Sound Source Localization by Asymmetrically Arrayed 2ch Microphones on a Sphere

Nobutaka Ono*, Soichiro Fukamachi*, Takuya Nishimoto* and Shigeki Sagayama*
* Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo, 113-8656
Email: {onono, s-fukamachi, nishi, sagayama}@hil.t.u-tokyo.ac.jp

***Abstract*— In this paper, we propose a novel system to localize a sound source in any 2D directions using only two microphones. In our system, the two microphones are asymmetrically placed on a sphere, thus, 1) the diffraction by the sphere and the asymmetrical arrangement of the microphones yield the localization cue including the front-back judgment, and 2) unlike the dummy head system, no previous measurements are necessary due to the analytical representation of the sphere diffraction. To deal with reverberation or ambient noises, we consider the maximum likelihood estimation of the direction of arrival with a diffused noise model on a sphere. We present a real system that we built through the investigation of the optimal microphone arrangement for speech, and give experimental results in real environment.**

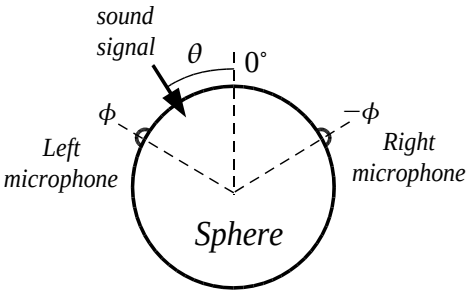

Fig. 1. Arrangement of the microphones

I. Introduction

Localization of sound source is very useful in various acoustic applications such as target tracking, environment monitoring, speaker indexing, and so on. Most of man-made systems are omni-directional pressure-sensitive microphones arrayed in free space, where three or more microphones are essential to localize the 2D direction since a pair of microphones placed in free space has an intrinsic axial symmetry, which brings a front-back ambiguity to the 2D localization.

On the other hand, many animals including human have the capability of localizing a sound source from any directions with only two ears [1]. The key is to exploit the frequency characteristic of reflection or diffraction by the pinna or the head. By mimicking the auditory systems, several previous works aim at realizing such abilities as monaural sound source localization [2], [3], [4], DOA estimation of elevation and azimuth using asymmetric reflectors like the ones found in barn owls [5], or human-like localization based on a dummy head system [7], [8]. Especially, the development of sound source localization by 2ch array should enrich PC applications since it can be easily installed on a PC through the standard audio input. It would also facilitate the sound source separation based on a stereo signal [9], [10].

In this paper, inspired from the human auditory mechanism, we propose a novel system to localize a sound source in any 2D directions using only two microphones. In our system, two microphones are asymmetrically arrayed on a sphere, where, instead of the pinna, the diffraction by the sphere and the asymmetrical arrangements of the microphones yield the localization cue including the front-back judgment. Unlike the dummy head, our system doesn't require consuming HRTF measurements due to the analytical representation of the diffraction by a sphere. A pilot work has been done by Handzel et al., where they examined through numerical simulations the asymmetric arrangement of two microphones on a sphere and a localization algorithm based on the metric of the interaural level and phase difference [11]. In this paper, to deal with speech sources and real environment including reverberation or ambient noises, we consider the maximum likelihood estimation of the direction of arrival with a diffused noise model on a sphere. Through the investigation of the optimal microphone arrangement for speech, we built a prototype system using a wooden ball with a 30mm radius. We describe the system and give experimental results obtained with it in real environment.

II. Investigation of Localization Possibilities

Suppose that two microphones are placed at angles of $\pm\phi$ on a sphere shown in Fig. 1. When a plane wave arrives from a direction θ, the observed signals $\boldsymbol{M}(\omega) = (M_L(\omega), M_R(\omega))^T$ can be written as

$$\boldsymbol{M}(\omega) = S(\omega)\boldsymbol{H}(\omega,\theta) + \boldsymbol{N}(\omega), \tag{1}$$

where $S(\omega)$ is the arriving source signal and $\boldsymbol{N}(\omega)$ the observation noise. The frequency characteristics $\boldsymbol{H}(\omega,\theta)$ depending on the direction of arrival can be written

$$\begin{aligned}\boldsymbol{H}(\omega,\theta) &= (H_L(\omega,\theta)\ H_R(\omega,\theta))^T \\ &= (D(\omega,\theta-\phi)\ D(\omega,\theta+\phi))^T .\end{aligned} \tag{2}$$

D in eq. (2) is called the diffraction coefficient, which can be expressed analytically as [12]

$$D(\omega,\psi) = \frac{1}{ka}\sum_{n=0}^{\infty}\frac{j^{(n+1)}(2n+1)P_n(\cos\psi)}{ka\ h_{n+1}^{(2)}(ka) - nh_n^{(2)}(ka)}, \tag{3}$$

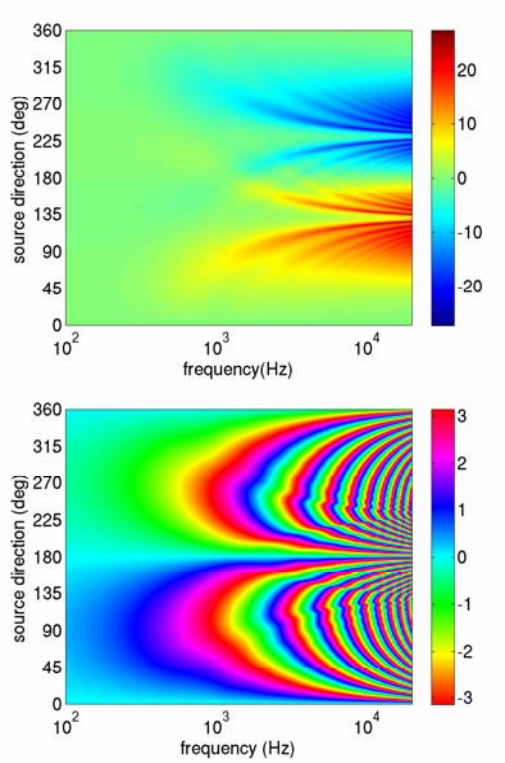

Fig. 2. Interchannel intensity ratio in [dB] (upper) and phase difference in [rad] (lower) for an asymmetrical arrangement ($\phi = 50°$) on a sphere of radius 85mm

where $P_n(x)$ is the Legendre function, $h_n^{(2)}(x)$ is the spherical Hankel function of the second kind, a is the radius of the sphere, ψ the direction of arrival, and $k = \omega/c$ where c is the sound velocity.

Even if $\boldsymbol{N}(\omega) = 0$ in eq. (1), $H_L(\omega, \theta)$, $H_R(\omega, \theta)$ cannot be directly observed due to the unknown source term $S(\omega)$. Thus, in sound source localization, the Inter-channel Transfer Function (ITF) defined by their ratio as

$$\mathrm{ITF}(\omega, \theta) = \frac{H_L(\omega, \theta)}{H_R(\omega, \theta)} = \frac{D(\omega, \theta - \phi)}{D(\omega, \theta + \phi)} \tag{4}$$

is the significant quantity, which is theoretically independent from $S(\omega)$. The pair of interchannel intensity and phase differences, which is often used as a cue of localization in a 2ch array, is equivalent to the ITF. For a localization without ambiguity, a one-to-one correspondence between the ITF and the source directions is necessary.

First, we show in Fig. 2 the amplitude and the phase of the ITF for a sphere of radius 85mm and an arrangement corresponding to $\phi = 50°$. Despite the absence of pinna-like structures there, we can see that the significant amplitude difference between both channels is introduced only by the diffraction of the sphere.

To explore the relationship between the ITF and the source direction, we show several loci of the ITF in the complex plane for the symmetric arrangement ($\phi = 90°$) and an asymmetric arrangement ($\phi = 50°$) at 1kHz, 2kHz, and 3kHz in Fig. 3.

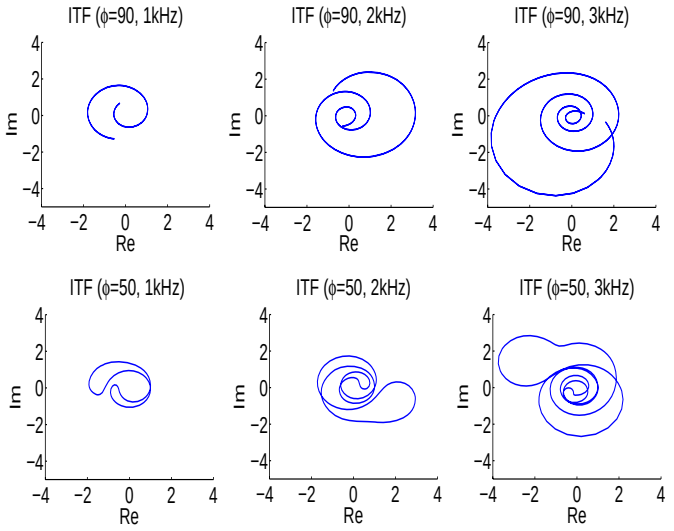

Fig. 3. Loci of ITF for 360° sound direction in the complex plane for the symmetric arrangement ($\phi = 90°$) (upper) and an asymmetric arrangement ($\phi = 50°$) (lower)

In the symmetric case, the loci are curves with two endpoints, which means that the ITF is completely overlapped between the front and back directions and these cannot be distinguished (the two endpoints correspond to $\theta = 90°$ and $\theta = 270°$). While in the asymmetric case, they depict closed loops with a few cross-points, which means that most directions can be localized. The ambiguity of the directions corresponding to the cross-points will be solved by integration of the localization results over frequencies. Note that all loci have a common cross-point at ITF$= 1$, which corresponds to $\theta = 0°$ and $\theta = 180°$. Our system thus cannot distinguish only these two directions because of the intrinsic left-right symmetry.

III. Localization Algorithm Based on ML Estimation and Diffused Sound Field Model

Because actual observations include noise, the ITF obtained from an observation $\boldsymbol{M}(\omega)$ does not coincide with the theoretical values shown in Fig. 3. Furthermore, the reliability of each frequency band should be different depending on each SN ratio. One reasonable way to perform localization in such a case is to apply maximum likelihood estimation based on some stochastic model of the observation noise [13]. Assuming that $\boldsymbol{N}(\omega)$ follows a complex-valued Gauss distribution with mean 0 and covariance matrix $V(\omega)$, the logarithmic likelihood is given by

$$\begin{aligned} \log p(\boldsymbol{M}(\omega); S(\omega), \theta) &= -\frac{1}{2}\log(2\pi|\det V(\omega)|) \\ &\quad -\frac{1}{2}\boldsymbol{E}(\omega)^h V(\omega)^{-1}\boldsymbol{E}(\omega), \end{aligned} \tag{5}$$

$$\boldsymbol{E}(\omega) = \boldsymbol{M}(\omega) - S(\omega)\boldsymbol{H}(\omega, \theta). \tag{6}$$

Since the unknown source term $S(\omega)$ is present in the expression, we replace it by the ML estimation:

$$S_{ML}(\omega) = \frac{\boldsymbol{H}(\omega, \theta)^h V(\omega)^{-1}\boldsymbol{M}(\omega)}{\boldsymbol{H}(\omega, \theta)^h V(\omega)^{-1}\boldsymbol{H}(\omega, \theta)}, \tag{7}$$

and integrate the log-likelihood over frequencies to obtain the total log-likelihood function as

$$\mathrm{LL}(\theta) = \frac{1}{2}\int_0^{\omega_H} \Big\{-\boldsymbol{M}(\omega)^h V(\omega)^{-1}\boldsymbol{M}(\omega) + \frac{|\boldsymbol{H}(\omega,\theta)^h V(\omega)^{-1}\boldsymbol{M}(\omega)|^2}{\boldsymbol{H}(\omega,\theta)^h V(\omega)^{-1}\boldsymbol{H}(\omega,\theta)}\Big\} d\omega, \quad (8)$$

where ω_H is the upper limit frequency of the observation and some terms which do not depend on the observation are omitted for simplicity.

In the ML estimation described above, the determination of the shape of $V(\omega)$ is important. When the noise powers of the left and right observation are equivalent, the covariance matrix $V(\omega)$ can be written

$$V(\omega) = \sigma^2(\omega)\begin{pmatrix} 1 & \eta(\omega) \\ \eta(\omega)^* & 1 \end{pmatrix}, \quad (9)$$

where $\sigma(\omega)^2$ is the noise power and $\eta(\omega)$ represents the noise correlation. In a reverberant or ambient-noisy environment, the correlation is indeed not negligible especially when the microphone distance is small. Under the diffused noise field assumption [14], [15] where plane waves of noise arrive randomly from any spherical direction, we here evaluate the correlation statistically as

$$\eta_D(\omega) = \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi} D(\omega,\psi_L(\psi_1,\psi_2)) \times D(\omega,\psi_R(\psi_1,\psi_2))^* \sin\psi_1 d\psi_1 d\psi_2, \quad (10)$$

$$\psi_L(\psi_1,\psi_2) = \cos^{-1}(\sin\psi_1\cos(\phi-\psi_2)), \quad (11)$$

$$\psi_R(\psi_1,\psi_2) = \cos^{-1}(\sin\psi_1\cos(\phi+\psi_2)). \quad (12)$$

In the case of 3D free space, $D(\omega,\psi) = e^{-j\omega L\sin\psi/c}$, where L is the distance between the two microphones, analytically yields $\eta_D(\omega) = \sin(\omega L/c)/(\omega L/c)$, that is, a sinc function [14]. In the case of the sphere diffraction, we can calculate it numerically using the analytical representation of D in eq. (3). According to our calculation, it is similar to a sinc function.

Actually, the real environment is not a perfect diffuse field. Thus, we set

$$\eta(\omega) = \alpha\cdot\eta_D(\omega), \quad (13)$$

where $0 \le \alpha \le 1$, and determine α experimentally.

IV. Localization Experiments in Real Environments

A. Optimization of the Microphone Arrangement

In order to obtain the optimum arrangement, we examined the relationship between the position of the microphones and the localization accuracy through simulation. In this experiment, the radius of the sphere was 30mm, the source signal was speech, and Gaussian white noises with 10dB or 0dB SN ratio (two conditions were used) were added into left and right observation signals. The source direction θ changed from 0° to 360°. The localization accuracy was evaluated by the ratio of estimations inside $\pm5^\circ$ from the source direction. The result is shown in Fig. 4.

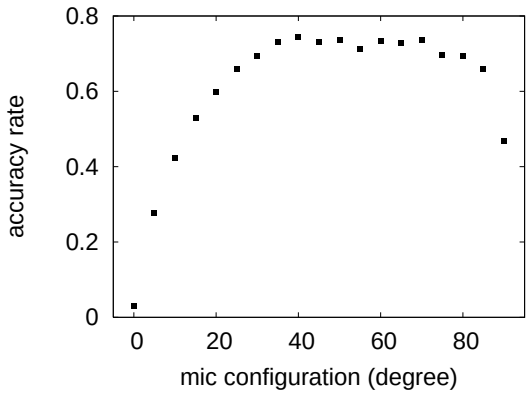

Fig. 4. Localization accuracy depending on microphone position

For $\phi > 70^\circ$ and $\phi < 30^\circ$, the localization accuracy is low. The reason of the former is the difficulty of the front-back judgment due to the arrangement near symmetry, while that of the latter is the smaller difference between the observation signals. Although the accuracy is almost flat between them, we adopted $\phi = 46^\circ$ which achieves the best accuracy in this experiment.

B. Fabrication of the Real System

To build the system, for small-size realization, we used two electret type microphones (AT805F; audio-technica) and a wooden ball with a 30mm radius. Microphone-size holes were carved in the ball at $\pm46^\circ$ in the equator, and the microphones were fixed there. An internal screw was attached at the south pole position and the ball was fixed to a small tripod. A photograph of the constructed system is shown in Fig. 5. The observed signals are readable from the standard audio input through a microphone amplifier (AT-MA2; audio-technica).

C. Experimental Conditions and Results

The speech signal generated by a speaker was recorded by the system. The distance between them was kept equal to 1m and angles every 30° from 0° to 180° were chosen as the source direction. In the environment, several ambient noises, from fan or HDD, and reverberation are present.

The sampling frequency of the recorded signals was 16kHz. The frame length of STFT was 512 samples (32ms), and a Hamming window was used. In the noise covariance of eq. (9), $\sigma^2(\omega) = C$ was used as a simple white noise model. As for the noise correlation model, both $\eta(\omega) = 0$ (independent noise model) and $\eta(\omega) = 0.5\cdot\eta_D(\omega)$ (partially diffused noise model) were examined. For a stable localization, the log-likelihood function eq. (8) was accumulated over 10 frames and the source direction was estimated from the maximum every 10 frames (one estimation per 320ms interval). The estimations from the nearly silent intervals are removed by thresholding.

The results using the independent noise model ($\eta(\omega) = 0$) are shown in Fig. 6. The front and back judgment was rather correctly performed but there are about 20% errors. Furthermore, we can see a tendency of the estimations for

Fig. 5. Photograph of the constructed system

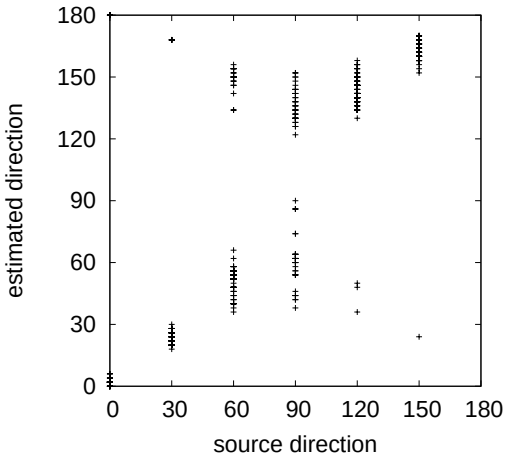

Fig. 6. Localization results using the independent noise model

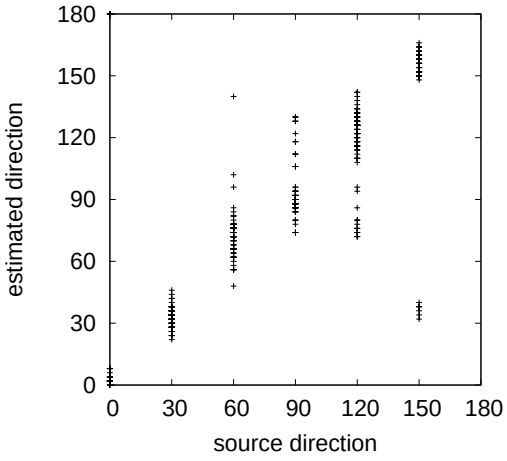

Fig. 7. Localization results using the partially diffused noise model

front sources ($0^\circ < \theta < 90^\circ$) to be biased to the front, and similarly for back sources. A reason for this bias is that the correlated component included in the noise is interpreted as the target signal from the direction to yield the highest correlation, that is, the front ($\theta = 0^\circ$) or the back ($\theta = 180^\circ$) since the noise model doesn't include any correlation components.

Whereas the results using the partially diffused noise model ($\eta(\omega) = 0.5 \cdot \eta_D(\omega)$) show that the biased estimations were compensated and the errors of front-back judgment were less than 10%, which means that our approach works well. The localization accuracy could be improved by increasing the number of accumulation frames. Thus, a trade-off between localization accuracy and temporal resolution should be determined in every application.

V. Conclusions

In this paper, we discussed a novel localization system for any 2D directions using two microphones asymmetrically placed on a sphere. Using a ML estimation based on the diffused noise model, we showed that the constructed system is able to localize any 2D direction almost correctly in real environment. 2ch BSS based on this system is an interesting future work which we plan to investigate.

References

[1] J. Blauert, Spatial Hearing, MIT Press, Cambridge, MA, 1983.

[2] J. G. Harris, C. J. Pu, and J. C. Principe, "A monaural cue sound localizer," in Analog Integrated Circuits and Signal Processing, vol. 23, pp. 163–172. 2000.

[3] N. Ono, Y. Zaitsu, T. Nomiyama, A. Kimachi, and S. Ando, "Biomimicry sound source localization with Fishbone," Trans. Inst. Electrical Engineers of Japan, vol. 121-E, no. 6, pp. 313–319, 2001.

[4] H. Nakashima, N. Ohnishi, and T. Mukai, "A learning system for estimating the elevation angle of a sound source by using a feature map of sperctrum," Trans. IEICE D-II, vol. J87-D-II, no. 11, pp. 2034–2044, 2004. (in Japanese)

[5] N. Ono and S. Ando, "Sound Source Localization Sensor with Mimicking Barn Owls," Proc. Transducers'01, vol.2, pp.1654-1657, Munich, 2001.

[6] N. Ono, A. Saito, and S. Ando, "Bio-mimicry Sound Source Localization with Gimbal Diaphragm," Trans. Inst. Electrical Engineers of Japan, vol. 123-E, no. 3, pp. 92-97, 2003.

[7] R. E. Irie, "Multimodal sensory integration for localization in a humanoid robot," Proc. the Second IJCAI Workshop on Computational Auditory Scene Analysis (CASA '97), pp 54–58, 1997.

[8] H. Nakashima, Y. Chisaki, T. Usagawa and M. Ebata, "Frequency domain binaural model based on interaural phase and level differences," Acoust. Science & Technology, vol. 24, no. 4, pp. 172–178, 2003.

[9] 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.

[10] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS Using the EM Algorithm in Reverberant Environment," Proc. WASPAA2007, 2007. (submitting)

[11] A. A. Handzel and P. S. Krishnaprasad, "Biomimetic Sound-Source Localization," IEEE Sensors Journal, vol. 2, no. 6, pp. 607–616, Dec. 2002.

[12] T. Hayasaka and S. Yoshikawa, Onkyo Shindoron (Theory of Acoustic Vibration), Maruzen, Tokyo, 1974. (in Japanese)

[13] M. Brandstein and D. Ward, Microphone Arrays, Springer, 2001.

[14] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelmann and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," JASA, vol. 27, no. 6, pp. 1072-1077, 1955.

[15] I. A. McCowan, H. Bourlard, "Microphone Array Post-Filter Based on Noise Field Coherence," IEEE Trans. on Speech and Audio Processing, vol. 11, no. 6, pp. 709-716, 2003.