The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Real-time sound source orientation estimation using a 96 channel microphone array

Hirofumi NAKAJIMA, Keiko KIKUCHI, Toru DAIGO,
Yutaka KANEDA, Kazuhiro NAKADAI and Yuji HASEGAWA

Abstract— **This paper proposes real-time sound source orientation estimation based on orientation-extended amplitude beamforming (OE-ABF). To recognize a sound source orientation (such as face orientation) is an important function for a robot who can achieve natural human-robot interaction because the function is required to distinguish the human target from a robot or another person. We developed a sound source orientation system using orientation-extended beamforming (OE-BF) and showed the system worked properly at least under a specific controlled environment. However, in practical use, this system does not work properly because the system doesn't take into account the differences between the supposed model in OE-BF and in practical situations. For example, the system model supposes that there is neither noise nor reverberation, however, this is not a realistic assumption. To solve this assumption mismatch problem, we propose sound source orientation estimation based on OE-ABF, and constructed a real-time sound source orientation estimation system with the proposed method using a 96ch microphone array. Evaluation results of our proposed system show that the average error of estimated angles is lower than 5°, while the error of our previously reported system was greater than 20°. With this system, the robot is able to distinguish that the utterance target of a person standing 1m in front is itself or another person standing 0.2m to the left of the robot. This is valuable for human-robot interaction.**

I. Introduction

"Robot Audition" is important to achieve natural human-robot interactions [1]. Sound source localization, separation and speech recognition are primary functions for robot audition. In fact, many algorithms and systems for these functions have been proposed [2], [3], [4]. Sound source orientation estimation is another important function because a robot should recognize a source's orientation such as face orientation. This function is required to distinguish the target of the human from a robot or another person. If we only focused on face orientation estimation, then visual processing would be effective [5]. However, even for face orientation, audio processing has several merits. One of the merits is that audio processing is not influenced by lighting conditions. When we use a short humanoid type robot having camera(s) on its face, for example Honda ASIMO, the elevation angle of the robot's face needs to be high to have eye-contact with a human. In this case, the "backlight" situation often occurs because the line of the robot's vision lies on ceiling lighting. In this "backlight" situation, to recognize face orientation using visual processing is difficult. Another merit is computational cost. Audio processing requires much less computational cost than visual processing, and therefore, it is suitable for real-time processing.

There are a few reports about sound source orientation estimation [6], [7]. In [7], we proposed a sound source orientation estimation system based on an orientation-extended beamforming (OE-BF) method using a microphone array. However, this system has several issues and it is difficult to use in practical environments. This paper proposes a new orientation estimation system based on the orientation-extended amplitude beamforming (OE-ABF) method this can solve these issues and it can be used practically in real-time.

II. Issue and approach

The primary issues of the reported system [7] are:

- Low precision
- Pre-measure requirement of transfer functions (TFs)
- Parameter adjustment required by a human
- Low real-time factor.

Because the orientation estimation system [7] supposes an ideal acoustic environment without any noises nor reverberations, the performance in practical environments is poor. Also [7] does not take into account the difference between TFs of the target human speaker and pre-measured TFs for OE-BF coefficient design. Therefore, if we used a loudspeaker to measure and make the OE-BF coefficient, the estimated orientation would include many errors because of the difference between a human speaker and a loudspeaker. In [7], an actual human speaker, who is the subject for evaluating this system, was used for pre-measurement. The system performance was good enough. However, in practical use, we can not specify who will use this system. Therefore, to get high performance estimation we should measure all possible speakers' TFs. This is unrealistic and impossible. Parameter adjustment is another problem of [7]. To distinguish whether the input signal is voice or not, [7] uses a simple power threshold method, that requires a threshold parameter. Because the appropriate parameter is changed according to each uttered voice level and orientation, we should decide the parameter after recording each speaker and orientation. This is also unrealistic because we can not know the uttered voice level and orientation in advance. Moreover, since [7] uses OE-BF for not only orientation estimation but localization simultaneously, therefore the calculation cost will be very high because it is proportional to the product of the number of digitized locations and orientations. Also, simultaneous

H. Nakajima, K. Nakadai and Y. Hasegawa are with the Honda Research Institute Japan Co., Ltd, 8-1, Honcho, Wako-shi, Saitama 351-0188, Japan. (email: {nakajima, nakadai, yuji.hasegawa}@jp.honda-ri.com)

K. Kikuchi, T. Daigo and Y. Kaneda are with the Tokyo Denki University, 2-2, Kanda-Nishiki-cho, Chiyoda-ku, Tokyo 101-8457, Japan. (email: {08gmc09@ms, 07gmc09@ms, kaneda@c}.dendai.ac.jp)

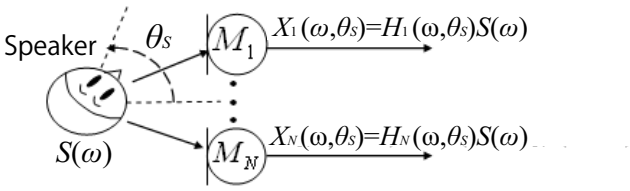


Fig. 1. Sound source orientation estimation model

calculation increases estimation errors by transferring the error of localization into orientation and vice versa.

To solve these issues we propose a new system introducing the following method and processes:

- Orientation-extended amplitude beamforming (OE-ABF)
- Time-frequency mask
- Temporal histogram

OE-ABF is a beamforming method focused on amplitude of TFs, which is robust to the difference of TFs, and solves the low precision problem which originates from the TF difference between the BF coefficient (loudspeaker) and the input signal (human speaker). A time-frequency mask is applied to exclude the low confidence time frames and frequency bands. To use this mask, this system prevents estimation precision from being degraded due to noises and reverberations. A temporal histogram is introduced instead of a simple power average. This improves the estimation precision when the source's spectrum is not widely spread. The details of the method and processes are described in the next section.

III. Orientation estimation based on OE-ABF

We used OE-ABF as a method for source orientation estimation. OE-ABF is an advanced version of OE-BF used in [7] regarding the input vectors and BF coefficient vectors. OE-ABF is robust to acoustical environmental changes, for example, source location and type. To explain OE-ABF theoretically, first, we formulated the orientation-extended transfer function and described OE-BF. Next, we proposed OE-ABF and explained the details by clarifying the difference between OE-ABF and OE-BF. Finally, we will show the orientation estimation using OE-ABF.

A. Formulation of orientation-extended transfer function

Acoustical transfer function (TF) is a prime function representing transfer characteristics from a source to a microphone and is used for almost all acoustical signal processing. In general, TF is treated as a function whose arguments are locations of sound source and microphone and not their orientations. However, in actual use, TF changes according to sound source orientation (and also microphone orientation if using a directional microphone) because the sound source has a non-uniform directivity pattern. Therefore, TF should be a function of not only the locations but the orientations. Orientation-extended TF is the TF including the source orientations as an argument. Fig. 1 shows a propagation model including sound source orientation using an N-elements microphone-array. $S(\omega)$ shows the sound source spectrum where ω represents the frequency. M_k is the k-th microphone ($k = 1, 2, ..., N$) and $H_k(\omega, \theta_S)$ denotes the transfer function between the kth microphone and the sound source with orientation θ_S. In this section, we suppose the source location is fixed and known for simplicity. The recorded signal with M_k is represented as

$$X_k(\omega, \theta_S) = H_k(\omega, \theta_S)S(\omega). \quad (1)$$

This equation can be simplified using a vector notation as

$$\begin{aligned} \mathbf{h}(\omega, \theta_S) &= [H_1(\omega, \theta_S), ..., H_N(\omega, \theta_S)]^T, \quad (2)\\ \mathbf{X}(\omega, \theta_S) &= [X_1(\omega, \theta_S), ..., X_N(\omega, \theta_S)]^T, \\ &= [H_1(\omega, \theta_S)S(\omega), ..., H_N(\omega, \theta_S)S(\omega)]^T, \\ &= \mathbf{h}(\omega, \theta_S)S(\omega), \end{aligned}$$

where T denotes a transpose operator, $\mathbf{X}$ is defined as an input vector and $\mathbf{h}$ is an orientation-extended TF vector.

B. Orientation-extended beamforming (OE-BF)

OE-BF is a beamforming method derived by extending conventional beamforming (BF) regarding a source's orientation. BF is a method to make spatial directivity, and it enables selective recordings. BF is able to make a sound power map by steering the focus point of BF's directivity, and also estimate the sound source location by searching for the maximum point of the power map[8].

By using orientation-extended TF when designing BF, we can also make an orientation-extended BF (OE-BF) method that can estimate sound source orientation by applying the same method regarding location[7].

The OE-BF output is formulated as

$$C(\omega, \theta_S, \theta_{BF}) = \mathbf{g}(\omega, \theta_{BF})^H \mathbf{X}(\omega, \theta_S), \quad (3)$$

where $\mathbf{g}(\omega, \theta_{BF})$ shows the BF coefficient whose focus angle is θ_{BF} and H shows the complex conjugate transpose operator. If we use a Delay-and-Sum method (DS-BF), the BF coefficient $\mathbf{g}(\omega, \theta_{BF})$ is represented as

$$\mathbf{g}(\omega, \theta_{BF}) = \frac{\mathbf{h}(\omega, \theta_{BF})}{||\mathbf{h}(\omega, \theta_{BF})||}. \quad (4)$$

C. Orientation-extended amplitude beamforming (OE-ABF)

OE-BF calculates $C(\omega, \theta_S, \theta_{BF})$ from an inner product of two complex vectors $\mathbf{g}(\omega, \theta_{BF})$ and $\mathbf{X}(\omega, \theta_S)$. If we use DS-BF, the output $C(\omega, \theta_S, \theta_{BF})$ is represented as

$$C(\omega, \theta_S, \theta_{BF}) = \frac{\mathbf{h}(\omega, \theta_{BF})^H \mathbf{h}(\omega, \theta_S)}{||\mathbf{h}(\omega, \theta_{BF})||} S(\omega). \quad (5)$$

This inner product part $\mathbf{h}(\omega, \theta_{BF})^H \mathbf{h}(\omega, \theta_S)$ shows the similarity between the TF for BF coefficient and the actual TF when the source produces the sound. Therefore, if the target sound source differs from a sound source used for the BF coefficient design, the output $C(\omega, \theta_S, \theta_{BF})$ decreases even if the orientation is the same. In the reported system [7] using OE-BF, BF coefficients using the same sound source, that is a voice utterer, is applied. However, it is impossible to make all BF coefficients of possible voice utterers. Therefore,

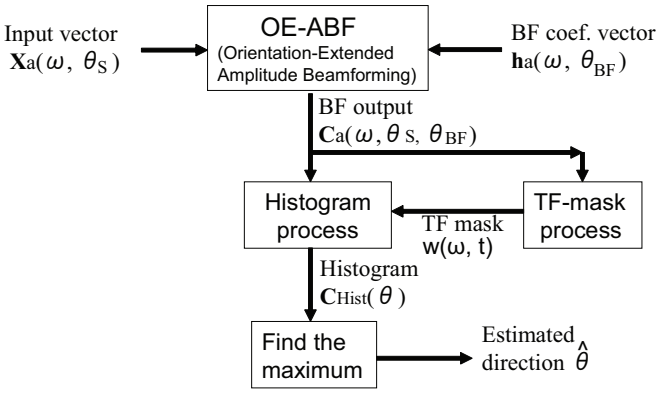


Fig. 2. Diagram for proposed method

the reported system [7] can not be used as a practical system at all. For this problem, a system that is robust for source differences is desired.

To solve this problem, we propose orientation-extended amplitude beamforming (OE-ABF) . According to a preliminary analysis of TF differences regarding source type, we found that the phase of TF was easily changed by the source type, but the amplitude was not strongly changed. This tendency is particular in high frequency because the phase difference regarding the geometrical difference is anti-proportional to the wavelength, while the amplitude difference is independent to the wavelength.

For OE-ABF, we use amplitude-based TF vector $\mathbf{h}_a(\omega, \theta_S)$ and input vectors $\mathbf{X}_a(\omega, \theta_S)$ as

$$\mathbf{h}_a(\omega, \theta_S) = [|H_1(\omega, \theta_S)|, ..., |H_N(\omega, \theta_S)|]^T, \quad (6)$$

$$\mathbf{X}_a(\omega, \theta_S) = [|X_1(\omega, \theta_S)|, ..., |X_N(\omega, \theta_S)|]^T \quad (7)$$

OE-ABF output is represented as

$$C_a(\omega, \theta_S, \theta_{BF}) = \mathbf{g}_a(\omega, \theta_{BF})^H \mathbf{X}_a(\omega, \theta_S), \quad (8)$$

where $\mathbf{g}_a(\omega, \theta_{BF})$ shows the amplitude BF coefficient. The DS-BF coefficient of OE-ABF is calculated as

$$\mathbf{g}_a(\omega, \theta_{BF}) = \frac{\mathbf{h}_a(\omega, \theta_{BF})}{||\mathbf{h}_a(\omega, \theta_{BF})||}. \quad (9)$$

With this BF coefficient, the output $C_a(\omega, \theta_S, \theta_{BF})$ can be calculated as

$$C_a(\omega, \theta_S, \theta_{BF}) = \frac{\mathbf{h}_a(\omega, \theta_{BF})^H \mathbf{h}_a(\omega, \theta_S)}{||\mathbf{h}_a(\omega, \theta_{BF})||} |S(\omega)|. \quad (10)$$

Therefore, OE-ABF output is proportional to the inner product between $\mathbf{h}_a(\omega, \theta_{BF})$ and $\mathbf{h}_a(\omega, \theta_S)$, which represents the TF similarity of the BF coefficient and the recorded signal vectors.

D. Orientation estimation based on OE-ABF

Fig. 2 shows the diagram for our proposed orientation estimation, that uses the OE-ABF method as a core process and also introduces a time-frequency mask and a histogram as additional processes for improving estimation performance. Basically, the estimated orientation is calculated as an angle $\hat{\theta}_{\omega t}(\omega, t)$ which maximizes $C_a(\omega, \theta_S, \theta_{BF})$ regarding θ_{BF}. This $\hat{\theta}_{\omega t}(\omega, t)$ is denoted as two functions: frequency ω and time t, because this value is calculated over all frequency bands and time frames although t had been ommitted before because of simplicity. To decide the final estimated orientation angle $\hat{\theta}$ from $\hat{\theta}_{\omega t}(\omega, t)$, we should apply some gathering process. For the reported system [7], we used a simple power average over all frequency bands and time periods. However, this simple average is not robust to noises and reverberations because no excluding process for them is applied in the average. Also the power average degrades the precision when the source signal does not have a wide power spectrum because the low power bands are almost ignored in the average even if the signal-to-noise ratio (SNR) is high enough for estimation. To solve these problems, a time-frequency mask is applied to exclude low confidence data for orientation estimation and a histogram is used as a gathering process to improve precision for a non-wide spectrum source, such as speech vowels.

The final estimated orientation angle is calculated as

$$\hat{\theta} = \underset{\theta}{\mathrm{argmax}}[C_{Hist}(\theta)], \quad (11)$$

where $\mathrm{argmax}_\theta[C(\theta)]$ represents the argument θ that maximizes the function $C(\theta)$ and $C_{Hist}(\theta)$ is a histogram of estimated orientation angle, whose derivation is described in the next subsection.

This process corresponds to a digital weight version of [9], which uses pre-whitening and spectral weighting. Since our method requires only multiplications, the calculation cost is much lower than [9], which requires divisions and exponentials with floating-point exponents.

1) Histogram: In the reported system [7], which uses a simple power average, the averaged result is strongly affected by high power frequency bands, but less affected by low power bands. This means that a limited part of the frequency bands is taken into account, even if the source spectrum includes all frequency bands. For example, for the speech source, high frequency bands (over 4kHz) are not counted because the power of low frequency bands (below 1kHz) is much larger than that of high frequency even if the SNR is high enough for estimation. To counter this phenomenon, we introduce a histogram $C_{Hist}(\theta)$ as

$$C_{Hist}(\theta) = \sum_{\omega,t} w(\omega, t) U(\omega, t, \theta) \quad (12)$$

$$U(\omega, t, \theta) = \begin{cases} 1 & if\ \theta = \mathrm{argmax}_{\theta_{BF}}[C_a(\omega, \theta_S(t), \theta_{BF})] \\ 0 & otherwise, \end{cases} \quad (13)$$

where $w(\omega, t)$ shows the time-frequency mask described in the next section.

2) Time-frequency mask: We calculated this mask as $w(\omega, t) = w_\omega(\omega, t) w_t(t)$, where $w_\omega(\omega, t)$ shows the frequency mask of time t and $w_t(t)$ is the time mask. The frequency mask is calculated based on an input auto-correlation function and the time mask is derived from the result of voice activity detection (VAD). First, we explain VAD for the time-mask. The conventional system uses a power-threshold method for VAD. This method has difficulties in deciding the threshold value because the appropriate value varies

according to the environment, for example, speaker location and background noise power. Therefore, the conventional system needs to be used in a fixed environment and the threshold value needs to be adjusted manually. To solve these issues, we used an auto-correlation based VAD, which evaluates a periodic property of vowels [10]. This method has the following advantages:

- Not influenced by source power
- Robust to non-periodic noises
- Not required to adjust parameters manually.

The process is as follows:

- Calculate auto-correlation function $\phi(\tau)$ of input signal with a truncated window of length L, where τ is the delay time.
- Search the minimum delay time τ_{min} that is the earliest delay time having the condition $\phi(\tau) < \beta$, where β is a parameter.
- Get the maximum correlation value $\phi(\tau)_{max}$ in the range $\tau > \tau_{min}$.
- Make the decision: if $\phi(\tau)_{max} > \alpha$ then "Voice" otherwise "Non-voice".

Based on this decision, the binary time-mask $w_t(t)$ is made as

$$w_t(t) = \begin{cases} 1 & \text{Voice active period} \\ 0 & \text{Pause period} \end{cases} . \quad (14)$$

After preliminary experiments of this VAD, we found that the estimated orientation included many errors in the last part of the voice period because of reverberation. To solve this problem, we excluded the last L_t part from the voice period. We decided these parameters heuristically as $\alpha = 0.5, \beta = -0.2, L = 1024, L_t = 1024$, and used the setting in all experiments.

Next, we describe the frequency mask. Conventional methods use all frequency bands without concerns for SNR. In this case, because the estimated orientations of low SNR frequency bands have many errors, estimation precision becomes poor. To solve this, we introduced a binary frequency mask $w_\omega(\omega, t)$ based on SNR as

$$w_\omega(\omega, t) = \begin{cases} 1 & if \ \max_\theta[C_a(\omega, \theta_S(t), \theta)] \\ & \quad > \text{mean}_t[\max_\theta[C_a(\omega, \theta_S(t), \theta)]]\delta \\ 0 & \text{Otherwise} \end{cases} , \quad (15)$$

where $\max_\theta$ shows the maximum value regarding θ, mean_t denotes the time-average and δ is the threshold parameter. In this paper, we decided $\delta = 1.5$ heuristically.

IV. Real-time sound source orientation estimation system

We made a real-time sound source orientation estimation system using previously described OE-ABF and additional processes. Fig. 3 shows the system diagram. Each module in the system is explained in the following subsections.

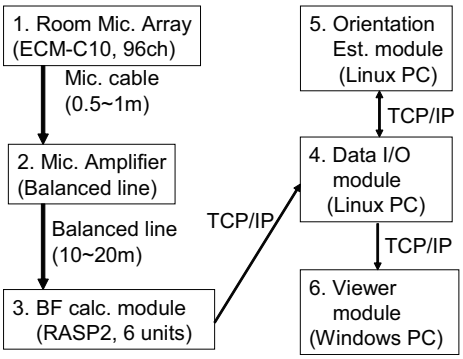


Fig. 3. Real-time location and orientation estimation system

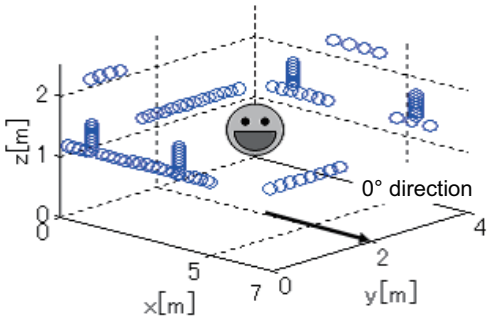


Fig. 4. Arrangement of microphones

A. Room microphone array (1)

The room microphone array configuration was composed of 96 microphones embedded in the walls. The size of the room was 4.0 m × 7.0 m× 3.0 m, and the reverberation time (RT_{20}) was 0.23 s. Fig. 4 shows the microphone locations. In this figure, the ∘ marks show the locations. 64 microphones were used mainly for horizontal plane estimation and located in the walls with almost the same spacing between each microphone. The other 32 microphones were for improving vertical estimation precision and used for 4 linear same-spaced arrays. All microphones were SONY ECM-C10, which are non-directive type microphones with 70dB SNR. All microphones were connected to amplifiers with cables less than 1m in order to prevent noise from environmental electro-magnetic waves.

B. Microphone amplifier (2)

Microphone amplifiers were AEMM-04 made by Nittobo Acoustic Engineering. AEMM-04 had 4ch microphone amplifiers with power supply and balanced output functions. The outputs were sent to BF calculation modules by balanced cables with shields. Because the calculation modules were outside the room, the cable length was 10-20m. Although the cables were long, signal degradation was limited because of the balanced lines and amplifiers.

C. BF calculation module (3)

We used 6 RASP2 made by JEOL system technology as a BF calculation module. RASP2 had 16ch A/D converters and a CPU, on which Linux OS could be run. 1) Sound source

localization and 2) orientation estimation up to similarity calculations of each frequency were processed.

1) Sound source localization: We used steered DS-BF for localization. The TF for the DS-BF was measured using an omni-directive loudspeaker (B&K 4295). The BF focus points were 221 points decided by digitizing the search space of the room (4m $\times$ 3m) using a 0.25m square mesh. To reduce calculation costs, dominant 5 frequency bands with high SNR were selected and processed. A maximum of 3 possible sources were detected by this process.

2) OE-ABF output calculations for orientation estimation: The OE-ABF output $C_a(\omega, \theta_S(t), \theta)$ was calculated using a prepared BF coefficient, only at the source positions estimated from the localization process. The BF coefficients are calculated and buffered using 1769 TFs measured by a directive loudspeaker (Generec 1029A) placed at all possible 221 positions and 8 source orientations (45° step). For further cost reduction, we limited the frequency band to 1kHz.

D. Data I/O module (4)

We used middle-ware MMI Ver.2 [11] as a Data I/O module that can manage various types of data and treat asynchronous data. Output data from BF calculation module and estimated orientation data of the estimation module were transferred through this module.

E. Sound source orientation estimation module (5)

This module calculated the final estimated orientation from BF output data and sent it to a viewer through MMI. First, this module made a time and frequency mask based on SNR of each time frame and frequency band. Next, using these masks this module made the histogram of the estimated orientation for each frequency band and time frame. Finally, this module decided the source orientation by taking the orientation in which the histogram becomes the maximum value. To smooth out estimation data, this module excluded the time frames in which the estimated source position was highly non-continuous. These non-continous time frames were decided as the frames in which the source moving speed was higher than the highest walking speed.

F. Viewer module (6)

We used an MMI viewer as the viewer module to display the estimated position and orientation with the experimental room graphically. This module used OpenSceneGraph [1]. Fig. 5 is an example of the viewer display. Using this viewer, we could confirm that the system could estimate the source position and orientation correctly and work in real-time.

V. Evaluation

We performed experiments to evaluate our proposed method. To show its effectiveness, we compared it to the reported system [7] under the same condition. To do so, we focused on their orientation estimation performance, and did not take into account the localization errors nor the errors originating from the frequency band limitation. In this section, after we described the BF coefficient dataset and speech data for evaluation, we analyzed each orientation estimation process to make sure that the algorithm was working properly (Exp. 1 in Sec. V-A). Next, we evaluated the estimated orientation errors for both methods (Exp. 2 in Sec. V-B). We also demonstrated our real-time sound source orientation estimation system with snapshots (Exp. 3 in Sec. V-C). Finally, we investigated the relationship between performance and the number of microphones (Exp. 4 in Sec. V-D).

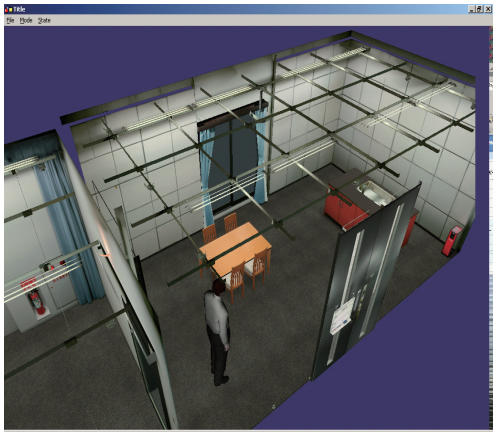


Fig. 5. Viewer display example

The BF coefficient dataset was made by measuring impulse responses with a loudspeaker (GENELEC 1029A) and using Fourier transformation. We fixed the source position to the center of the room in this experiment. The orientation of the loudspeaker was rotated from 0° to 360° with a 15° step (total 24 orientations) as shown in 4. The sampling frequency was 16kHz, the output signal for measurement was Time Stretched Pulse (TSP) with a length of 2^{14}. To exclude the reflections of the later reverberations from the impulse responses, we truncated the responses into 1024 points.

Speech data was recorded by an actual male speaker standing at the same position where the loudspeaker had been. During the recording, we did not place any noise sources to make the same condition as the previous report [7]. However, there was some background noise sources such as an air-conditioner. SNR changed depending on the microphone channel, the range of SNR was 15-30dB. The orientations were 4 directions (0°, 90°, 180° and 270°). The uttered voice used 5 continuous Japanese-language vowels: "A, I, U, E, O."

A. Experiment 1: Analysis of each process

First, we evaluated the VAD part. Fig. 6 shows a recorded voice spectrogram. We found that the voice power was included in the recorded signal up to 8kHz, in the frequency range below 500Hz the background noise was dominant, and there were low level spectrums in the last part of the voice spectrogram because of room reverberations. Fig. 7 represents its VAD result. We confirmed that the correlation-based VAD could detect the voice period correctly in spite of

[1] http://www.openscenegraph.org/projects/osg

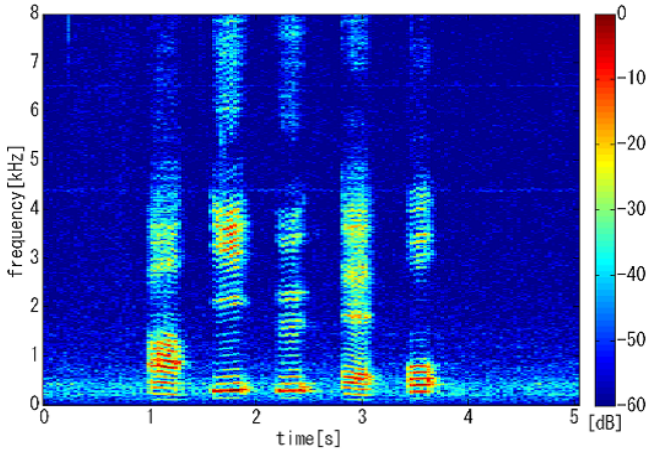


Fig. 6. Input signal spectrogram

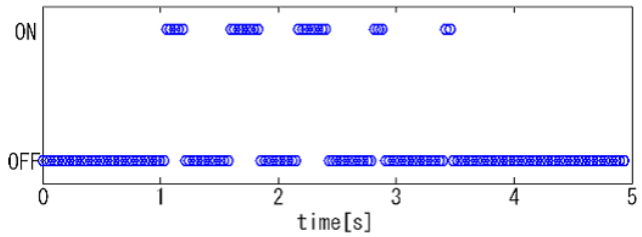


Fig. 7. Voice activity detection result

such noises and reverberations. The reason the time length detected by VAD was shorter than the strong power period in the voice spectrogram was that the VAD excluded the reverberation-dominant area in which the orientation was estimated inaccurately.

Next, we evaluated the frequency mask process. Fig. 8 shows the maximum BF output, which is used to calculate the frequency mask as described in Sec. III. By comparing this figure to Fig. 6, the maximum similarity is emphasized in the high frequency range (2kHz or higher), oppositely, in the low frequency (1kHz or lower) deemphasized. To use the low frequency band for estimation would degrade the precision because the BF output difference is small in this band although the SNR is high. Fig. 9 shows the frequency mask made by the proposed method. Black pixels show the masked areas. We found that the high frequency area, that is effective for estimation because of having large BF output differences, is not masked, oppositely, the low frequency area is masked.

Fig. 10 shows the BF output histogram using the frequency mask. The correct orientation is 270°. We found that the histogram became the maximum value at 270° in almost all time areas.

B. Experiment2: Evaluation of orientation estimation

To show the effectiveness of our proposed method, we evaluated the estimation errors of proposed and conventional systems. Also, to analyze the contribution of the newly introduced method and processes, we switched the method and processes:

- Basic estimation method:
  OE-BF (Amplitude & Phase) or OE-ABF (Amplitude)

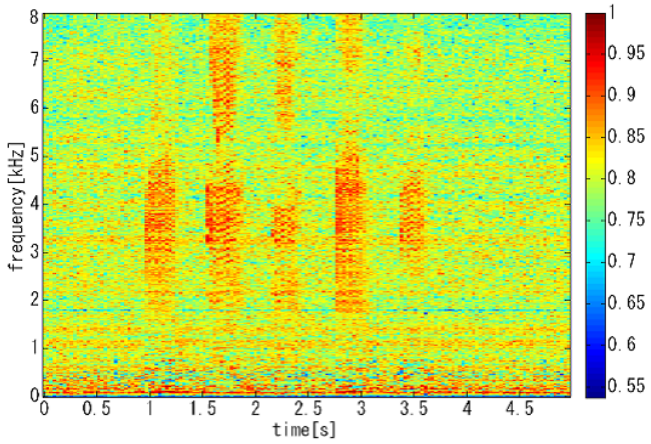


Fig. 8. Maximal value of the BF output in time-frequency domain

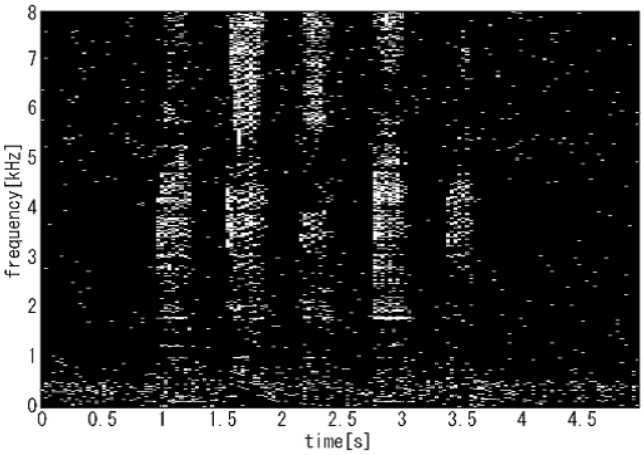


Fig. 9. Time-frequency mask

- Time mask (VAD):
  Power threshold (Manual VAD) or Correlation based (Automatic VAD)
- Frequency mask:
  Not used or Applied (Frequency mask)
- Histogram:
  Not used or Applied (Histogram).

Total 16 systems were tested. The conventional system used no new method nor processes, and the proposed system used a new method and three new processes. Fig. 11 shows the estimation error and its standard deviation of all systems. Fig. 11(a) is the result with manual VAD, and (b) is with automatic VAD. In both figures, the dark color bars represent results based on OE-BF, and the bright color bars are based on OE-ABF. The left two bars are without frequency masks nor histogram, next to the left bars are only with a frequency mask, next to them are only with a histogram, and the right bars are with a frequency mask and histogram. Compared to 11(a) and (b), we found that the difference of errors between manual and automatic VAD is small. Therefore, the proposed VAD could achieve almost the same performance as the manual VAD without any parameter adjustments. In both figures, all the dark color bars have larger errors compared to the bright color bars. This shows that the OE-ABF is the most important factor to improve the estimation performance

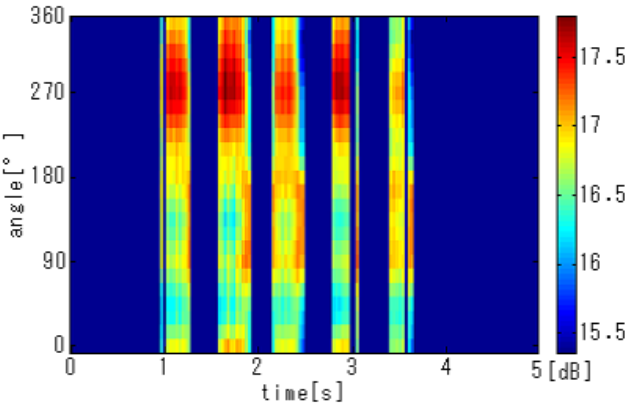
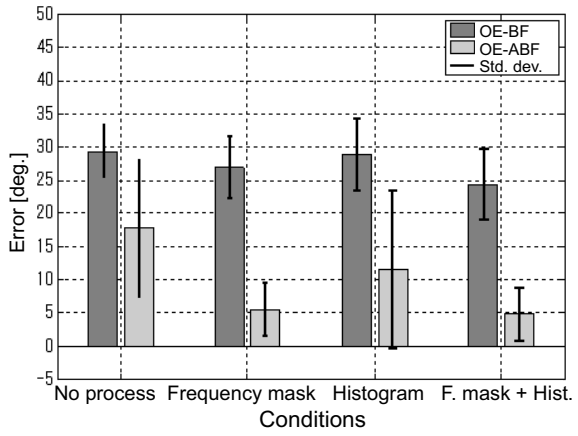
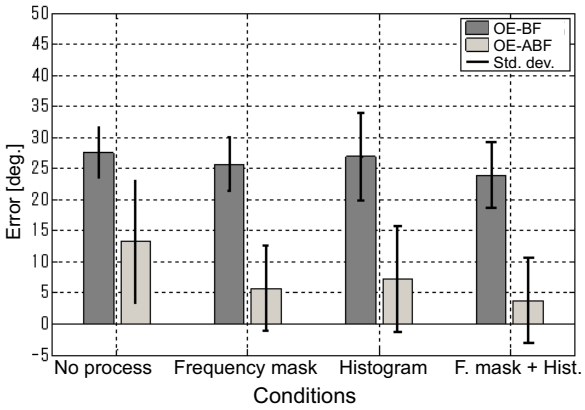


Fig. 10. BF output histogram for each angle



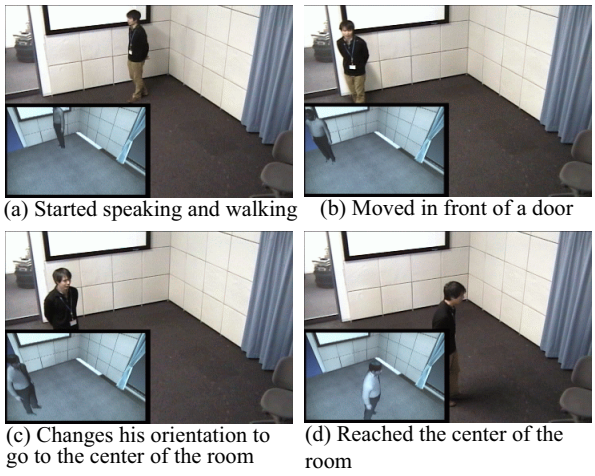
(a) Manual VAD (power threshold based)



(b) Automatic VAD (correlation based)

Fig. 11. Estimation error and standard deviation

in this experiment. Comparing the 4 conditions regarding frequency mask and histogram in both figures, we found that the method with the least errors was the method using both frequency mask and histogram. We made sure that the frequency mask and histogram processes were both effective for reducing errors. The errors of conventional and proposed methods were 29° and 4°, respectively. We achieved a 25° error reduction by introducing the new method. Considering the BF coefficient dataset was prepared using every 15 ° step TFs, the 4° error was very small and also regarding standard deviation 7° is considered to be almost the limit value.



(a) Started speaking and walking (b) Moved in front of a door

(c) Changes his orientation to go to the center of the room (d) Reached the center of the room

Fig. 12. Snapshots of real-time sound source orientation estimation system

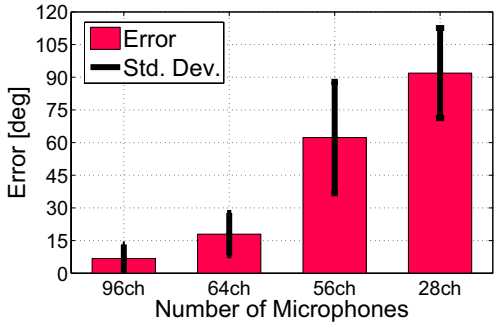


Fig. 13. Estimation error by number of microphones

C. Experiment3: Snapshots of real-time sound source orientation estimation system

Fig. 12 shows the snaphots of our real-time system when a person was walking around a room while speaking. The person started at the corner of the room, moved in front of a door, and then went to the center of the room. We confirmed that both localization and orientation were successfully estimated with our system in real-time.

D. Experiment4: Performance by the number of microphones

Fig. 13 shows the estimation error by the number of microphones used in our proposed method. The 64ch arrangement is the same as the one used in [7], which is made by removing 4 vertical 8ch microphone arrays from the full 96ch arrangement (Fig. 4). We found that the estimation error almost tripled. The 56ch arrangement is made by removing 2 horizontal 4ch microphone arrays located on high positions ($z > 2$) from 64ch. The error of 56ch system was much larger than the 64ch system even though it has 8 less chanels. The 28ch arrangement is made by 1/2 alternative decimation of 56ch. The error of 28ch is almost 90°. We found that both number and arrangement of micrphones is important to reduce estimation errors.

VI. Conclusion

In this paper, we reported a real-time sound source orientation estimation system. This system is based on an

orientation-extended amplitude beamforming method (OE-ABF) and we introduce time-frequency mask and temporal histogram processes to improve the system performance. Evaluation results showed that the average error of the estimated angles was lower than 5°, while that of the previously-reported system was greater than 20°. Automatically setting of parameters, reduction in the number of microphones and further improvements of robustness by associating with visual information are challenging future work.

REFERENCES

[1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *17th National Conf. on Artificial Intelligence (AAAI2000)*. AAAI, 2000, pp. 832–839.

[2] J. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *2004 International Conference on Intelligent Robots and Systems (IROS2004)*. IEEE/RSJ, 2004, pp. 2123–2128.

[3] I. Hara, F. Asano, Y. Kawai, F. Kanehiro, and K. Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid hrp-2," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2004, pp. 2404–2410.

[4] S. Yamamoto, K. Nakadai, J.-M. Valin, J. Rouat, F. Michaud, T. Ogata, K. Komatani, and H. G. Okuno, "Making a robot recognize three simultaneous sentences in real-time," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2005, pp. 897–892.

[5] G. Dedeoglu, T. Kanade, and S. Baker, "The asymmetry of image registration and its application to face tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 807–823, 2007.

[6] P. Meuse and H. Silverman, "Characterization of talker radiation pattern using a microphone array," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1994, vol. 2, pp. 257–260.

[7] K. Nakadai, H. Nakajima, K. Yamada, Y. Hasegawa, T. Nakamura, and H. Tsujino, "Sound source tracking with directivity pattern estimation using a 64ch microphone array," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2005, pp. 196–202.

[8] D. Johnson and D. Dudgeon, *Array Signal Processing*, Prentice hall, 1993.

[9] J. Valin, F. Michaud, and J. Rouat, "Robust sound source localization using a microphone array on a mobile robot," in *2003 International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, 2003, vol. 2, pp. 1228–1233.

[10] R. Tucker, "Voice activity detection using a periodicity measure," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[11] K. Nakadai, H. Nakajima, M. Murase, S. Kaijiri, K. Yamada, Y. Hasegawa, and H. Tsujino H. Okuno, "Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2006, pp. 852–859.