2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Design and Evaluation of Two-Channel-Based Sound Source Localization over Entire Azimuth Range for Moving Talkers

Hyun-Don Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno

Abstract— **We propose a way to evaluate various sound localization systems for moving sounds under the same conditions. To construct a database for moving sounds, we developed a moving sound creation tool using the API library developed by the ARINIS Company. We developed a two-channel-based sound source localization system integrated with a cross-power spectrum phase (CSP) analysis and EM algorithm. The CSP of sound signals obtained with only two microphones is used to localize the sound source without having to use prior information such as impulse response data. The EM algorithm helps the system cope with several moving sound sources and reduce localization error. We evaluated our sound localization method using artificial moving sounds and confirmed that it can well localize moving sounds slower than 1.125 rad/sec. Finally, we solve the problem of distinguishing whether sounds are coming from the front or back by rotating a robot's head equipped with only two microphones. Our system was applied to a humanoid robot called SIG2, and we confirmed its ability to localize sounds over the entire azimuth range.**

I. Introduction

Recently, sound source localization has been applied to robots as a way of improving human-robot interactions [1-3]. Also, some sound source separation methods such as beamformer [4,5] need the location of target and noise signals in order to separate target signals for speech recognition. In fact, many methods of sound source localization for humanoid robots have been developed, and their performance has generally improved over time. However, the following three items should still be developed or improved:

1) Robots should be able to localize moving sound sources as well as fixed sound sources. Also, since robots should be able to move and rotate their bodies and heads in order to track someone, a sound source localization method should be able to localize moving sounds while coping with the effects created by moving microphones.

2) To design and evaluate sound localization systems to cope robustly with moving sounds, we first need database for various moving sounds. Thus, as a conventional way to create database for moving sounds, we have recorded sound signals and their positions while manually moving the speaker. Therefore, it is difficult to create moving sounds which have accurate track information and to repeatedly create the same database with the same condition in order to compare a developed system with other ones for sound localization regardless of a kind of methods.

3) Robots need to improve their cognition abilities (active perception) concerning changing location of sounds while they are in motion. For example, robots should be able to distinguish whether sound signals are coming from the front or back if they rotate or move only two microphones placed in the robot's head or body.

In this study, we accordingly improved the sound source localization system for humanoid robots by implementing three principal techniques:

1. Our two-channel-based system can reliably localize two moving sounds without prior information.
2. To evaluate our system, we proposed the new way to construct a database for moving sounds.
3. Robots implementing our system can localize sounds over the entire azimuth range by rotating their head or body with two microphones.

In detail, first, we already developed two channel based sound source localization system [12]. This one used cross-power spectrum phase (CSP) analysis [6] of sound signals obtained by only two microphones to localize the sound source without impulse response data. We also applied an expectation-maximization (EM) algorithm [7] to localizing several sound sources and reducing localization errors. In this paper, we quantitatively evaluated this system for moving sounds with various velocities. Then, after evaluating the developed system using created database for moving sounds, we determined the best frame number for training the EM algorithm of our system to cope with moving sounds.

Second, to evaluate our sound source localization system for moving sounds, we developed the moving sound creation tool by using the API library called SoundLocus of Arinis's technology (http://www.arns.com/english/index.html). The conventional ways to evaluate moving sounds required a database to be made by recording moving speakers. This way is a hard way to construct a good database with accurate information about sound tracks. Moreover, since it is very difficult to construct the same database repeatedly, it is difficult to evaluate various sound source localization systems under the same conditions. In contrast, since our moving sound creation tool can create moving sounds including azimuth and distance information according to

Hyund-Don Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno are with Speech Media Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan (e-mail: {hyundon, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp).

the created frame or time, it can be used to evaluate various sound localization systems under the same condition.

Finally, in spite of using only two microphones, a robot implementing our system can distinguish between sounds from the front and sounds from the back by simply rotating its head at least 10 degrees. We evaluated our system's ability to localize moving sounds created by developed moving sound creation tool. The results helped us to determine the rotation speed and rotation angle of the robot's head in order to localize sounds over the entire azimuth range.

The rest of this paper is organized as follows. Section II describes the sound source localization that we developed. Section III describes the new way to evaluate sound localization systems for moving sounds and the results of evaluating our system. Section IV describes two-channel sound localization over entire azimuth range for humanoid robots. Section V concludes this paper.

II. Sound Source Localization

For sound source localization, the latest systems for robots mostly use one of three methods: head-related transfer function (HRTF) [1,8,9], multiple signal classification (MUSIC) [2,10], and CSP [6]. Although HRTF and MUSIC typically need impulse response data and an array of microphones in order to localize several sound sources, CSP does not need impulse response data and can accurately determine the direction of a sound using only two microphones. Using CSP with two microphones can locate only one sound source each frame even if several sound sources are present. This is because CSP obtains the sound localization information from the spatial correlation between two signals. Besides, CSP is usually unreliable in noisy environments. To overcome these weaknesses, we developed a new method based on probability for estimating the number and location of sound sources. First, the CSP results for three frames (shifting every half frame) are collected. Then, an EM algorithm [7] is used to estimate the distribution of the data. In this way, our method can localize several sound sources using the distribution of CSP results and can reduce the error in sound source localization.

A. Cross-power spectrum phase analysis

The direction of a sound source can be obtained by estimating the Time Delay Of Arrival (TDOA) between two microphones [3]. When there is a single sound source, the TDOA can be estimated by finding the maximum value of the cross-power spectrum phase (CSP) coefficients [6] derived by

$$csp_{ij}(k) = IFFT\left[\frac{FFT[s_i(n)]\,FFT[s_j(n)]^*}{\left|FFT[s_i(n)]\right|\left|FFT[s_j(n)]\right|}\right] \quad (1)$$

$$\tau = \arg\max\left(csp_{ij}(k)\right) \quad (2)$$

where k and n are the sampled number for the delay of arrival between two microphones, $s_i(n)$ and $s_j(n)$ are signals entering into microphones i and j, respectively. FFT (or IFFT) is the fast Fourier transform (or inverse FFT), * is the complex conjugate, and τ is the estimated TDOA. The sound source direction is derived by

$$\theta = \cos^{-1}\left(\frac{v\cdot\tau}{d_{\max}\cdot F_s}\right) \quad (3)$$

where θ is the sound direction, v is the sound propagation speed, Fs is the sampling frequency, and d_{max} is the distance with the maximum delay between two microphones. The sampling frequency of our system was 16 kHz. CSP has to consider the diffraction of sounds if microphones are not located in a free space. Therefore, we estimated TDOA for our CSP method after assuming that the shape of the robot's head is a circle. Figure 1 shows the parameters used in equation (3). Here, we assume that waves of sounds received at a pair of microphones become plane waves.

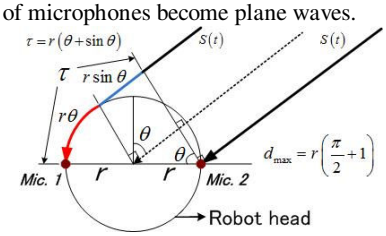

Fig. 1. Localization of multiple sound sources.

B. Localization of multiple sound sources by EM

Figure 2 (A) shows sound source localization events extracted by CSP according to time or frame lapses. Events that lasted 192 ms are used to train the EM algorithm to estimate the number and localization of sound sources.

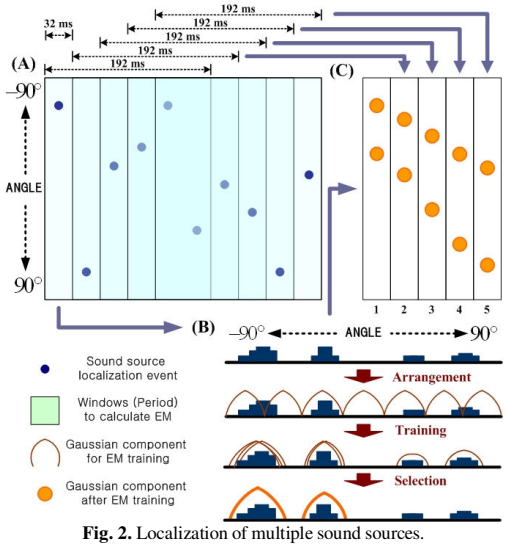

Fig. 2. Localization of multiple sound sources.

The interval for the EM algorithm was experimentally determined as shown in the upper part of Figure 6. Figure 2 (B) shows the training process for the EM algorithm to estimate the distribution of sound source localization events. The EM training results in Figure 2 (C) indicate refined localizations by iterating processes (A) and (B). The interval for EM training is shifted every 32 ms.

Here, we explain the process of applying EM algorithm. Figure 3 describes the process in Figure 2 (B) in detail. In (A) of figure 3, as the first step of EM training, sound source localization events were gathered for 192 ms. Next, Gaussian components defined by using equation (4) for training the EM algorithm were uniformly arranged on whole angles.

$$P\left(X_m|\theta_k\right)=\frac{1}{\sqrt{2\pi\sigma_k^{\,2}}}e^{\frac{\left(X_m-\mu_k\right)^2}{2\sigma_k^{\,2}}} \tag{4}$$

where μ_k is the mean, $\sigma^2{}_k$ is the variance, θ_k is a parameter vector, m is the number of data, and k is the number of mixture components. At that time, in (A) of Figure 3, the μ and σ parameters in Gaussian components are the respective center and radius values of each component. Then, the sound localization events are applied to the arranged Gaussian components to find the parameter vector, θ_k, describing each component density, $P(X_m|\theta_k)$, through iterations of the E and M steps. This EM step is described as follows:

1) E-step: The expectation step essentially computes the expected values of the indicators, $P(\theta_k|X_m)$, where each sound source localization event X_m is generated by component k. Given N is the number of mixture components, the current parameter estimates θ_k and weight w_k, using Bayes' Rule derived as

$$P\left(\theta_k|X_m\right)=\frac{P\left(X_m|\theta_k\right)\cdot w_k}{\sum_{k=1}^{N}P\left(X_m|\theta_k\right)\cdot w_k} \tag{5}$$

2) M-step: At the maximization step, we can compute the cluster parameters that maximize the likelihood of the data assuming that the current data distribution is correct. As a result, we can obtain the recomputed mean using Equation (6), the recomputed variance using Equation (7), and the recomputed mixture proportions (weight) using Equation (8). The total number of data is indicated by M.

$$\mu_k=\frac{\sum_{m=1}^{M}P\left(\theta_k|X_m\right)X_m}{\sum_{m=1}^{M}P\left(\theta_k|X_m\right)} \tag{6}$$

$$\sigma_k^2=\frac{\sum_{m=1}^{M}P\left(\theta_k|X_m\right)\cdot\left(X_m-\mu_k\right)^2}{\sum_{m=1}^{M}P\left(\theta_k|X_m\right)} \tag{7}$$

$$w_k=\frac{1}{N}\sum_{m=1}^{M}P\left(\theta_k|X_m\right) \tag{8}$$

After the E and M steps are iterated an adequate number of times, the estimated mean, variance, and weight based on the current data distribution can be obtained.

Then, in (B) of Figure 3, the weight and mean of Gaussian components are reallocated based on the density and distribution of the histogram data. Finally, in (C) of Figure 3, if the components overlap, each weight value of overlapping Gaussian components will be added. After that, if the weight value is higher than a threshold value, the system can determine the localization of the sound source by computing the average mean of the overlapping Gaussian components. In contrast, components with small weights are regarded as noise and will be removed.

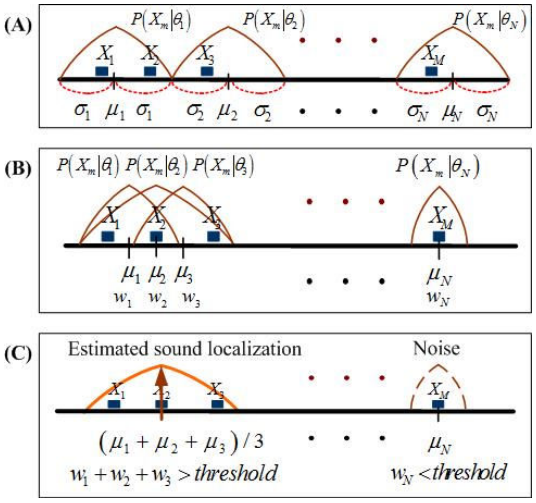

Fig. 3. Process of EM algorithm for estimating sound sources.

C. *Experiments and Results*

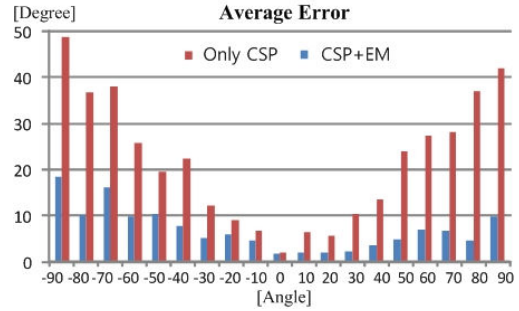

Fig. 4. Localization errors for CSP only and CSP+EM method.

To evaluate the EM algorithm, we experimentally compared the CSP method together with the EM algorithm with the CSP only. We recorded five commands, "sig", "ohayogozaimasu", "konnichiwa", "konbanwa", and "oyasuminasai" trans. the name of our robot, "good morning", "good afternoon", "good evening", and "good

night". They were produced at every 10° from -90° to 90°, at a distance of 1.5 m from the head of the robot, and at a magnitude of 85 dB. Since the robot was at the center of a square room whose side was 5 m where background noise was about 55 dB (A), the reverberation effect was neglected. We calculated the average CSP results of all frames within the interval of five commands for each measurement point. As shown in Figure 4, the average errors with the CSP method and the EM algorithm were less than those with the CSP only method for every angle where the average error indicates the average of difference values between the original point angle and the observed localization angle.

III. Evaluation using Moving Sound Creation Tool

A. SoundLocus Tool

We developed the moving sound creation tool by using the API library called SoundLocus Lite from Arnis's technology. We assumed that the validity of this tool was confirmed because theses and patents of Arnis's technology were already presented in its website (http://www.arns.com/english/index.html). This tool can convert an audio data of a wav file form into a stereo wav file according to the track of desired as shown in Figure 5. Therefore, by designating the velocity and track of moving sounds beforehand, we could freely make moving sounds of stereo wave file form. Since this tool based on head-related transfer function (HRTF) is to create moving sounds for a headphone set, this one does not consider reverberation and ambient noise. Nevertheless, that is effective to evaluate a proposed sound localization method and compare it with other methods under the same condition, i.e., it is unnecessary to consider the error of a track for moving sounds and to reflect dynamically changed resonance and background noses in real environments whenever doing experiments.

To evaluate our method for single moving sounds, we created eight moving speech signals, which were rotated from 0° to 359° at 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1, 1.25 rad/sec at about 2.0 m from the center position with SIG2. The length of each created moving sound was 30 sec. We performed sound localization using these sounds, as shown in (A) of Figure 5. Also, to make certain of the effect of propagation in the air, we have tried to evaluate our sound localization system using created moving sounds emitted by a pair of speakers, as shown in (B) of Figure 5. We used two omni-directional microphones installed at the left and right ear position of the humanoid robot SIG2 (refer to Figure 9) and used two fixed speakers at 0.5 m from the left and right sides of the microphones. To evaluate our method for two moving sounds, we mixed two moving sounds. One rotated at 2 m from the center at 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1, 1.125 rad/sec and the other one rotated at 1 m at the half the angular velocity lagging 90° behind the other one. The middle part of Figure 6 shows the track of moving sounds and the results of localizing two moving.

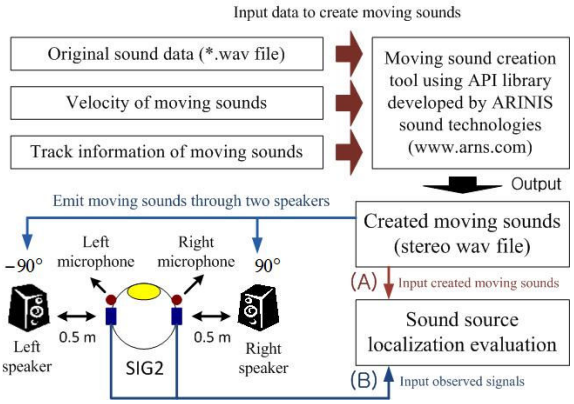

Fig. 5. Creating moving sound sources and experimental conditions.

B. Evaluation

The top part of Figure 6 shows the average error and success rate of localizing single moving sounds according to the number of frames for training the EM algorithm. The success rate is the total percentage when the difference between the original location of the created moving sound source and the estimated sound localization was within 30°. All dotted lines in Figure 6 indicate the results of localizing sounds as observed from speakers shown in (B) of Figure 5. Here, in 6 frames for EM, the average error was the least and the success rate was the best. Therefore, we could experimentally determine that the appropriate interval for our system was 192 ms (6 frames) as shown in Figure 2. Moreover, we learned that our system can cope with moving sounds slower than 1.125 rad/s. Since one of purposes of this study help robots to localize the voices of walking people, we confirmed that our system can cope with moving speech at the average walking speed, 1.0 m/s (1.0 rad/sec at 1 m), of healthy adults. The middle left part of Figure 6 shows that our system localized sounds moving at 1.125 rad/sec for 30 seconds at 2 m. The middle right part of Figure 6 shows that our system localized two sounds moving at 1.125 rad/sec and at 0.563 rad/sec for 30 seconds. The bottom part of Figure 6 shows the average error and success rate of localizing two moving sounds when the number of frames for training the EM algorithm was 6 (192 ms). The two sound sources rotated at different angular velocities. One (source 1) rotated twice as fast as the other one (source 2). The average error and success rate was better for the slower one than for the faster one. The overlapped line, in the graph of success rate of localizing two moving sounds, indicated the percentage of accurate sound localization where two sound sources occurred at the same time. Here, two sound sources have some silent intervals severally because we used the sources recorded from common dialogues. In case of when two moving sounds were emitted from two speakers as shown in the bottom part of Figure 6, the performances were not good because two sounds interfered with each other in the air space.

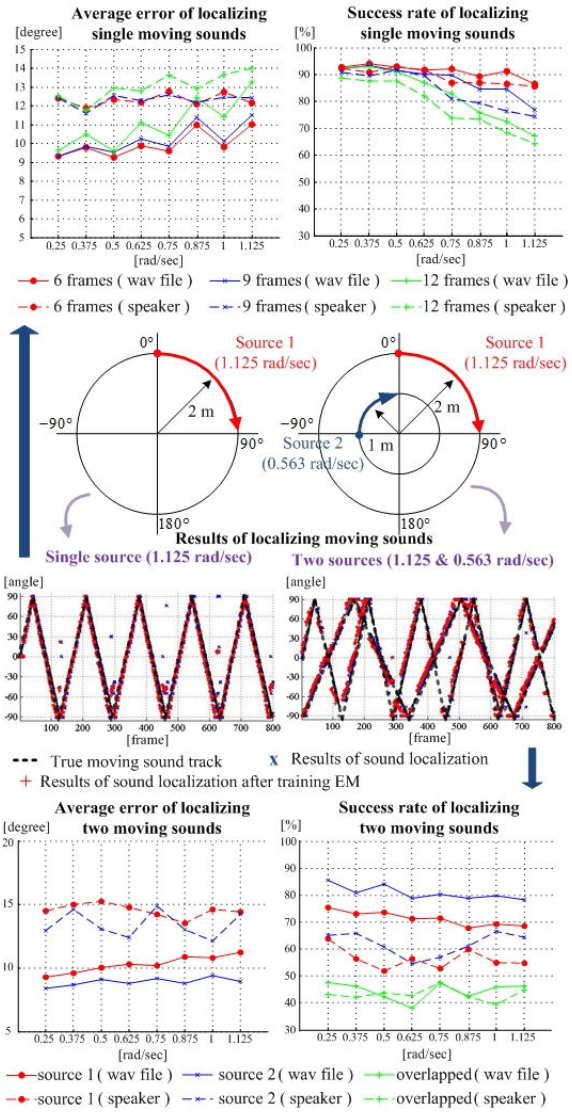

Fig. 6. Results of evaluating our system for moving sounds.

IV. Sound Localization for Humanoid Robots

The target application of our sound source localization method is robots, and it is natural that robots move and rotate their bodies and heads in order to track someone. Therefore, even though the orientation of the microphones in the robot's head or body will constantly change, the sound source localization method must be able to cope with the effects created by the moving microphones. Moreover, if moving robots can track sound sources, they may be able to distinguish whether sound signals are coming from their front or back with only two microphones. This is because the TDOAs and powers obtained for equivalent sound signals coming from the front and back are the same, as shown in (A) of Figure 7.

We can overcome this problem by rotating the robot's head while the sound signals are being generated. For example, as shown in (B) of Figure 7, if sound signals are coming from the front, the robot can determine their direction by reducing the angle of the sound localization while turning its head. As shown in (C), if sound signals are coming from the back, the angle of sound localization will be increased by turning the robot's head. Given this difference, our method can localize the actual source after the robot's head has turned more than 10 degrees.

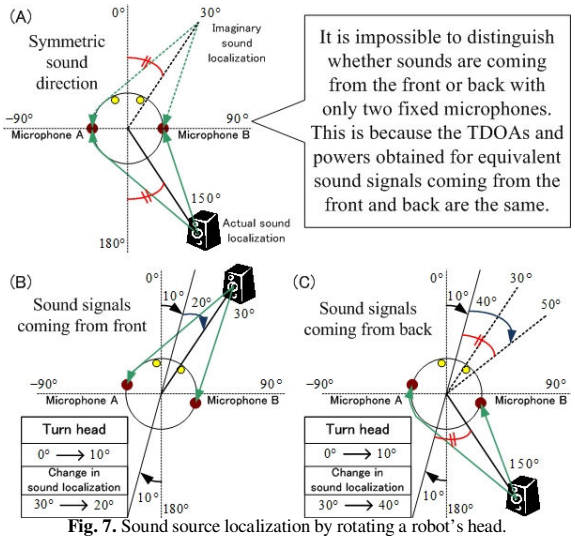

Fig. 7. Sound source localization by rotating a robot's head.

A. Voice Activity Detection using GMM

To localize sounds over the entire azimuth range with two microphones, after the robot first classified speech signals, it has to rotate two microphones during the periods of speech signals. Therefore, we developed a voice activity detection (VAD) based on Gaussian mixture model (GMM). GMM is a powerful statistical method widely used for speech classification [11]. Here, we applied the 0 to 12th coefficients (total 13 values) and the Δ1 to Δ12th coefficients (total 12 values) of Mel Frequency Cepstral Coefficients (MFCCs) to GMM defined by Equation (9) and the weight as denoted by Equation (10).

$$P_{mixture}\left(X_{1\sim25}\middle|\theta_{1\sim25}\right)=\sum_{L=1}^{25}P_L\left(X_L\middle|\theta_L\right)w(L) \tag{9}$$

$$\sum_{L=1}^{25}w(L)=1,\quad 0\le w(L)\le 1 \tag{10}$$

where P is the component density function, L is the number of MFCC parameters, X is the value of the MFCC data of the 0 to 12th and the Δ1 to Δ12th coefficients, and θ is the parameter vector concerning each MFCC value. Moreover,

to classify speech signals robustly, we designed two GMM models for speech and noise derived as

$$f = \log\left(P_s\left(X_s|\theta_s\right)\right) - \log\left(P_n\left(X_n|\theta_n\right)\right) \quad (11)$$

where P_s is the GMM related to speech, and X_s is the MFCC data set at the t-th frame belonging to the speech parameters, θ_s. On the other hand, P_n is the GMM related to noise and X_n is the MFCC data set at the t-th frame belonging to the noise parameters, θ_n. Finally, if the final value, f, denoted as Equation (11), is higher than the value of the threshold to discriminate the speech signal from GMM, signals at the t-th frame will be regarded as speech signals.

$$\begin{aligned} &\text{IF } f(t) > thresold \text{ THEN } f(t) = 1 \ (speech) \\ &\text{ELSE } f(t) = 0 \ (noise) \end{aligned} \quad (12)$$

We used 30 speech data (15 males and 15 females) for the speech parameters to train the GMM parameters, and 77 noise data generated in home environments such as the sounds of a door opening or shutting and those of electrical home appliances (e.g., a vacuum cleaner, a hair drier, and a washing machine) for the noise parameters. To verify the performance of GMM parameter training, we classified the sound sources using speech and noise data for training. As a result, we obtained a success rate for speech classification of 95.5% and a success rate for noise classification of 72.8%.

B. System overview

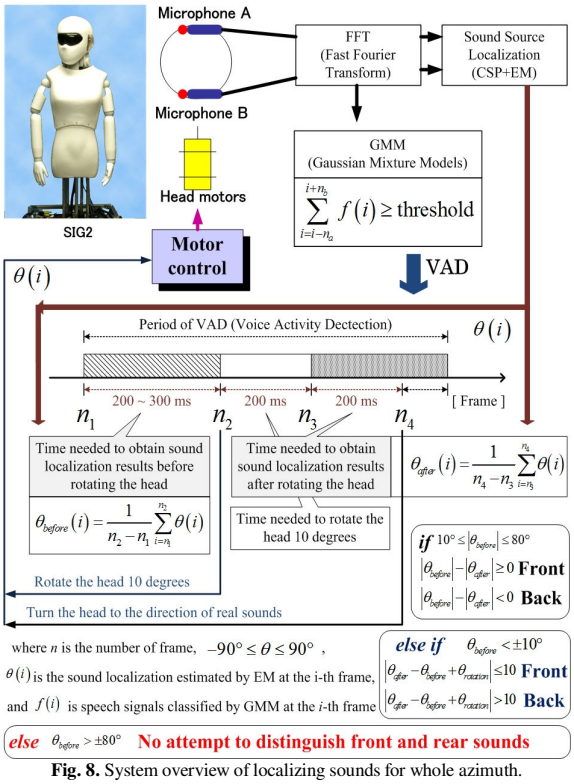

Fig. 8. System overview of localizing sounds for whole azimuth.

In spite of using only two microphones, our system can distinguish between sounds from the front and sounds from the back by simply rotating its head at least 10 degrees. The reason that it can distinguish front from back sources by rotating 10 degrees is that the error margin for a single moving sound is about 10 degrees, as shown in Figure 6. Figure 8 shows the process to localize sounds over the entire azimuth range for a humanoid robot performing the following steps:

1) The robot detects speech signals classified by Gaussian mixture model (GMM). Our voice activity detection (VAD) requires at least 200 ms in order to discriminate between speech signals and noises. The robot can then detect the period of these signals by using

$$\sum_{i=i-n_a}^{i+n_b} f(i) \geq \text{threshold} \quad (13)$$

where $f(i)$ is the i-th speech frame classified by equation (12). If some speech frames exist within the interval of designated frames from the n_a-th frame to the n_b-th frame, we can decide that the i-th frame is within the interval of the target speech.

2) Before turning its head 10 degrees in the direction of the detected signals, the robot calculates the average of the sound localization events by using

$$\theta_{before}(i) = \frac{1}{n_2 - n_1}\sum_{i=n_1}^{n_2}\theta(i) \quad (14)$$

where $\theta(i)$ is the estimated sound localization event of the i-th frame and θ_{before} is the average angle between the n_1-th frame and the n_2-th frame.

3) After turning its head 10 degrees, the robot obtains the average of the sound localization events between the n_3-th frame and the n_4-th frame for 200 ms by using

$$\theta_{after}(i) = \frac{1}{n_4 - n_3}\sum_{i=n_3}^{n_4}\theta(i) \quad (15)$$

4) Finally, using the difference between the initial average angle calculated by equation (14) and the final average angle calculated by equation (15), the robot can localize sounds over the entire azimuth range and turn its head to that direction.

The system can logically distinguish between front and back localization if sound signals are continuously generated for longer than 0.7 seconds (Figure 8). This is because our system has a delay of more than 200 ms for detecting speech signals, 200 ms for rotating the motor 10 degrees, and more than 200 ms for localizing sounds after turning its head. Here, we rotated the head motor less than 0.25 rad/sec (0.25 rotations per 1 second) in order to avoid the effect of motor noises. We confirmed that the magnitude of our motor noise is less than 55 dB(A) when rotating that less than 0.25 rad/sec, at that time, our sound localization system could work without the disturbing noise generated from the motor. In addition, within $\pm 80°$ to $\pm 100°$, our

system does not try to distinguish front and back localizations because sounds are coming from the side in these cases as shown in Figure 9. Besides, although our sound localization system over entire azimuth range has been evaluated for fixed sounds, it would be able to cope with linearly moving sounds slower than the average walking speed, 1 m/s, of healthy adults. In the future work, we are considering the evaluation of our system to distinguish whether the direction of linearly moving sounds is the front or rear by rotating two microphones.

C. *Experiments and results*

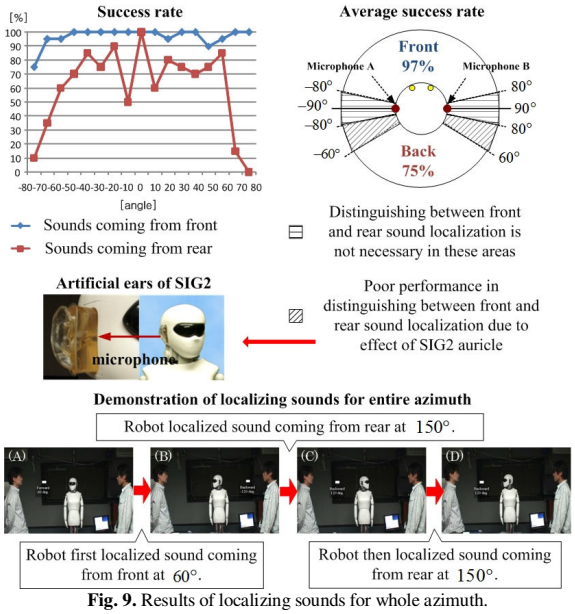

Fig. 9. Results of localizing sounds for whole azimuth.

Figure 9 shows the results of applying our system to the robot called SIG2. In this experiment, the robot distinguished between sounds coming from the front and back whenever speech signals of "sig", its name, were generated. The length of the speech signals was about 0.75 seconds, and speech signals were generated 20 times at each position. The left part of Figure 9 shows the success rate of distinguishing the right sound localization. In right part of Figure 9, the robot obtained a success rate of 97% in the forward area, and The success rate in the backward areas excluding ±70° and ±80° was 75%. We analyzed that the performance in the backward areas was not good because of the effect of the artificial ears installed at SIG2. The bottom part of Figure 9 shows that SIG2 performed entire azimuth sound localization by rotating its head. In this experiment setup, two talkers were at 60° and -60° and when the talker 1 who was at 60° called SIG, it first localized the front sounds at 60° and the talker 2 was located at -150° in (A). Next, the robot localized the back sounds at -150° from (B) to (C) when the talker 2 called "sig". It then localized the back sound at 150° from (C) to (D) when the talker 1 called "sig" again.

V. Conclusion

We proposed the way that can repeatedly evaluate sound source localization under the same conditions regardless of the kind of localization method and number of microphones. We developed a two-channel sound source localization method incorporating a cross-power spectrum phase (CSP) analysis and the EM algorithm. Tests showed that our method can reliably locate sounds moving slower than 1.125 rad/sec. Also, to localize sounds over the entire azimuth with two microphones, we developed a system that can distinguish whether sound signals are coming from the front or back of a robot by rotating the robot's head. In the future work, we will design robots that can communicate with people by adding speech recognition with a source separation function and voice synthesis to our system.

References

[1] Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi Mizoguchi, Hiroshi G. Okuno, and Hiroaki Kitano, "Real-Time Auditory and Visual Multiple-Object Tracking for Humanoids," in Proc. of 17th International Joint Conference on Artificial Intelligence (IJCAI-01), Seattle, Aug. (2001) pp. 1425-1432.

[2] 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," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2004), Oct. (2004) pp. 2404-2410.

[3] H-D. Kim, J. S. Choi, and M. S. Kim, "Speaker localization among multi-faces in noisy environment by audio-visual integration", in Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA2006), May (2006) pp. 1305-1310.

[4] J-M. Valin, J. Rouat, and F. Michaud, "Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2004), Sep. (2004) pp. 2123-2128.

[5] H.L. Van Trees, Ed., *Optimum arrays processing*, John Wiley & Sons, 2002.

[6] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of multiple sound sources based on a CSP analysis with a microphone array," IEEE/ICASSP Int. Conf. Acoustics, Speech, and Signal Processing, June (2000) pp 1053-1056.

[7] T. K. Moon. "The Expectation-Maximization algorithm," IEEE Signal Processing Magazine, Nov. (1996) 13(6) pp. 47-60.

[8] C. I. Cheng & G. H. Wakefield, "Introduction to Head-Related transfer Functions (HRTFs): Space," Journal of the Audio Engineering Society, vol. 49, no. 4, pp.231-248, 2001.

[9] S. Hwang, Y. Park, and Y. Park, "Sound Source Localization using HRTF database," in Proc. Int. Conf. on Control, Automation, and Systems (ICCAS2005), June, 2005, pp.751-755.

[10] R. O. Schmidt, "Multiple Emitter Location and Signals Parameter Estimation," IEEE Trans. Antennas Propag., AP-34, 1986, 276-280.

[11] M. Bahoura and C. Pelletier, "Respiratory Sound Classification using Cepstral Analysis and Gaussian Mixture Models," IEEE/EMBS, pp. 9-12, Sep. 2004.

[12] H-D. Kim, K. Komatani, T. Ogata, H. G. Okuno, "Auditory and visual integration based localization and tracking of humans in daily-life environments", IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2007), Oct. (2007) pp. 2021-2027.