-j\omega t}\big]_{-T/2}^{T/2}$ can also be eliminated by choosing $\omega T = 2n\pi$ $(n = 0, 1, 2, \cdots)$ which is actually the orthogonal complete condition of $\{e^{-j\omega t}\}$. It is because

$$
\begin{aligned}
\big[f(\boldsymbol{r},t)e^{-j\omega t}\big]_{-T/2}^{T/2} &= f(\boldsymbol{r},T/2)e^{-j\pi n} - f(\boldsymbol{r},-T/2)e^{j\pi n} \\
&= (-1)^n(f(\boldsymbol{r},T/2) - f(\boldsymbol{r},-T/2)) \\
&= (-1)^n\big[f(\boldsymbol{r},t)\big]_{-T/2}^{T/2},
\end{aligned}
$$

which shows the integral boundary term for all weight frequencies $\omega = 2n\pi/T$ $(n = 0, 1, 2, \cdots)$ are equal except for the sign $s_n \equiv (-1)^n$. Therefore we obtain a complex vector equation

$$
\nabla g_\omega(\boldsymbol{r}) = \{\big(\frac{1}{R} + \frac{j\omega s_n}{c}\big)g_\omega(\boldsymbol{r}) + \frac{s_n}{c}\big[f(\boldsymbol{r},t)\big]_{-T/2}^{T/2}\}\boldsymbol{n} \tag{8}
$$

with unknown variables $\boldsymbol{n}, R, \big[f(\boldsymbol{r},t)\big]_{-T/2}^{T/2}$.

We call this equation the (temporal) integral form of sound source constraint (iSSC). The iSSC, as well as the SSC-PDE, provides an exact relation to determine the source location. Furthermore, it provides an algebraic relation for a direct solution based on the weighted integral measurements in ever so a small interval $[-T/2, T/2]$. Indeed the weighted integral is the well-known Fourier series coefficient, but no smooth window function (except for the do-nothing window) is required to obtain them.

3.2 Linearized algorithm

Since the iSSC has $2 + 1 + 1 = 4$ non-observable degrees of freedom (dof) whereas the number of independent equations are 6 or 4 for 3-D or 2-D gradient measurement of $\nabla g_\omega(\boldsymbol{r})$, we can essentially solve for the the source location using only one complex vector equation for a single weight frequency ω.

However, Eq.(8) is nonlinear for $R, \big[f(\boldsymbol{r},t)\big]_{-T/2}^{T/2}$ and $\boldsymbol{n}$ because of their multiplications involved. To obtain rapidly an initial guess, desired is the linearization of Eq.(8) based on the extension of unknown variables into $\boldsymbol{n}/R, \boldsymbol{n}/c, \boldsymbol{n}\big[f(\boldsymbol{r},t)\big]_{-T/2}^{T/2}$. In this case, the number of unknowns is $3 + 3 + 3 = 9$ for three-dimensional (3-D) gradient measurement and $2 + 2 + 2 = 6$ for 2-D

gradient measurement. This means two frequencies are sufficient for both cases to localize a single source.

Two simple cases are the use of DC frequency and adjacent frequency as one of the paired frequencies. For the DC frequency case, since $s_0 = 1$, it follows that

$$\nabla g_0(\boldsymbol{r}) = \{\frac{1}{R} g_\omega(\boldsymbol{r}) + \frac{1}{c}\big[f(\boldsymbol{r},t)\big]_{-T/2}^{T/2}\}\boldsymbol{n}$$

hence, eliminating $\big[f(\boldsymbol{r},t)\big]_{-T/2}^{T/2}$, we obtain simply as

$$\nabla(g_\omega(\boldsymbol{r}) - s_n g_0(\boldsymbol{r})) \;=\; \{\frac{1}{R}(g_\omega(\boldsymbol{r}) - s_n g_0(\boldsymbol{r})) + \frac{j\omega}{c} g_\omega(\boldsymbol{r})\}\boldsymbol{n}. \tag{9}$$

For the use of adjacent frequencies ω_1, ω_2, it follows that

$$\nabla(s_2 g_{\omega 1} - s_1 g_{\omega 2}) \;= \{\frac{s_2}{R} g_{\omega 1} - \frac{s_1}{R} g_{\omega 2} + \frac{j\omega_1 s_2}{c} g_{\omega 1} - \frac{j\omega_2 s_1}{c} g_{\omega 2}\}\boldsymbol{n}, \tag{10}$$

where s_1, s_2 are the values of s_n for ω_1, ω_2 respectively. In both equations, the integral boundary terms are eliminated and $\boldsymbol{n}/R, \boldsymbol{n}/c$ can be solved directly. Succeeding nonlinear least squares estimation from $\boldsymbol{n}/R, \boldsymbol{n}/c$ (4 dof) to $R, \boldsymbol{n}$ (3 dof) reduces the noise and improves the accuracy.

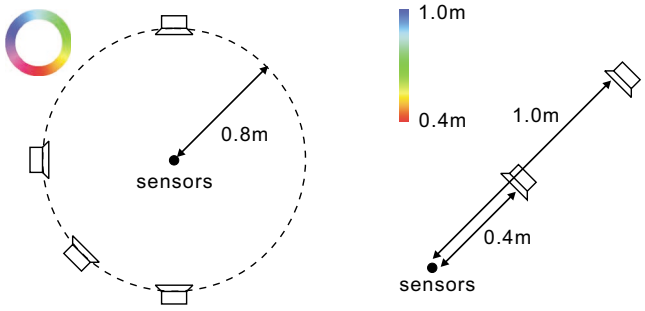
4. FINE GRAIN LOCALIZATION IN TIME AND FREQUENCY

Most established methods for the multiple sound source localization, e.g. the eigenspace methods[3, 4], require long observation time to obtain narrow frequency bands and well-converged statistical quantities, which can reduce significantly the temporal resolution. Recently, another approach appears that relies on the time-frequency sparseness of sound power distributions to separate their sources. For example, the power of speech sounds is concentrated on some fragments of the time-frequency domain because of the harmonic structure, formants, glottal closures, etc. Where the source signals have sparseness, they are rarely overlapped thus one source is only active almost everywhere. This means in a small time-frequency, i.e. a grain, single source localization is sufficient. Accumulation of those results yields the multiple localization with the finest granularity both in time and space.

One important point of this approach is that the nature of sparseness can change source by source. For impulsive sounds like mechanical noise, high temporal resolution (short frame length) is appropriate. Some kind of slowly changing periodic noise are, however, isolated well by a narrow band analysis (long frame length). In the short time Fourier transform (STFT) analysis for speech, the optimal frame length is about 1024 sample for the 16kHz sampling frequency[21] because of its moderately changing nature. The PDE-based framework that provides always an exact formula irrespective of the length of analysis frame is therefore suitable inherently for the sparseness-based multiple source localization. To adjust the time-frequency resolution to the target signal, we can use the following relations

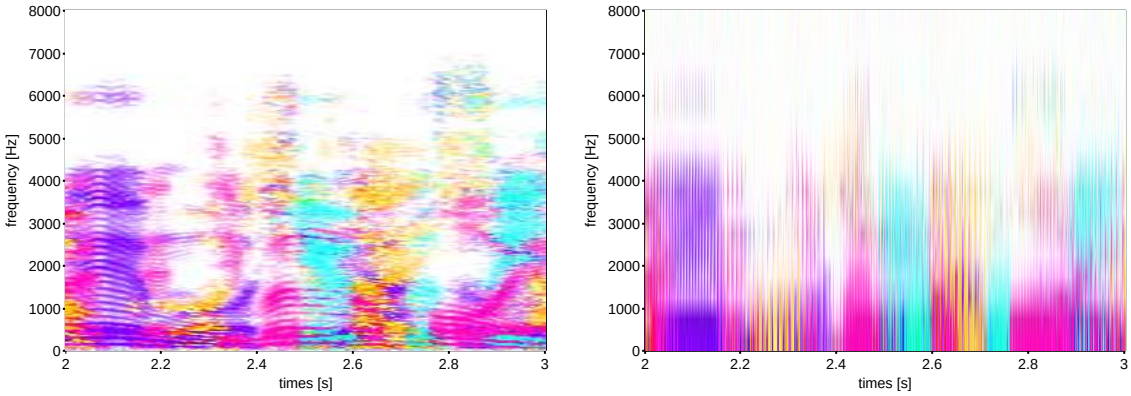
$$\Delta\omega = 2\pi/T, \quad \Omega = 2\pi/\Delta t, \tag{11}$$

where T is the time resolution of time-varying spectra (observation interval), $\Delta\omega$ is the frequency resolution, and Ω is the maximal analysis frequency of sounds.



(A) (B)

Figure 1: Two experimental setups and the color samples for display of results.



(a) 512 points STFT (T =32ms) (b) 32 points STFT (T =2ms)

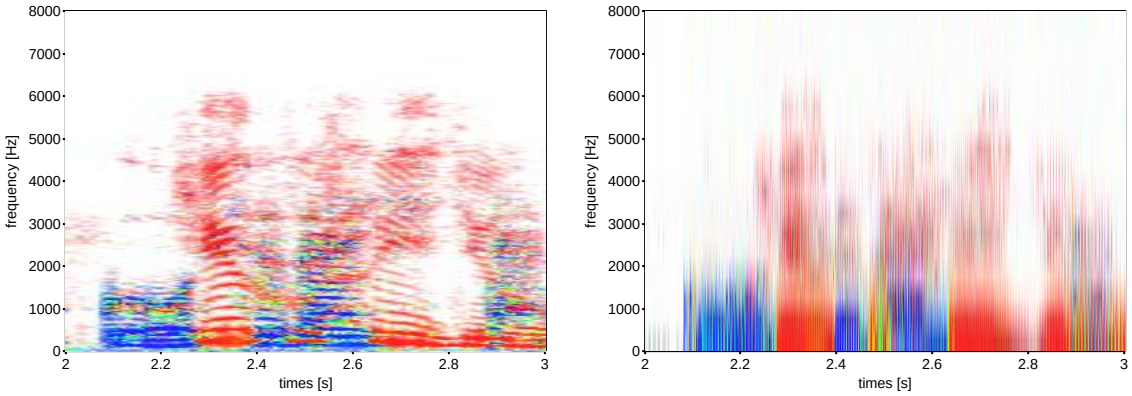
Figure 2: The estimated source direction maps in time-frequency plane for (a) setup A with the 512 points (32ms) STFT, and for (b) setup A with the 32 points (2ms) STFT. The source direction in each grain is indicated by a color shown by Fig.2 left circle.

5. NUMERICAL EVALUATION

Fig. 1 shows simulated two environments for experiments of sound source localization. In the setup A, four sources are located in the same distance and the different directions. While in the setup B, two sources are located in the different distances and the same direction. In both of them, source signals were different speech utterances sampled by 16kHz. Spatial gradients were obtained by the difference of observed signals at 5cm-spaced points. The sound source direction and distance were estimated at every time-frequency bin by our method. For each setup, narrow band and wide band analysis were examined.

Fig. 2(a) and Fig. 2(b) show the estimated source directions for the setup A as color maps. They were obtained through 512 or 32 points STFT, respectively. Mainly, due to the harmonic structure in the narrow band analysis, or the temporal pitch structure in the wide band analysis, the energies of the speech utterances are sparsely distributed and rarely overlapped. Thus, in the two figures, four colors corresponding to the four sources are clearly dominant, which indicates that the sources are successfully localized. Apart from differences of time and frequency resolutions,

the estimated directions are almost consistent in the two figures. Fig. 3(a) and Fig. 3(b) also show the estimated distances for the setup B in the same way. It illustrates that the discrimination by the source distance is well performed.



(a) 512 points STFT (T =32ms) (b) 32 points STFT (T =2ms)

Figure 3: The estimated source distance maps in time-frequency plane for (a) setup B with the 512 points (32ms) STFT, and for (b) setup B with the 32 points (2ms) STFT. The source direction in each grain is indicated by a color shown by Fig.2 right bar.

Actually, the optimal time and frequency resolution for the sparse representation depends on statistical properties of target, noise, or surrounding environment like reverberation. However, in the conventional sparseness-based localization, the high frequency resolution is essential to estimate time-delay from phase difference, which limits the performance of the sparseness-based localization. While in our method, arbitrary resolution is allowable as shown in this experiments because of the rigorous formulation of finite observation. The advantage should yield the capability of the robust localization in the real environment.

6. SUMMARY

We proposed a novel algorithm of sound source localization with following properties: 1) use of not statistical but wideband waveform data directly, 2) exactness of algorithm for any finite and short observation, and 3) directness of the algorithm without iterations or DOA sweeps. We examined the algorithm with numerically simulation and an actual data taken in a room with significant reverberation.

References

[1] Y. Senjo, K. Miyamoto, T. Kurihara, and S. Ando, “A novel coupling scheme of speech separation and face recognition: Implementation with “SmartHead” autonomous vision sensor with ears,” Proc. Int. Workshop Networked Sensing Systems (INSS 2006), Chicago, 2006.

[2] T. Horiuchi, M. Mizumachi, and S. Nakamura, “Iterative estimation and compensation of signal direction for moving sound source by mobile microphone array,” IEICE Trans. Fundamentals, vol.E87-A, no.11, pp.2950-2956, 2004.

[3] R. O. Schmidt, “Multiple Emitter Location and Signal Parameter Estimation,” IEEE Trans. Antennas and Propagation, vol.34, no.3, 1986.

[4] A. Paulraj and R. R. Kailath, “ESPRIT-Estimation of Signal Parameters via Rotational Invariance Techniques,” IEEE Trans. Acoust., Speech, Signal Processing, vol.37, no.7, pp.984-995, 1989.

[5] J. O. Smith and J. S. Abel, “Closed-form least-squares source location estimation from range-difference measurements,” IEEE Trans. Acoust., Speech, Signal Processing, vol.35, no.12, pp.1661-1669, 1987.

[6] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, “A closed-form location estimation for use with room environment microphone arrays,” IEEE Trans. Speech and Audio Processing, vol.5, no.1, pp.45-50, 1997.

[7] S. Ando and N. Ono, “Partial differential equation (PDE)-based theory of sound source localization,” 4th Joint Meeting ASA and ASA, Honolulu, 2006.

[8] Y. Fujita, N. Ono, and S. Ando, “Partial-differential-equation-based sound source localization: Finite Fourier integral approach and its application to multiple source localization,” 4th Joint Meeting ASA and ASJ, Honolulu, 2006.

[9] T. Nara and S. Ando, “Projective Method for the Inverse Source Problem of the Poisson Equation,” Inverse Problems, vol.19, no.2, pp.355-370, 2003.

[10] S. Ando, H. Shinoda, K. Ogawa, and S. Mitsuyama, “A Three-Dimensional Sound Localization System Based on the Spatio-Temporal Gradient Method,” Trans. Soc. Instrumentation and Control Engineers, vol.29, no.5, pp.520-528, 1993. (in Japanese)

[11] S. Ando, “An Intelligent Three-Dimensional Vision Sensor with Ears,” Sensors and Materials, vol.7, no.3, pp.213-231, 1995.

[12] S. Ando, “An Autonomous Three-Dimensional Vision Sensor with Ears,” Trans. IEICE, vol.E78-D, no.12, pp.1621-1629, 1995.

[13] S. Ando and H. Shinoda, “Ultrasonic Emission Tactile Sensing,” IEEE Control Systems, vol.15, no.1, pp.61-69, 1995.

[14] N. Ono and S. Ando, “Sound Source Localization Sensor with Mimicking Barn Owls,” Proc. Transducers’01, vol.2, pp.1654-1657, Munich, Jun. 2001.

[15] N. Ono, T. Hirata, and S. Ando, “A Study on Sound Source Localization Sensor with Mimicking Ormia Ochracea’s Ears,” Tech. Digest 18th Sensor Symp., pp.351-354, May 2001.

[16] N. Ono, Y. Zaitsu, T. Nomiyama, A. Kimachi and S. Ando, “Biomimicry Sound Source Localization with Fishbone,” Trans. IEEJ(E), vol.121-E, no.6, pp.313-319, Jun. 2001.

[17] M. Kurihara, N. Ono, and S. Ando, “Theory and experiment of dual sound source localization with five proximate microphones,” Proc, SICE Annual Conference, pp.353-354, 2002.

[18] N. Ono, A. Saito, and S. Ando, “Bio-mimicry sound source localization with gimbal diaphragm,” Proc. 19th Sensor Symp., pp.441-446, Kyoto, May 2002.

[19] N. Ono, A. Saito, and S. Ando, “Bio-mimicry Sound Source Localization with Gimbal Diaphragm,” Trans. IEEJ, vol.123-E, no.3, pp.92-97, 2003.

[20] N. Ono, A. Saito, and S. Ando, “Design and Experiments of Bio-mimicry Sound Source Localization Sensor with Gimbal-Supported Circular Diaphragm,” Proc. 12th Int. Conf. Solid State Sensors and Actuators (Transducers’03), pp.939-942, 2003.

[21] O. Yilmaz and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking,” IEEE Trans. Signal Processing, vol. 52, no. 7, pp 1830-1847, 2004.