Proceedings of the 2001 IEEE/RSJ
International Conference on Intelligent Robots and Systems
Maui, Hawaii, USA, Oct. 29 - Nov. 03, 2001

Epipolar Geometry Based Sound Localization and Extraction for Humanoid Audition

Kazuhiro Nakadai†, Hiroshi G. Okuno†*, and Hiroaki Kitano†‡
†Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.
Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya-ku, Tokyo 150-0001, Japan
Tel: +81-3-5468-1661, Fax: +81-3-5468-1664
* Department of Intelligence Science and Technology, Kyoto University
‡Sony Computer Science Laboratories, Inc.
nakadai@symbio.jst.go.jp, okuno@nue.org, kitano@csl.sony.co.jp

Abstract

Sound localization for a robot or an embedded system is usually solved by using Interaural Phase Difference (IPD) and Interaural Intensity Difference (IID). These values are calculated by using Head-Related Transfer Function (HRTF). However, HRTF depends on the shape of head and also changes as environments changes. Therefore, sound localization without HRTF is needed for real-world applications. In this paper, we present a new sound localization method based on auditory epipolar geometry with motion control. Auditory epipolar geometry is an extension of epipolar geometry in stereo vision to audition, and auditory and visual epipolar geometry can share the sound source direction. The key idea is to exploit additional inputs obtained by motor control in order to compensate damages in the IPD and IID caused by reverberation of the room and the body of a robot. The proposed system can localize and extract simultaneous two sound sources in a real-world room.

Keywords : sensor fusion, humanoid, localization, active audition

1 Introduction

Auditory processing in robots is important for understanding the surrounding environment and compensating narrow visual field. However, auditory processing in robotics has not been studied so much. Some robots developed so far had a capability of sound source localization [1, 2]. They can localize only a sound source, while we usually hear mixture of sounds. *Computational Auditory Scene Analysis (CASA)* has been studied for understanding mixture of sounds. We apply CASA concepts to humanoid robot and aim to realize a robot which can understand mixture of sounds. Then, in this paper, we focus on sound source extraction and localization, which can work even when multiple sound sources exist.

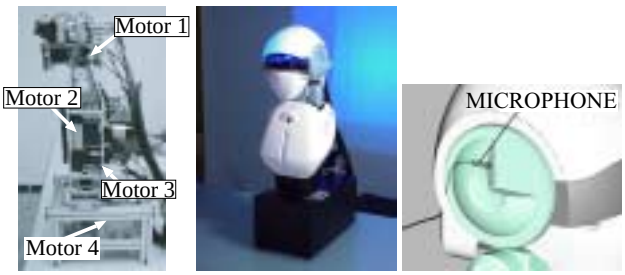
Generally, *Head Related Transfer Function* (HRTF) is used for sound source localization in binaural processing. HRTF depends on each person or embedded system due to the difference of its head shape. It also changes as the environment changes. In some cases, HRTF's change is quite drastic and mobile systems may fail in sound localization based on HRTF. Therefore, sound source localization without HRTF is necessary for real-world or at least dynamically varying environments.

As a testbed of integration of perceptual information in the real world with motor control, we designed a humanoid robot (hereafter, referred to as *SIG*) [3].

The mechanical structure of *SIG* is shown in Fig. 1(a). *SIG* has 4 DOFs of body driven by 4 DC motors, a pair of CCD cameras of Sony EVI-G20, and omnidirectional microphones of Sony electret condenser microphone ECM-77S. It has the cover and simple pinnae as shown in Fig. 1(b). Microphones are installed inside the pinnae, but they have small holes to capture sounds directly from outside the cover as shown in Fig. 1(c).

The paper is organized as follows: Section 2 presents auditory epipolar geometry for sound source localization without HRTF. Section 3 presents issues in auditory epipolar geometry. Section 4 proposes a new sound source localization system by integration of auditory epipolar geometry and motor control. Section 5 shows evaluation of the system, and last two sections give discussion and conclusion.

0-7803-6612-3/01/$10.00©2001 IEEE



(a) Mechanical Structure (b) Cover (c) Microphone

Figure 1: *SIG* the humanoid

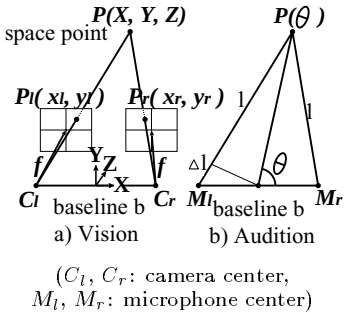


Figure 2: Epipolar geometry for localization

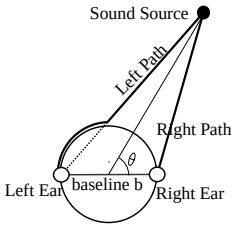


Figure 3: Influence of Cover

2 Auditory Epipolar Geometry

Auditory Epipolar Geometry is proposed to extract directional information of sound sources without using HRTF [4]. In stereo vision research, epipolar geometry is one of the most common localization method [5]. Auditory epipolar geometry is an extension of epipolar geometry in vision (hereafter, *visual epipolar geometry*) to audition as shown in Fig 2. Since auditory epipolar geometry extracts directional information by using the geometrical relation, it can dispense with HRTF.

Auditory epipolar geometry works as follows. First, it extracts peaks by using *Fast Fourier Transforms* (FFT) for each subband, and then calculates the *Interaural Phase Difference (IPD)* as the difference of phases between right and left peaks. The sound source direction is estimated by Eq. (1):

$$\cos\theta = \frac{v}{2\pi f b}\triangle\varphi \tag{1}$$

where v is the velocity of sound, b is the distance (baseline) between left and right microphones, $\Delta\varphi$ is *IPD* and f is the frequency of sound. In this paper, the velocity of sound is fixed to 340m/sec and is invariant to the temperature and humidity.

Since the baselines for vision and audition are in parallel in *SIG*, whenever sound source is localized by visual epipolar geometry, it can be easily converted into the angle θ. This means that symbolic representation of direction is used as a clue to integrate visual and auditory information, and we reported the feasibility of such integration based on epipolar geometry [4].

Auditory epipolar geometry has two problems. One is the influence of the cover of *SIG*, and the other is influence of real-world environments such as reverberation of the room and the body of *SIG*. As the first problem, a sound can reach at most an ear directly. For example, in Fig. 3, a sound has to travel along the cover because left path is not direct to the left ear from a sound source. The problem is solved by adjusting the formula of auditory epipolar geometry by taking the shape of *SIG* into account. The new formula is specified as follows:

$$\theta = F^{-1}\left(\frac{v}{2\pi f}\Delta\varphi\right) \tag{2}$$

where F represents the difference of distance between left and right ears from an infinite sound source.

$$F(\theta) = \begin{cases} \left(\theta - \frac{\pi}{2}\right)\frac{b}{2} + \frac{b}{2}cos\,\theta & \left(0 \le \theta \le \frac{\pi}{2}\right) \\ \left(\theta - \frac{\pi}{2}\right)\frac{b}{2} - \frac{b}{2}cos\,\theta & \left(\frac{\pi}{2} < \theta \le \pi\right) \end{cases} \tag{3}$$

The second problem is solved by measurement and simulation of various parameters, which is described in the next section.

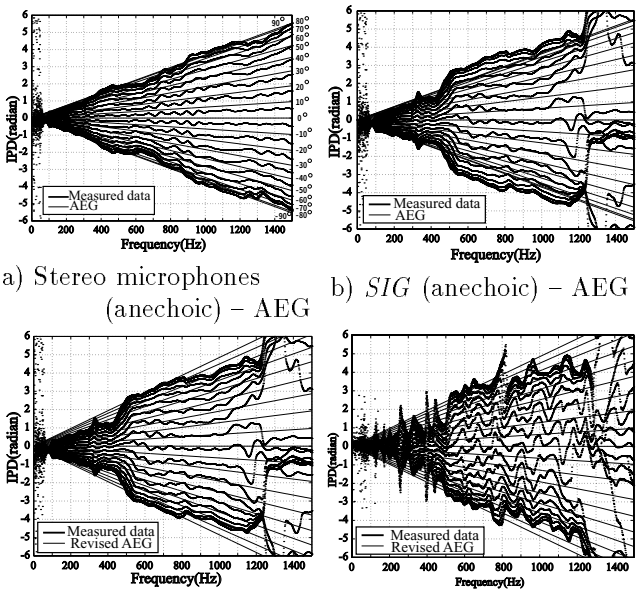
3 Issues in Auditory Epipolar Geometry

IPD and IID(*Interaural Intensity Difference*) depend on the following three major factors:

1. Distance difference between left and right ears from a sound source
2. Reverberation of robot body and head
3. Reverberation of room

To investigate influences of these factors in real-world, we measured acoustics of *SIG*. In addition, to clear influence by the factor which is difficult to analyze by measurements, we simulated the acoustics by Boundary Element Method (BEM) using SYSNOISE[1] .

[1] Computational Vibro-Acoustics software, Copyright LMS International 1999.



a) Stereo microphones (anechoic) – AEG

b) *SIG* (anechoic) – AEG

c) *SIG* (anechoic) – AEG(revised)

d) *SIG* (non-anechoic) – AEG (revised)

Figure 4: IPD Measurement(AEG: auditory epipolar geometry)

We measured impulse responses in the range of $\pm 90°$ from the median plane of *SIG* in the horizontal plane by 10° in an anechoic room. Fig. 4(a) shows the result of acoustic measurement using exposed stereo microphones without the cover. In this case, we do not need to take any influence of the cover into account. Thin lines labeled "AEG" means IPD estimated by Eq.(1), that is, auditory epipolar geometry. The estimation by auditory epipolar geometry fits corresponding measured data well. This proves that the principle of auditory epipolar geometry is correct.

Fig. 4(b) shows comparison between IPD of *SIG* in an anechoic room and IPD estimated by Eq. (1). The estimation by auditory epipolar geometry does not fit corresponding data at frequencies of higher than 300 Hz. This misfit is caused by reverberation of *SIG* body and head, where HRTF models all influence the input.

Fig. 4(c) shows a comparison between measured IPD of *SIG* in an anechoic room and estimated IPD by Eq. (2). IPD is estimated better than one in Fig. 4(b). This indicates that the misfit problem of cover influence is overcome by the revised auditory epipolar geometry by Eq. (2).

Fig. 4(d) shows the result of measurements in a prepared non-anechoic room. This room is $10m^2$ and has sound absorbing material attached to walls, ceiling and floor. Measured IPD is distorted by room acoustics. The revised auditory epipolar geometry does not work well at the frequencies of higher than 1200 Hz because the range of IPD exceeds $\pm 2\pi$ corresponding to the baseline of *SIG*'s ears. In every case, the sensitivity against the directional change of around 0° is higher.

Next, we analyzed the influence of room reverberation using SYSNOISE. Fig. 5 shows IPD and IID at 30°. IPD and IID labeled "SYSNOISE (no floor)" are calculated using 3D mesh data of *SIG* head, and have peaks between 300 and 400 Hz. These peaks are caused by *SIG* head. IPD and IID measured with *SIG* also have peaks between 300 and 400 Hz by the cover.

We also calculated IPD and IID under the condition that a floor exists at the distance of $1m$ below *SIG* . In comparison with the case of no floor, more peaks are found. Thus, even a simple floor causes undulations of IPD and IID. Therefore, it is necessary for sound source localization to consider acoustic environments.

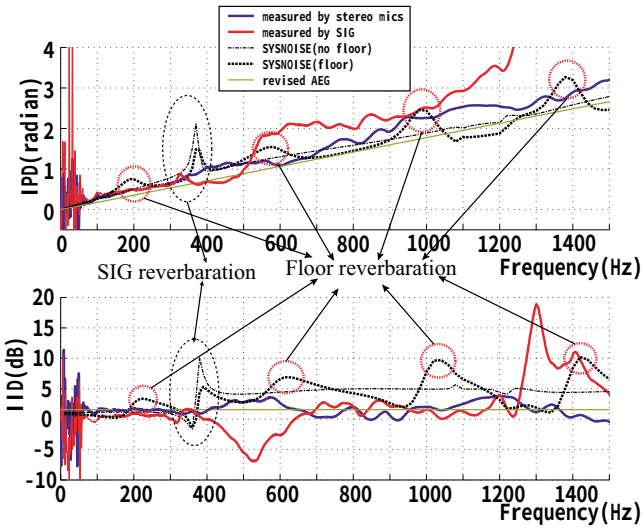


Figure 5: Measured and simulated IPD and IID at 30°

We showed auditory epipolar geometry method is robust against the first and second factors by using revised auditory epipolar geometry. Although we can localize sound sources using HRTF, HRTF needs measurement in an anechoic room. On the other hand, auditory epipolar geometry can estimate sound source direction mathematically and it does not need measurement any longer.

As for the last factor, we cannot always measure room acoustics in advance, especially in unknown acoustic environments. It is difficult and time-consuming to simulate room acoustics by SYSNOISE. We propose

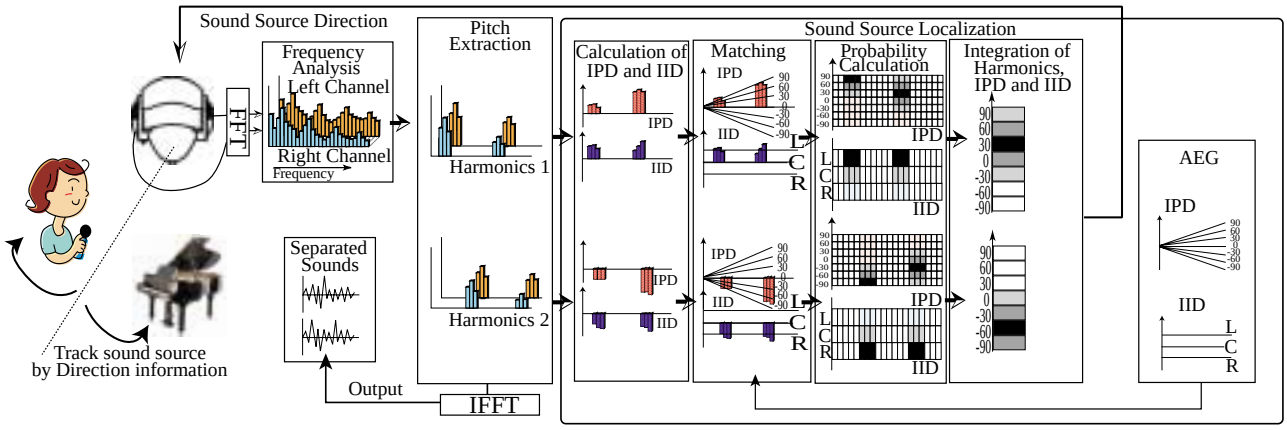


Figure 6: Sound source localization system by auditory epipolar geometry with motion control

a new sound source localization method to obtain accurate direction by integrating with robot motion in weak reverberation environments. The next section explains a system with the new localization method.

4 Localization by Auditory Epipolar Geometry with Motion Control

In order to trace a specific sound source where multiple sound sources exist, we develop sound source localization system by integrating auditory epipolar geometry with motion control.

Fig. 6 shows the sound source localization system by auditory epipolar geometry with motion control. The system takes as input a mixture of sounds which come from different directional sound sources. The input signal is sampled with sampling frequency of 48 KHz and 16-bit quantization, and converted into spectrum by FFT. By pitch extraction and sound source localization, sound sources are separated with the direction, and *SIG* can turn to the direction of a specific sound source.

4.1 Pitch Extraction

Pitches are extracted as follows:

- Filtering a power spectrum
  The filter is designed as a band-pass filter from 90 to 3000 Hz. In the pass band, frequencies where the power is more than a threshold th are to be passed. The threshold is determined experimentally. The filtered spectrum is represented as

$$S_f(t) = \{S_p(f_i,t) | 90 < f_i < 3000\text{Hz}, S_p(f_i,t) > th\} \quad (4)$$

  where f_i is frequency, t is time and Sp is power spectrum of the input sound.

- Clustering peaks in a filtered spectrum
  First, clusters are created every sub-band in S_f. Then, when there is neighboring sub-bands between two clusters, they are unified as one cluster. This processing is repeated until no neighboring subbands between any two clusters are found. As a result, C is obtained as a set of cluster C_ks.

$$C(t) = \{C_k(t) | 1 \le k \le n\} \quad (5)$$

  where n is the number of clusters in C.

- Extracting peaks from clusters
  A sub-band which has the strongest power in each cluster C_k is extracted as a peak P_k. Then P is extracted as a set of peaks in C.

$$P(t) = \{P_k(t) | P_k(t) = max(C_k(t)), 1 \le k \le n\} \quad (6)$$

- Extracting sounds by harmonic relationships
  Peaks which have a harmonic relationship with each other are extracted as a sound by

$$S_j(t) = \{P_i(t) | Harmonics(P_i(t), P_{F0}(t))\} \quad (7)$$

  where S_j is the jth sound and P_{F0} is a peak which has a fundamental frequency in S_j.

4.2 Matching using Auditory Epipolar Geometry

The sound source direction θ is assumed within $\pm 90°$ by $10°$ from the median plane of *SIG*. The system uses IPD and IID for sound source localization. As for sound source localization by IPD, the system creates hypotheses of IPD by Eq. (2) for each θ. Then it calculates the distance between each hypothesis and

IPD of input harmonics at frequencies of lower than 1200 Hz. The cost function is as follows:

$$d(\theta) = \frac{1}{n_{f<1200\text{Hz}}} \sum_{f=\text{F0}}^{1200\text{Hz}} (P_h(\theta, f) - P_s(f))^2/f) \quad (8)$$

where P_h and P_s are IPD of hypotheses and S_j, respectively. And $n_{f<1200\text{Hz}}$ is number of harmonics lower than 1200Hz.

Then θ of hypothesis which has a minimum value of d is regarded as sound source direction of the input harmonics.

As for harmonics with more than 1200Hz, IPD is not efficient in localization as mentioned in Section 3. IID has the feature that it is emphasized at high frequency by the head because the wavelength is short. So, we use IID for localization of such harmonics, that is, the system judges whether a sound source is from left or right side of *SIG* by summation of IID at frequency of higher than 1200 Hz. The direction is left side of *SIG* if d has positive value, otherwise it is right side. We can use a matching technique similar to IPD for localization by IID. However, hypothesis creation of IID requires a lot of calculation. We do not use hypothesis matching in localization by IID to make the system work in the real-time.

4.3 Integration of IPD and IID by Dempster Shafer theory

By integration of IID and IPD, the system can get more accurate direction information. Our integration is based on Dempster-Shafer theory, and to apply localization by IPD and IID to this theory, belief factors of IPD are calculated from the distances using probability density function.

$$p(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{d(\theta)-m}{\sqrt{s/n}}} \exp\left(-\frac{x^2}{2}\right) dx \quad (9)$$

where m and s are average and variance of $d(\theta)$, respectively. And n is the total number of d.

Our method judges only whether a sound source is left or right from IID. So, we decided belief factors of IID experimentally as shown in Tab. 1.

d	$90° \rightarrow 40°$	$30° \rightarrow -30°$	$-40° \rightarrow -90°$
+	0.35	0.5	0.65
-	0.65	0.5	0.35

Table 1: IID belief factor

Then, belief factors of IID and IPD are integrated using Dempster-Shafer theory.

$$\begin{aligned} P_{IPD+IID}(\theta) &= P_{IPD}(\theta)P_{IID}(\theta) + (1 - P_{IPD}(\theta))P_{IID}(\theta) \\ &\quad + P_{IPD}(\theta)(1 - P_{IID}(\theta)) \end{aligned} \quad (10)$$

As a result of the integration, θ with maximum $P_{IPD+IID}$ is regarded as a sound source direction.

4.4 Sound Source Tracking

SIG can track a specific sound source using the direction information obtained by sound source localization. When a harmonic sound with a specific pitch is detected, *SIG* acts as follows:

1. *SIG* estimates the direction of the sound using epipolar based sound source localization method.
2. *SIG* turns to the direction of the sound source. The angle to turn is a half of estimated direction. When the angle is less than 10°, the angle is made 10° compulsorily.
3. This process is repeated until one of the following conditions are satisfied.
   - The estimated direction becomes 0°.
   - The number of iteration reaches 10.
   - The estimated direction exceeds the range of ±90°.

Thus, by tracking a sound source, *SIG* can turn to the direction of the sound source. This means that the accuracy of localization is improved because the sensitivity of sound localization is the maximum in the front of the sound source. The next section shows experiments to prove the effectiveness of the system.

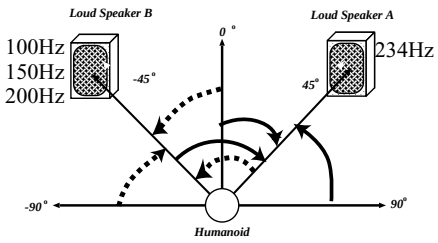


Figure 7: Experiment: Sound source localization with motion

5 Experiments

We evaluate our system through 2 experiments. One is sound source localization and tracking using 2 static sound sources of harmonic sounds with constant $F0$. This experiment proves the effectiveness of basic functions such as sound source extraction, localization and

tracking. The other uses moving human voices for sound sources. It shows that our system can work even when sound source is in motion and its fundamental frequency is changing.

5.1 Experiment 1: static sound sources with constant fundamen tal frequency

There are two sound sources: two B&W Noutilus 805 loud speakers located in our prepared room of 10 square meters as described before.

One sound source A (Loud Speaker A) plays a harmonic sound of 234 Hz. The other sound source B (Loud Speaker B) plays that of 100, 150 or 200 Hz. The initial direction of SIG is either of $0°, \pm 45°, \pm 90°$ as shown in Fig. 7. It turns to the direction of A or B according to focus-of-attention given in advance. W e tracked sound sources using auditory epipolar geometry as well as HRTF for comparison. W e use IPD, IID or both on direction estimation.

In HRTF, we use the same cost function as Eq. (8) for IPD. And for IID, Eq. (11) is used as a cost function.

$$d(\theta) = \frac{1}{n_{f>1200\mathrm{Hz}}} \sum_{f=1200\mathrm{Hz}}^{3000\mathrm{Hz}} (I_h(\theta, f) - I_s(f))^2 f) \quad (11)$$

where I_h and I_s are IID of hypotheses and S_j, respectively. And $n_{f>1200Hz}$ is number of harmonics with higher than 1200 Hz.

Table 2: result of sound source tracking

		Initial Error			Final Error		
		IPD	IID	Both	IPD	IID	Both
0	AEG	1	100%	1	0	35	0
	HRTF	0	12	1	0	6	0
45	AEG	19	100%	19	3	30	3
	HRTF	-10	5	5	-3	7	-2
-45	AEG	-11	100%	-11	0	-5	0
	HRTF	5	-8	5	0	5	0
90	AEG	-10	100%	-10	3	15	3
	HRTF	-30	0	0	-5	-1	-5
-90	AEG	50	100%	3	5	-5	5
	HRTF	53	53	3	60	47	0

(AEG: auditory epipolar geometry)

The total number of benchmarks is 202. For two cases the system failed in tracking because of the range excess of $\pm 90°$. Except for them, the system succeeds in tracking of sound sources. The result is summarized in Tab. 2. The value in a field shows error between true and estimated direction. Because the system finds only information of left or right by estimation using IID in auditory epipolar geometry, percentage of correct direction in the fields is shown.

The results are as follows:

- Sound source tracking is improved by the integration of IID and IPD. This tendency is remarkable when the estimated direction is far from $0°$.
- Localization by auditory epipolar geometry is less accurate than HRTF at initial state. However, this is solved by turning towards the sound source direction.
- Localization for a sound which can be used either IPD or IID works well by sound source localization with motion.
- The sensitivity of both auditory epipolar geometry and HRTF improves by turning to the sound source direction.

5.2 Experimen t 2 : moving human voices

Mr. A and B stand at $15°$ left and $60°$ right to SIG, respectively. Mr. A starts talking at $t = 1.5(s)$. And Mr. B starts talking at $t = 3(s)$. They continue to talk until $t = 17(s)$. We do not have any constraints for their speeches. This time, they speak arbitary phrases such as "hello SIG " and "please turn to me" in Japanese and English continuously. Mr. A starts moving at $t = 6(s)$, while Mr. B does not change his position through the experiment. SIG is configured to track the first speaker, Mr. A. So, SIG turns its body according to Mr. A's direction.

Fig. 8 shows the time sequence of the experiment. A lighter line means motor direction of SIG, and darker two lines mean estimated sound directions of speakers.

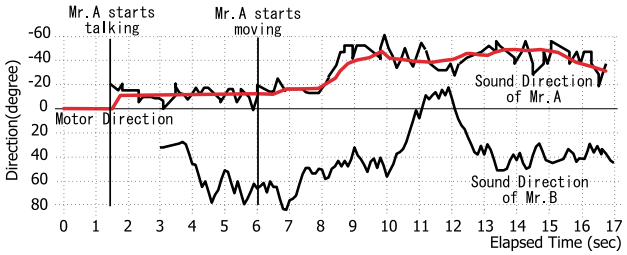


Figure 8: Sound source tracking using SIG

Two sound sources by voice are separated well during the experiment. The error of sound source localization of Mr. A is about $\pm 15°$, judging from difference between motor and sound direction. The error is within the same range even when Mr. A is moving,

motor control works well because *SIG* is tracking Mr. A smoothly.

The error of sound source localization of Mr. B is $\pm 30°$ on average. It is bigger than Mr. A while Mr. B does not move. And it becomes bigger after Mr. A starts moving.

The reasons are as follows:

- The accuracy of sound source localization is low as shown in Experiment 1 since Mr. B is far from the *SIG* front direction.
- The fundamental frequency is changing continuously because sound source is voice. This reduces the robustness against frequency change.
- *SIG* captures motor noises created by its active motion while tracking. This affect sound source localization badly.

This indicates that generally sound source localization is not so accurate, but it can improve by facing and tracking a sound source. Therefore, our sound source localization works well under the natural environment such as moving sound sources with changing fundamental frequency.

6 Discussion and Future work

We tried to localize the same benchmarks by ICA[6], Direction Pass Filter[7], and BiHBSS[8]. However these methods could not separate and localize sound sources in the real world. Our system currently does not cancel inevitable motor noises made by robot itself to perceive sounds in motion. Such noises should be canceled. We already proposed such an active audition system[4, 9]. Integration with active audition system should be needed to be applied in the real world. Evaluation of the system in noisy environments, front-back problem and vertical localization (elevation) without using HRTF are remained as future work.

7 Conclusion

We presented a new sound localization method based on auditory epipolar geometry with motion control. It can localize simultaneous two sound sources in a real-world room. The key idea is to use additional inputs obtained by motor control in order to compensate damages in the IPD and IID caused by reverberation of the room and the body of *SIG*. This idea leads to the active audition system where a robot or embedded system combines motor control, vision, and other sensors for better auditory scene analysis.

Acknowledgments

We thank Dr. Tino Lourens of Kitano Symbiotic Systems Project for his valuable discussions. We also thank Cybernet Systems Co. ,Ltd. for SYSNOISE simulation, and Nittobo Acoustic Engineering Co. ,Ltd. for anechoic room experiments.

References

[1] J. Huang, N. Ohnishi, and N. Sugie, "Separation of multiple sound sources by using directional information of sound source," *Artificial Life and Robotics*, vol. 1, no. 4, pp. 157–163, 1997.

[2] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proceedints of the Sixteenth International Joint Conference on Atificial Intelligence (IJCAI-99)*, pp. 1146–1151, 1999.

[3] H. Kitano, H. G. Okuno, K. Nakadai, T. Sabish, and T. Matsui, "Design and architecture of sig the humanoid," in *Proceedings of International Conference on Robotics and Systems 2000 (IROS 2000)*, pp. 181–190, 2000.

[4] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*, p. *to appear*, AAAI, 2000.

[5] O. D. Faugeras, *Three Dimensional Computer Vision: A Geometric Viewpoint*. MA.: The MIT Press, 1993.

[6] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," in *Proceedings of 1998 International Symposium on Nonlinear Theory and its Applications (NOLTA-98)*, pp. 923–927, 1998.

[7] Y. Nakagawa, H. G. Okuno, and H. Kitano, "Using vision to improve sound source separation," in *Proceedings of 16th National Conference on Artificial Intelligence (AAAI-99)*, pp. 768–775, AAAI, 1999.

[8] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communication*, vol. 27, no. 3-4, pp. 209–222, 1999.

[9] K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano, "Active audition system and humanoid exterior design," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS 2000)*, IEEE, 2000. 1453-1461.