Humanoid Active Audition System

Kazuhiro Nakadai[1], Hiroshi G. Okuno[1,2], Tino Laurens[1], and Hiroaki Kitano[1,3]

[1] Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.,
Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya-ku, Tokyo 150-0001, Japan
[2] Department of Information Sciences, Science University of Tokyo, Noda, Chiba 247-8510, Japan
[3] Sony Computer Science Laboratories, Inc., Shinagawa, Tokyo 141-0022, Japan

Abstract. Perception for humanoid should be active, e.g., by moving its body or by controlling parameters of sensors such as cameras or microphones, to understand environments better. Active vision is one of the common capabilities of a humanoid. Active perception usually makes sounds of actuators, which makes auditory processing more difficult. A conventional solution of this problem in audition processing is the "*stop-perceive-act*" principle; that is, humanoid keeps still to listen to sounds, extracts features from input sounds, makes a plan to act and executes the plan.

In this paper, we propose *active audition* to overcome the shortage of the "*stop-perceive-act*" principle. Its main issues are (1) the cancellation of internal sounds made by actuators to enhance external sounds in input sounds; (2) sound source separation from a mixture of sounds, since input sounds is not a single sound source; (3) sensor fusion because sound source separation is an ill-posed problem.

A cover of a humanoid is used to cancel its internal sounds to enhance outer sounds. Usually sound source separation assumes the influence of a humanoid head, or a *Head-Related Transfer Function* (HRTF), which is measured in an anechoic room for many distinct spatial positions. However, this is time-consuming and the HRTF needs to be remeasured when acoustic environments change. Therefore, we present a new method based on the *epipolar geometry* without using any HRTF.

We present an active audition system for humanoid *"SIG the humanoid"*. The audition system of a humanoid requires localization of sound sources and identification of meanings of the sound in the auditory scene. The active audition reported in this paper focuses on improving sound source tracking by integrating audition, vision, and motor movements. Given the multiple sound sources in the auditory scene, *SIG* actively moves its head to improve localization by aligning microphones orthogonal to the sound source and by capturing the possible sound sources by vision. However, such an active head movement inevitably makes motor noises. The system must adaptively cancel motor noises using motor control signals. The experimental result demonstrates that the active audition by integration of audition, vision, and motor control enables sound source tracking in various conditions.

1 Introduction

The goal of the research reported in this paper is to establish a technique of multi-modal integration for improving perception capabilities of a humanoid. We use an upper-torso humanoid as a platform of the research, because we believe that multi-modality of perception and high degree-of-freedom is essential to simulate intelligent behavior of human. Among various perception channels, this paper reports active audition that integrates audition with vision and motor control.

Active perception is an important research topic that requires integration of perception and behavior. A lot of research has been carried out in the area of active vision, because it will provide a framework for obtaining necessary additional information by integrating vision with behaviors, such as control of optical parameters or actuating camera mount positions. For example, an observer controls the geometry parameters of the sensory apparatus in order to improve the quality of the perceptual processing [1]. Such activities include moving a camera or cameras (vergence), changing focus, zooming in or out, changing camera resolution, widening or narrowing iris and so on. Therefore, active vision system is always integrated with servo-motor system, which means that active vision system is in general associated with motor noises.

The concept of active perception also can be extended to audition, which is "*active audition*" we propose [15, 16, 18]. Binaural researches report that we can localize sounds using not only *Head-Related Transfer Function* (*HRTF*) but subtle changes of *Interaural Time / Phase Difference* (*ITD / IPD*) caused by continuously head movements. Their experiments show that it is difficult to localize sound sources especially in front when we completely stop, that is, we can not use information of the ITD/IPD changes. This claims that motion is indispensable to

[0] Email: nakadai@symbio.jst.go.jp, tino@symbio.jst.go.jp, okuno@nue.org, kitano@csl.sony.co.jp

understand and simulate human audition. Actually, this kind of knowledge is utilized to synthesize 3D surround sounds in virtual reality field.

Audition is also always active since people hear a mixture of sounds and focus on some parts of input. Usually, people with normal hearing can separate sounds from a mixture of sounds and focus on a particular voice or sound even in a noisy environment. This capability is known as the *cocktail party effect*. While traditionally, auditory research has been focusing on human speech understanding, understanding auditory scene in general is receiving increasing attention. Computational Auditory Scene Analysis (CASA) studies a general framework of sound processing and understanding [5, 7, 20, 24]. Its goal is to understand an arbitrary sound mixture including speech, non-speech sounds, music and other sounds in various acoustic environment. It requires not only understanding of meaning of specific sound, but also identification of spatial relationship of sound sources, so that sound landscapes of the environment can be understood. This leads to the need of active audition that has capability of dynamically focusing on specific sound in a mixture of sounds, and actively controlling motor systems to obtain further information using audition, vision, and other perceptions.

1.1 Audition for Humanoids in Daily Environments

Our ultimate goal is to deploy our robot in daily environments. For audition, this requires the following issues to be resolved:

- Ability to localize sound sources in unknown acoustic environment.
- Ability to actively move its body to obtain further information from audition, vision, and other perceptions.
- Ability to separate and identify sound sources even in motion.
- Ability to continuously perform auditory scene analysis under noisy environment, where noises come from both environment and motor noises of robot itself.

First of all, deployment to the real world means that the acoustic features of the environment is not known in advance. In the current computational audition model, the HRTF was measured in the specific room environment, and measurement has to be redone if the system is installed at different room. It is infeasible for any practical system to require such extensive measurement of the operating space. Thus, audition system without HRTF is an essential requirement for practical systems. The system reported in this paper implements epipolar geometry-based sound source localization that eliminates the need for HRTF. The use of epipolar geometry for audition is advantageous when combined with stereo vision systems because many stereo vision systems use epipolar geometry for visual object localization.

Second, active audition that integrates audition, vision, and motor control system is critical. Active audition can be implemented in various aspects. Take the most visible example, the system should be able to dynamically align microphone positions against sound sources to obtain better resolution. Consider that a humanoid has a pair of microphones. Given the multiple sound sources in the auditory scene, the humanoid should actively move its head to improve localization (getting the direction of a sound source) by aligning microphones orthogonal to the sound source. Aligning a pair of microphones orthogonal to the sound source has several advantages:

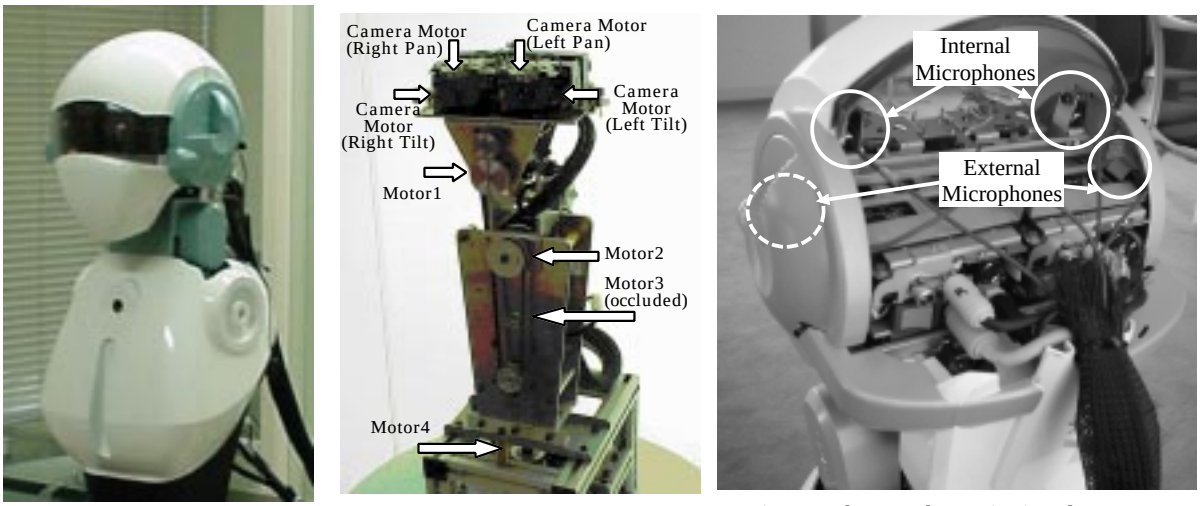
- Each channel receives the sound from the sound source at the same time.
- It is rather easy to extract sounds originating from the center by comparing subbands in each channel.
- The problem of front-behind sound from such sound source can be solved by using direction-sensitive microphones.
- The sensitivity of direction in processing sounds is expected to be higher along the center line, because sound direction is represented by a *sine* function.
- Zooming of audition can be implemented by using nondirectional and direction-sensitive microphones.

Therefore, *gaze stabilization* for microphones is very important to keep the same position relative to a target sound source.

Active audition requires movement of the components that mounts microphone units. In many cases, such a mount is actuated by motors that make considerable noises. In a complex robotic system, such as humanoid, motor noises are complex and often irregular because a number of motors may be involved in the head and body movement. Removing motor noise from auditory system requires information on what kind of movement the robot is making in real-time. In other words, motor control signals need to be integrated as one of the perception channels. If dynamic noise canceling of motor noise fails, one may end-up using "*stop-perceive-act*" principle reluctantly, so that the audition system can receive sound without motor noises. To avoid using such an implementation, we implemented an adaptive noise canceling scheme that uses motor control signal to predict and cancel a motor noise.

For humanoid audition, active audition and the CASA approach is essential. In this paper, we investigate a new sound processing algorithm based on epipolar geometry without using HRTF, and internal sound cancellation algorithms.

1.2 SIG the humanoid



a) Cover b) Mechanical structure c) Internal microphones (top) and cameras

Fig. 1. *SIG* the Humanoid

As a testbed of integration of perceptual information to control motor of high degree of freedom (DOF), we designed a humanoid (hereafter, referred as *SIG*) with the following components [12]:

- 4 DOFs of body driven by 4 DC motors — Its mechanical structure is shown in Fig. 1b. Each DC motor is controlled by a potentiometer.
- A pair of CCD cameras of Sony EVI-G20 for visual stereo input — Each camera has 3 DOFs, that is, pan, tilt and zoom. Focus is automatically adjusted. The offset of camera position can be obtained from each camera (Fig. 1b).
- Two pairs of nondirectional microphones (Sony ECM-77S) (Fig. 1c). One pair of microphones are installed at the ear position of the head to gather sounds from the external world. Each microphone is shielded by the cover to prevent from capturing internal noises. The other pair of microphones are installed very close to the corresponding microphone to gather sounds from the internal world.
- A cover of the body (Fig. 1a) reduces sounds to be emitted to external environments, which is expected to reduce the complexity of sound processing.

The paper is organized as follows: In Section 2, humanoid audition is discussed from the viewpoints of computational auditory scene analysis for humanoid. Section 3 presents our active audition system with new sound source separation to the problems of active perception. Section 4 shows experiment to prove the effectiveness of our active audition system. Last two sections give discussion, future work and conclusions.

2 New Issues of Humanoid Audition

This section describes our motivation of humanoid audition and some related work. We assume that a humanoid or robot will move even while it is listening to some sounds. Most robots equipped with microphones developed so far process sounds without motion [10, 14, 26]. This "*stop-perceive-act*" strategy, or hearing without movements, should be conquered for real-world applications. For this purpose, hearing with robot movements imposes us various new and interesting aspects of existing problems.

The main problems with humanoid audition during motion are understanding general sounds, sensor fusion, active audition, and internal sound cancellation.

2.1 General Sound Understanding

Since computational auditory scene analysis (CASA) research investigates a general model of sound understanding, input sound is a mixture of sounds, not a sound of single source. One of the main research topics of CASA is *sound stream separation*, a process that separates sound streams that have consistent acoustic attributes from a mixture of sounds. Three main issues in sound stream separation are

1. Acoustic features used as clues of separation,
2. Real-time and incremental separation, and
3. Information fusion — discussed separately.

In extracting acoustic attributes, some systems assume the humans auditory model of primary processing and simulate the processing of cocklear mechanism [5, 25]. Brown and Cooke designed and implemented a system that builds various auditory maps for sound input and integrates them to separate speech from input sounds [5].

Nakatani *et al*. used harmonic structures as the clue of separation and developed a monaural-based harmonic stream separation system, called HBSS [20]. HBSS is modeled by a multi-agent system and extracts harmonic structures *incrementally*. They extended HBSS to use binaural (stereo microphone embedded in a dummy head) sounds and developed a binaural-based harmonic stream separation system, called Bi-HBSS [21]. Bi-HBSS uses harmonic structures and the direction of sound sources as clues of separation. Okuno *et al.* extended Bi-HBSS to separate speech streams, and uses the resulting system as a front end for automatic speech recognition [23].

2.2 Sensor Fusion for Sound Stream Separation

Separation of sound streams from perceptive input is a nontrivial task due to ambiguities of interpretation on which elements of perceptive input belong to which stream [19]. For example, when two independent sound sources generate two sound streams that are crossing in the frequency region, there may be two possibilities; crossing each other, or approaching and departing. The key idea of Bi-HBSS is to exploit spatial information by using a binaural input.

Staying within a single modality, it is very difficult to attain high performance of sound stream separation. For example, Bi-HBSS finds a pair of harmonic structures extracted by left and right channels similar to stereo matching in vision where camera are aligned on a rig, and calculates the ITD / IPD, and/or the *IID* / *IAD* (*Interaural Intensity / Amplitude Difference*) to obtain the direction of sound source. The mapping from ITD, IPD, IID and IAD to the direction of sound source and vice versa is based on the HRTF associated to binaural microphones. Finally Bi-HBSS separates sound streams by using harmonic structure and sound source direction.

The error in direction determined by Bi-HBSS is about $\pm 10^\circ$, which is similar to that of a human, i.e. $\pm 8^\circ$ [6]. However, this is too coarse to separate sound streams from a mixture of sounds.

Nakagawa *et al*. improved the accuracy of the sound source direction by using the direction extracted by image processing, because the direction by vision is more accurate [19]. By using an accurate direction, each sound stream is extracted by using a *direction-pass filter*. In fact, by integrating visual and auditory information, they succeeded to separate three sound sources from a mixture of sounds by two microphones. They also reported how the accuracy of sound stream separation measured by automatic speech recognition is improved by **adding more modalities**, from monaural input, binaural input, and binaural input with visual information.

Some critical problems with Bi-HBSS and their work for real-world applications are summarized as follows:

1. **HRTF is needed for identifying the direction**. It is time-consuming to measure an HRTF, and it is usually measured in an anechoic room. Since it depends on auditory environments, re-measurement or adaptation is needed to apply it to other environments.
2. **HRTF is needed for creating a direction-pass filter**. Their direction-pass filter needs HRTFs to compose. Since an HRTF is usually measured in *discrete* azimuth and elevation, it is difficult to implement sound tracking for continuous movement of sound sources.

Therefore, a new method without using HRTF should be invented for localization (sound source direction) and direction (by using a direction-pass filter). We will propose a new auditory localization based on the epipolar geometry.

2.3 Sound Source Localization

Some robots developed so far had a capability of sound source localization. Huang *et al.* developed a robot that had three microphones [10]. Three microphones are installed vertically on the top of the robot, composing a triangle.

Comparing the input power of microphones, two microphones that have more power than the other are selected and the sound source direction is calculated. By selecting two microphones from three, they solved the problem that two microphones cannot determine the place of sound source in front or behind. By identifying the direction of sound source from a mixture of an original sound and its echoes, the robot turns the body towards the sound source. The idea using three microphones for sound source localization, however, means this robot is based on "*stop-perceive-act*" strategy. If the robot behavior is based on the concepts of active audition, it needs only two microphones for sound source localization. The combination of two microphones and active head movement can solve the front-behind problem because active head movements generate motion disparity in audition, which is corresponding to a parallax. For tracking a moving sound source, robots based on "*stop-perceive-act*" strategy do not track such a moving sound source properly while active audition based robots can do it easily because auditory processing with motion is necessary.

Humanoids of Waseda University can localize a sound source by using two microphones [14, 26]. These humanoids localize a sound source by calculating IID or IPD with HRTF. These robot can neither separate even a sound stream nor localize more than one sound source. The Cog humanoid of MIT has a pair of omni-directional microphones embedded in simplified pinnae [3, 11]. In the Cog, auditory localization is trained by visual information. This approach does not use HRTF, but assumes a single sound source. To summarize, both approaches lack for the CASA viewpoints.

2.4 Active Audition

A humanoid should be active in the sense that it tries to do some activity to improve perceptual processing. Such activity includes to change the position of cameras and microphones by motor control.

When a humanoid hears sound by facing the sound source in the center of the pair of microphones, ITD and IID is almost zero if the pair of microphones are correctly calibrated. In addition, sound intensity of both channels becomes stronger, because the ear cover makes a omni-directional microphone directional. Given the multiple sound sources in the auditory scene, a humanoid actively moves its head to improve localization by aligning microphones orthogonal to the sound source and by capturing the possible sound sources by vision.

However, a new problem occurs because gaze stabilization is attained by visual servo or auditory servo. Sounds are generated by motor rotation, gears, belts and ball bearings. Since these internal sound sources are much closer than other external sources, even if the absolute power of sounds is much lower, input sounds are strongly influenced. This is also the case for the SONY AIBO entertainment robot; AIBO is equipped with a microphone, but internal noise mainly caused by a cooling fan is too large to utilize sounds unless a speaker is located closely to it.

2.5 Internal Sound Cancellation

Since active perception causes sounds by the movement of various movable parts, internal sound cancellation is critical to enhance external sounds (see Fig. 2). A cover of humanoid body reduces sounds of motors emitted to the external world by separating internal and external world of the humanoid. Such a cover is, thus expected to reduce the complexity of sound processing caused by motor sounds. Since most robots developed so far do not have a cover, auditory processing cannot become first-class perception of a humanoid.

Internal sound cancellation may be attained by one or a combination of the following methodologies:

1. noise cancellation,
2. independent component analysis (ICA),
3. case-based cancellation,
4. model-based cancellation, and
5. learning and adaptation.

To record sounds for case-based and model-based cancellation, each sound should be labeled appropriately. We use data consisting of time and motor control commands as label for sound. In the next section, we will explain how these methods are utilized in our active audition system.

3 Active Audition System

An active audition system consists of two components; internal sound cancellation, and sound stream separation.

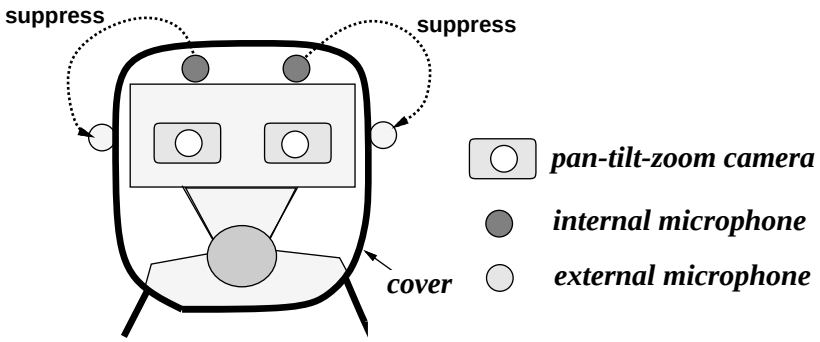


Fig. 2. Internal and external microphones for internal sound cancellation

3.1 Internal Sound Canceling System

Internal sounds of *SIG* are caused mainly by the followings:

- Camera motors — sounds of movement are quiet enough to ignore, but sounds of standby is loud (about 3.7 dB).
- Body motors — sounds of standby and movement are loud (about 5.6 dB and 23 dB, respectively).

Comparison of noise cancellation by adaptive filtering, ICA, case-based cancellation and model-based cancellation, we concluded that only adaptive filters work well. Four microphones are not enough for ICA to separate internal sounds. It is difficult to construct case-based or model-based for cancellation because the same movement generates a lot of different sounds. And even if successfully constructed, case-based and model-based cancellation would affect the phase of original inputs, which causes errors of IPD.

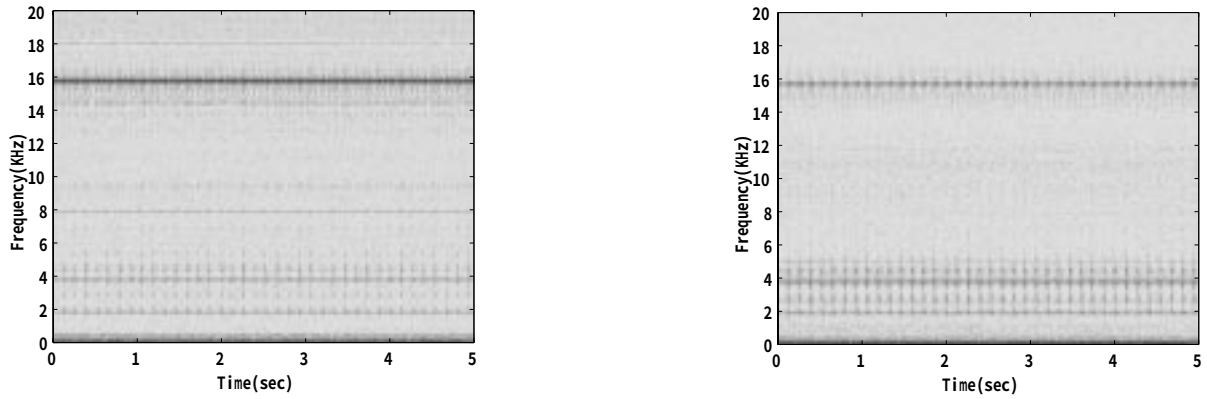
Figs. 4a and 4b show spectrograms of sound captured by humanoid internal and external microphones, respectively. Both figures have peaks in frequencies of 500 and 600Hz which are originating from external sound sources. The camera motor sound on standby has time continuous peaks at each frequency of around 2KHz, 4KHz and 16KHz on the spectrogram as shown in Fig. 3 and the power of the body motor sound on standby is concentrated only at low frequency. On the other hand, the body motor makes stronger noises in motion than the camera and motor sounds on standby. The noises are observed as burst noises between 2 and 4 seconds in each figure, and spread over broad bands. This means the body motor noises in motion affect the system worse.

Then, we designed an adaptive filter, which is often used as a filter for *active noise control* [22, 8], by a FIR (Finite Impulse Response) digital filter of order 100, because FIR filter is a linear phase filter. This property is essential to localize the sound source by IID/IAD or ITD/IPD. The coefficients of the FIR adaptive filter is calculated by least-mean-square (LMS) algorithm. It was expected that the adaptive filter would cancel such burst noises mainly in motion and emphasize sounds of 500 and 600Hz, which are originating from outer sources. Fig. 5 show the result of noise cancellation by the adaptive filter. Indeed, camera noises with frequency around 16KHz are canceled well, but Fig. 5 shows as follows:

- The body motor noises in motion, which we expected to cancel, are still remained.
- Sounds of 500 and 600Hz from external sources, which we did not want to cancel, are also canceled with other noises.

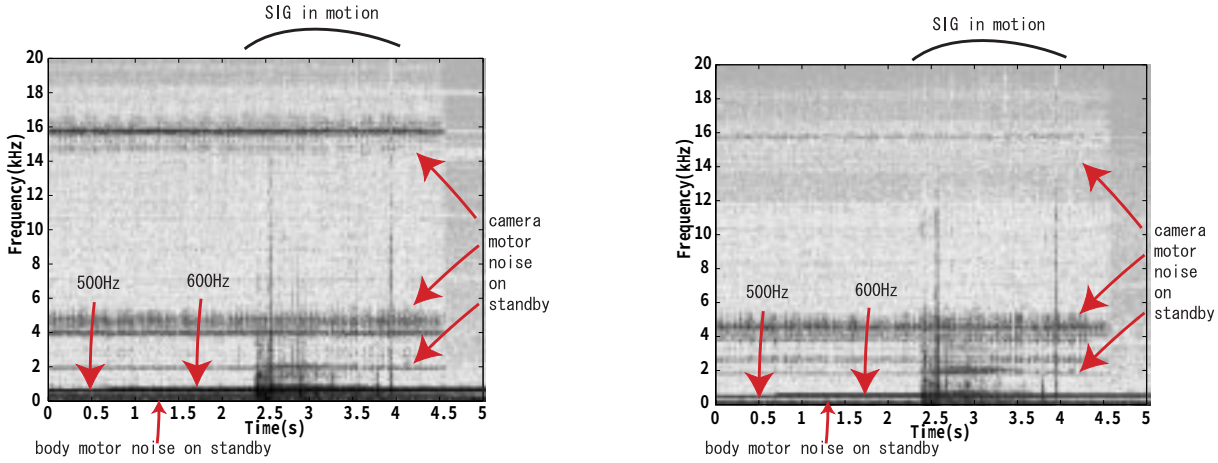
Thus, it is difficult to cancel internal motor noises by the adaptive filter. This insufficient cancellation makes poor localization compared to results of localization without internal sound cancellation. The reasons why internal motor noise cancellation fails are summarized below.

1. Adaptive filter is unsuitable for unexpected abrupt noises such as a burst noise because of generating an estimated value from past values of the number of the filter order.
2. The system can not assume that internal microphones capture only motor noises while the assumption is a premise in most applications of active noise control, because external sounds leak into inside cover through gaps and joints between cover components.
3. Some errors occur in noise estimation by the adaptive filter. These errors deteriorate the noise cancellation.



a) internal microphone sound b) external microphone sound

Fig. 3. Spectrogram of camera noise



a) internal microphone sound b) external microphone sound

Fig. 4. Spectrogram of input sounds

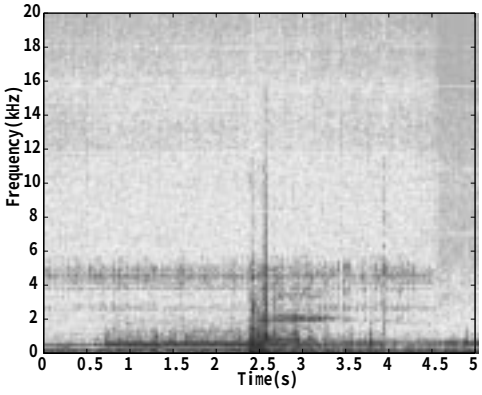


Fig. 5. Spectrogrem after the adaptive filter processing

Instead, our adaptive filter uses *heuristics with internal microphones*, which specifies the condition to cut off burst noises mainly caused by motors. For example, sounds at stoppers, by friction between cable and body, creaks at joints of cover parts may occur. The heuristics orders that localization by sound or direction-pass filter ignore a subband if all the following conditions hold:

1. The power of internal sounds is much stronger than that of external sounds.
2. Twenty adjacent subbands have strong power (30 dB).
3. A motor motion is being processed.

3.2 Sound Stream Separation by Localization

We design a new direction-pass filter with a direction which is calculated by epipolar geometry.

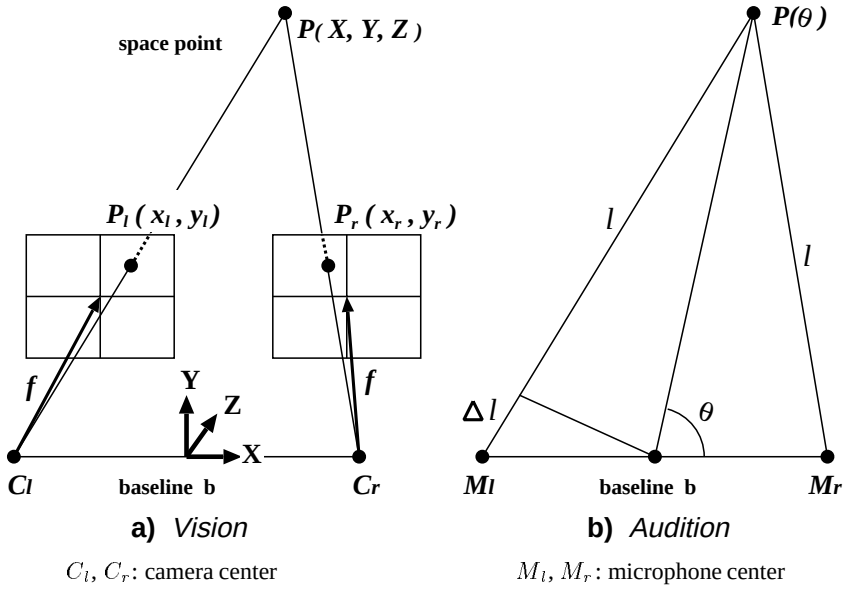


Fig. 6. Epipolar geometry for localization

Localization by Vision using Epipolar Geometry Consider a simple stereo camera setting where two cameras have the same focal length, their light axes are in parallel, and their image planes are on the same plane (see Fig. 6a). We define the world coordinate (X, Y, Z) and each local coordinate. Suppose that a space point $P(X, Y, Z)$ is projected on each camera's image plane, (x_l, y_l) and (x_r, y_r). The following relations hold [9]:

$$X = \frac{b(x_l + x_r)}{2d}, Y = \frac{b(y_l + y_r)}{2d}, Z = \frac{bf}{d}$$

where f is the focal length of each camera's lens and b is the baseline. Disparity d is defined as $d = x_l - x_r$.

The current implementation of common matching in *SIG* is performed by using corner detection algorithm [13]. It extracts a set of corners and edges then constructs a pair of graphs. A matching algorithm is used to find corresponding left and right image to obtain depth.

Since the relation $y_l = y_r$ also holds under the above setting, a pair of matching points in each image plane can be easily sought. However, for general setting of camera positions, matching is much more difficult and time-consuming. Usually, a matching point in the other image plane exists on the epipolar line which is a bisecting line made by the epipolar plane and the image plane.

Localization by Audition using Epipolar Geometry Auditory system extracts the direction by using epipolar geometry. First, it extract peaks by using FFT (Fast Fourier Transformation) for each subband, 47Hz in our implementation, and then calculates the IPD.

Let $Sp^{(r)}$ and $Sp^{(l)}$ be the right and left channel spectrum obtained by FFT at the same time tick. Then, the IPD $\triangle\varphi$ is calculated as follows:

$$\triangle\varphi = \tan^{-1}\left(\frac{\Im[Sp^{(r)}(f_p)]}{\Re[Sp^{(r)}(f_p)]}\right) - \tan^{-1}\left(\frac{\Im[Sp^{(l)}(f_p)]}{\Re[Sp^{(l)}(f_p)]}\right)$$

where f_p is a peak frequency on the spectrum, $\Re[Sp]$ and $\Im[Sp]$ are the real and imaginary part of the spectrum Sp. The angle θ is calculated by the following equation:

$$\cos\theta = \frac{v}{2\pi f_p b}\triangle\varphi$$

where v is the velocity of sound. For the moment, the velocity of sound is fixed to 340m/sec and remains the same even if the temperature changes.

This peak extraction method works at 48 KHz sampling rate and calculates FFT for 1,024 points, but runs much faster than Bi-HBSS (12 KHz sampling rate with HRTF) and extracted peaks are more accurate [17].

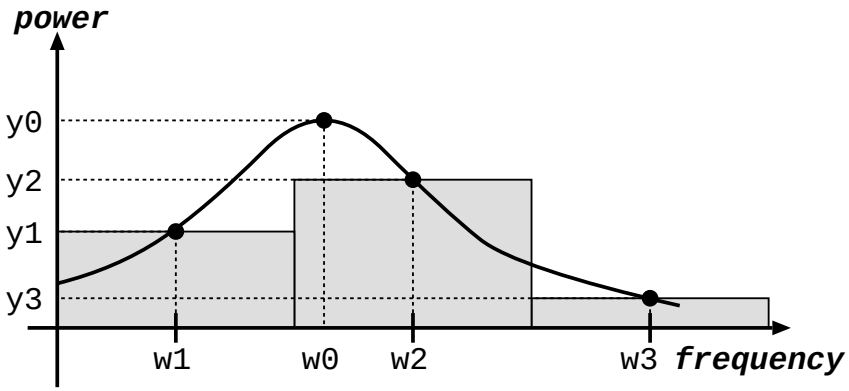


Fig. 7. A spectral peak by Fourier Transformation

Pitch Extraction Pitches are extracted by a kind of spectral subtraction [2]. It uses peak approximation method based on characteristics of FFT and window function. Consider that the peak $[\omega_2, y_2]$ is detected, and the values of both neighbors are $[\omega_1, y_1]$ and $[\omega_3, y_3]$ as shown in Fig. 7. Then, the true peak $[\omega_0, y_0]$ is estimated as follows:

$$\omega_0 = \begin{cases} \omega_2 + \dfrac{2\pi\,(2|y_1| - |y_2|)}{T\,(|y_1| + |y_2|)} & (\omega_1 < \omega_0 \le \omega_2) \\ \omega_2 - \dfrac{2\pi\,(-|y_2| + 2|y_3|)}{T\,(|y_2| + |y_3|)} & (\omega_2 < \omega_0 < \omega_3) \end{cases} \tag{1}$$

$$\begin{aligned} Arg(y_0) &= \tan^{-1}\left(\frac{\Im[y_0]}{\Re[y_0]}\right) \\ &= \tan^{-1}\left(\frac{\Im[y_2]}{\Re[y_2]}\right) + \frac{T}{2}(\omega_2 - \omega_0) \end{aligned} \tag{2}$$

$$\begin{aligned} |y_0| &= \frac{\Delta\omega\left(-T^2\Delta\omega^2 + 4\pi^2\right)}{2\pi^2 \sin\frac{T}{2}\Delta\omega}|y_2|\,, \\ \Delta\omega &= \omega_2 - \omega_0 \end{aligned} \tag{3}$$

ω_0 is estimated as Equation (1). And the phase and amplitude of the true peak y_0 are estimated as Equations (2) and (3), respectively.

Because the above equations require relatively the small number of calculation, our method can run faster. And it can extract more accurate pitches because the approximation is suitable to the frequency analysis methods, that is, the FFT and hanning window. For example, in comparison with Bi-HBSS [21], which is known as a sound source separation system using a pitch extraction method by spectral subtraction based on harmonic structures, our method needs only $1/200$ of amount of calculation per a peak [17].

New Direction-Pass Filter using Epipolar Geometry As mentioned earlier, HRTFs are usually not available in real-world environments, because it changes when a new furniture is installed, a new object comes in the room, or humidity of the room changes. In addition, HRTFs should be interpolated for auditory localization of a moving sound source, because HRTFs are measured for discrete positions. Therefore, a new method must be invented. Our method is based on the direction-pass filter with epipolar geometry.

As opposed to localization by audition, the direction-pass filter selects subbands that satisfies the IPD of the specified direction. The detailed algorithm is describes as follows:

1. The specified direction θ is converted to $\Delta\varphi$ for each subband (47 Hz).
2. Extract peaks and calculated IPD, $\Delta\varphi'$.
3. If IPD satisfies the specified condition, namely, $\Delta\varphi' = \Delta\varphi$, then collect the subband.
4. Construct a wave consisting of collected subbands.

By using the relative position between camera centers and microphones, it is easy to convert from epipolar plane of vision to that of audition (see Fig. 6b). In *SIG*, the baselines for vision and audition are in parallel.

Therefore, whenever a sound source is localized by epipolar geometry in vision, it can be converted easily into the angle θ as described in the following equation:

$$\cos\theta = \frac{\boldsymbol{P}\cdot\boldsymbol{M_r}}{|\boldsymbol{P}||\boldsymbol{M_r}|} = \frac{\boldsymbol{P}\cdot\boldsymbol{C_r}}{|\boldsymbol{P}||\boldsymbol{C_r}|}.$$

Localization by Servo-Motor System The head direction is obtained from potentiometers in the servo-motor system. Hereafter, it is referred as *the head direction by motor control*. Head direction by potentiometers is quite accurate by the servo-motor control mechanism. If only the horizontal rotation motor is used, horizontal direction of the head is obtained accurately, about $\pm 1^\circ$. By combining visual localization and the head direction, *SIG* can determine the position in world coordinates.

Accuracy of Localization Accuracy of extracted directions by three sensors: vision, audition, and motor control is measured. The results for the current implementation are $\pm 1^\circ$, $\pm 10^\circ$, $\pm 15^\circ$, for vision, motor control, and audition, respectively.

Therefore, the precedence of information fusion on direction is determined as below:

$$\text{vision} > \text{motor control} > \text{audition}$$

Sensor Integrated System The system contains a perception system that integrates sound, vision, and motor control (Fig. 8). The association module maintains the consistency between information extracted by image processing, sound processing and motor control subsystems. For the moment, association includes the correspondence between images and sounds for a sound source; loud speakers are the only sound sources, which can generate sound of any frequency. Focus of attention and action selection modules are described in [13].

4 Experiment — Motion Tracking by Three Kinds of Sensors

In this section, we will demonstrate how vision, audition and head direction by potentiometers compensate each missing information to localize sound sources while *SIG* rotates to see an unknown object.

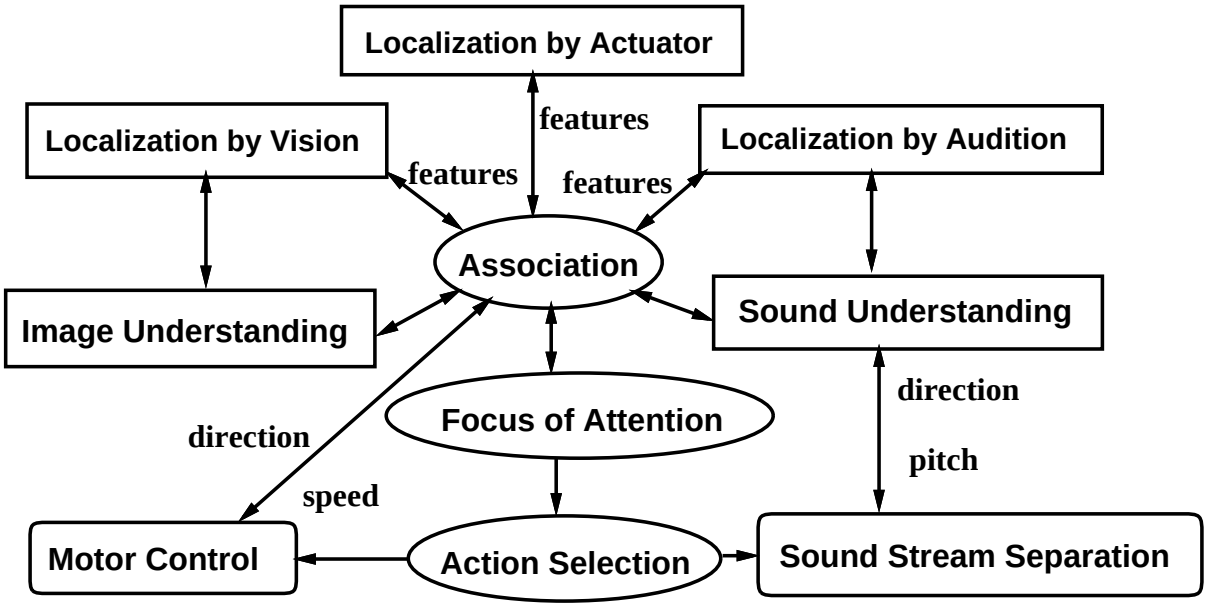


Fig. 8. Integrated humanoid perception system

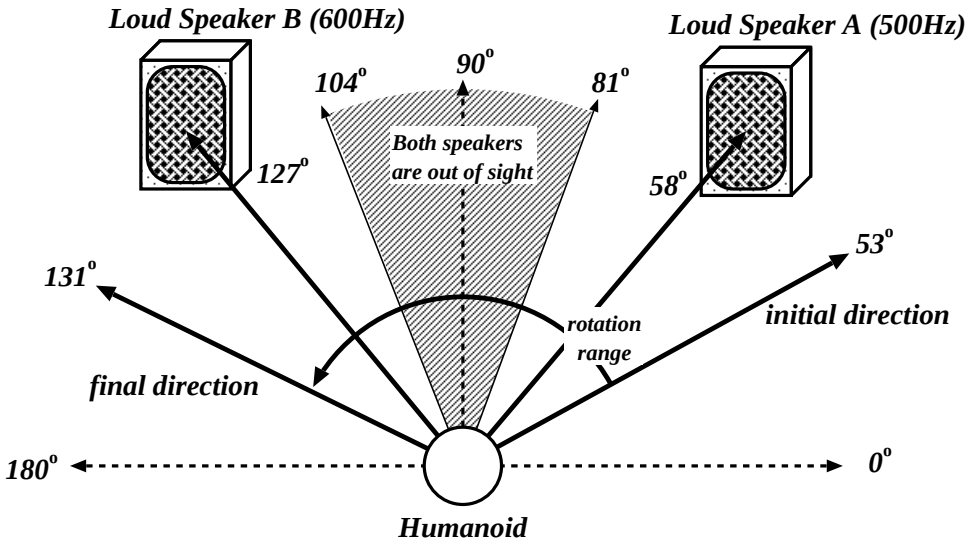
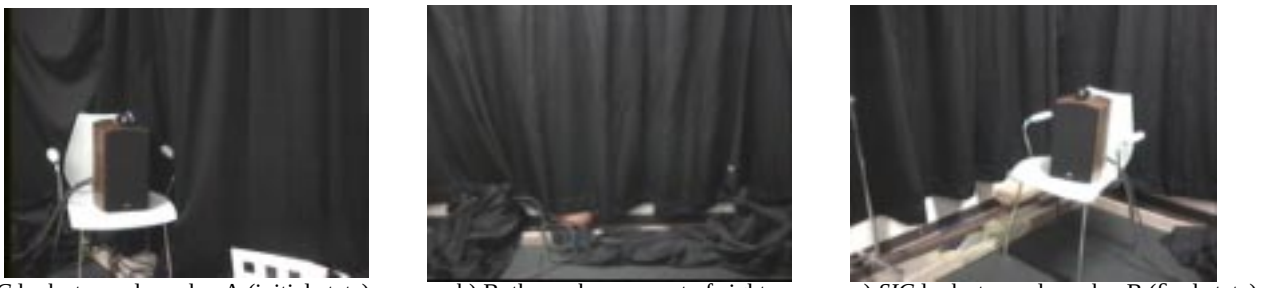
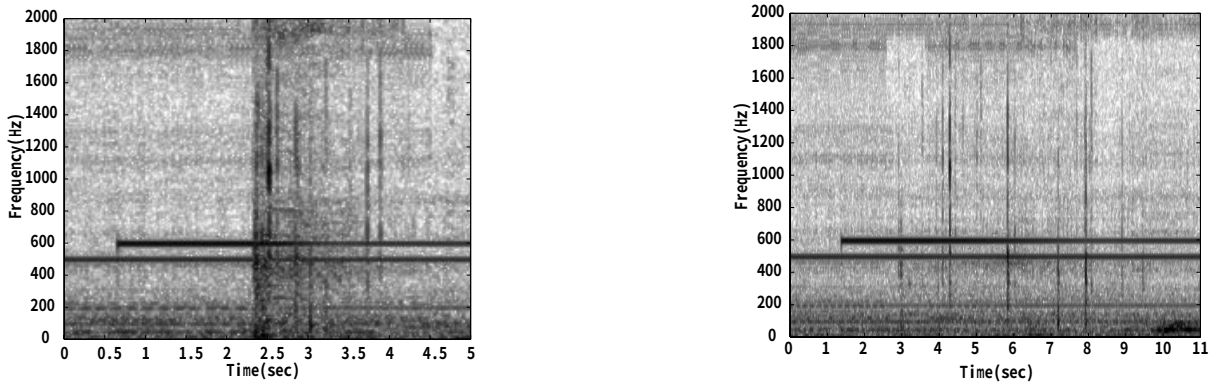


Fig. 9. Experiment: Motion tracking by vision and audition while *SIG* moves.



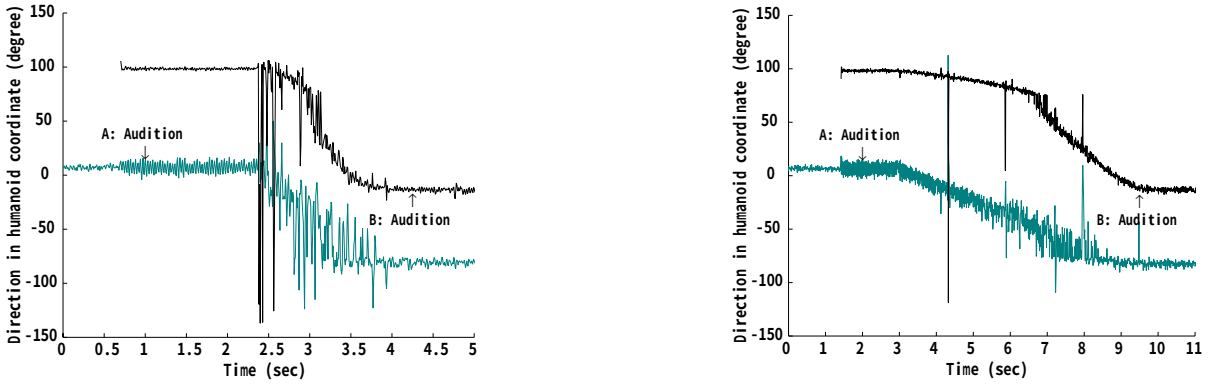
a) *SIG* looks toward speaker A (initial state). b) Both speakers are out of sight c) *SIG* looks toward speaker B (final state).

Fig. 10. Tracking by moving head



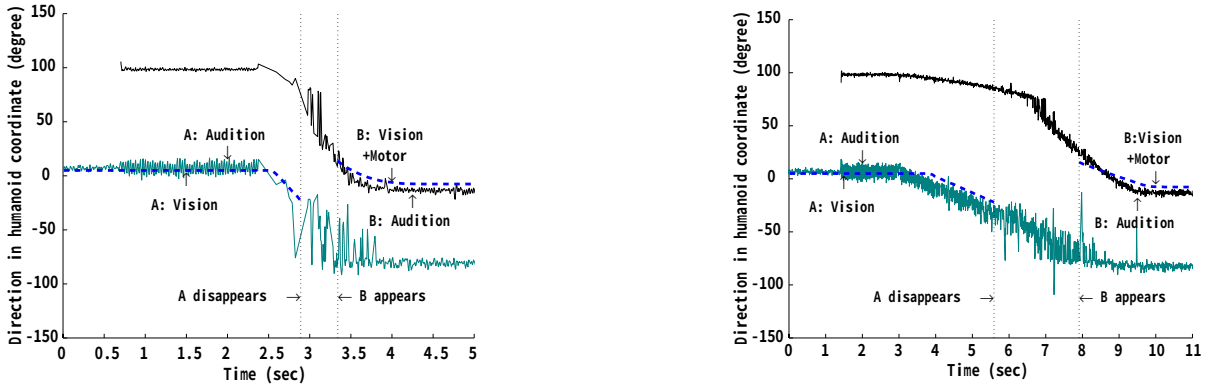
a) fast movement of *SIG* b) slow movement of *SIG*

Fig. 11. Spectrogram of input sounds



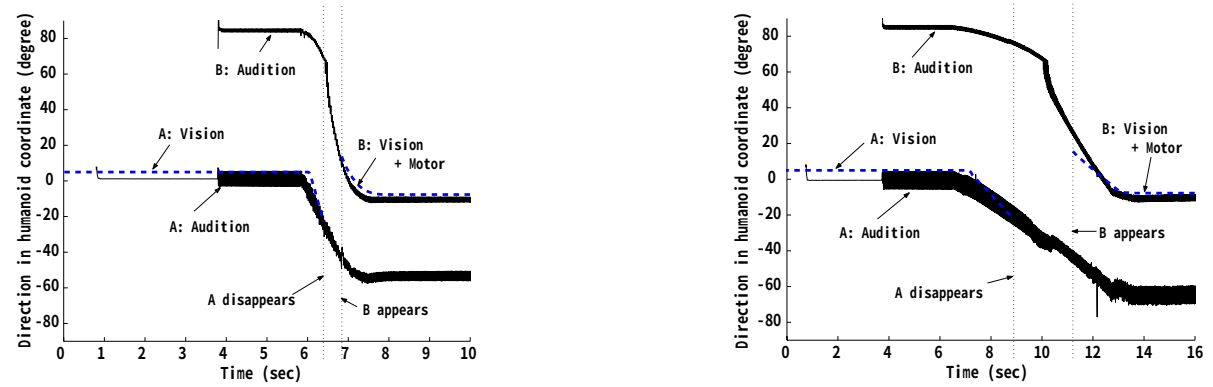
a) fast movement of *SIG* b) slow movement of *SIG*

Fig. 12. Localization without heuristics of cancellation



a) fast movement of *SIG* b) slow movement of *SIG*

Fig. 13. Localization by vision and audition



a) fast movement of *SIG* b) slow movement of *SIG*

Fig. 14. Localization for strong signal

Scenario: There are two sound sources: two B&W Noutilus 805 loud speakers located in a room of 10 square meters. The room where the system is installed is a conventional residential apartment facing a road with busy traffic, and exposed to various daily life noises. The sound environment is not at all controlled for experiment to ensure feasibility of the approach in daily life.

One sound source A (Speaker A) plays a monotone sound of 500 Hz. The other sound source B (Speaker B) plays a monotone sound of 600 Hz. A is located in front of *SIG* (5° left of the initial head direction) and B is located 69° to the left. The distances from *SIG* to both sound sources are about 210cm. Since the visual field of camera is only 45° in horizontal angle, *SIG* cannot see B at the initial head direction, because B is located at 70° left to the head direction, thus it is outside of the visual fields of the cameras. Fig. 9 shows this situation.

1. A plays a sound at 5° left of the initial head direction.
2. *SIG* associates the visual object with the sound, because their extracted directions are the same.
3. Then, B plays a sound about 3 seconds later. At this moment, B is outside of the visual field of the *SIG*. Since the direction of the sound source can be extracted only by audition, *SIG* cannot associate anything to the sound.
4. *SIG* turns toward the direction of the unseen sound source B using the direction obtained by audition.
5. *SIG* finds a new object B, and associates the visual object with the sound.

Four kinds of benchmark sounds are examined; fast (68.8 degree/sec) and slow (14.9 degree/sec) *SIG* movements ,which need 0.14 and 0.18 seconds to reach stable velocity from stationary state respectively. Weak signals (similar power to internal standby sounds, which makes signal to noise ratio 0dB) and strong signals (about 50 dB). Spectrogram of each input is shown in Fig. 11. Motion tracking by vision and audition, and motion information are evaluated.

Results: Results of the experiment were very promising. First, accurate sound source localization was accomplished without using the HRTF under the real world environment.

Errors in localization before and after the rotation are shown in Table 1. The errors can be estimated using the direction information described in Fig. 9. The errors are about 13.7 $^\circ$ on average, and at most 28.4 $^\circ$. The errors are acceptable; The reasons are as follows:

– The experiment is done not in a simulated environment, but in the residential room where HRTF measured in an anechoic room is of little use.
– The error of Bi-HBSS in a simulated environment is $\pm 10^\circ$. That of our localization method in the residential room is 13.7° on average. The accuracy of our localization method is compared with that in Bi-HBSS.
– The accuracy in localization of sound sources from the front of *SIG*(around 0°) is equal or superior to that in Bi-HBSS as shown in average errors of “speaker A(initial state)” and “speaker B(final state)” in Table 1. This means that our method is more sensitive to sources in front, because such a sound is less distorted by the head. Therefore, our system refine aligning microphones orthogonal to the sound source.

Although, in motion, motor noises disturb accurate localization, the use of epipolar geometry for audition was proven very effective. The velocity of rotation can be regarded as almost constant because it needs at most 0.17 seconds to reach stable velocity from stationary state. Hence, the direction in humanoid coordinate should change linearly in motion. In Figs. 12 and 14, time series data for estimated sound source direction using epipolar based non-HRTF method is plotted with an ego-centric polar coordinate where 0° is the direction dead front of the head, minus is right of the head direction. Actually, the results of localization with such linearity in motion are shown.

The effect of adaptive noise canceling is clearly shown. Fig. 12 shows estimated sound source directions without motor noise cancellation. Sound direction estimation is seriously hampered when the head is moving (around time 5 - 6 seconds). The spectrogram (Fig. 11) clearly indicate such motor noises. When *SIG* is constantly moving to track moving sound sources or to move itself for a certain position, the robot continues to generate such a noise that makes audition almost impossible to use for perception.

The effects of internal sound cancellation by heuristics are shown in Figs. 13, and 14. The time series of estimated sound source directions for weak and strong signals were localized by vision and audition. Dotted lines indicate the localization results by vision / vision and motor control. The localization information is more accurate than that obtained by audition because the error in vision is within $\pm 1^\circ$. The information by vision can be used for auditory processing such as direction-pass filters to improve sound source separation. However, *SIG* cannot attain such accurate localization information in motion (indicated as between vertical dotted lines) because both speakers are out of sight and any clue for localization by vision would not be found. In such cases, localization information by audition can compensate for missing information caused by narrow visual fields and occluded sound sources.

Furthermore, such accurate localization by audition enables association between audition and vision. While *SIG* is moving, sound source B comes into its visual field. The association module checks the consistency of localization by vision and audition. If the discovered loud speaker does not play sounds, inconsistency occurs and the visual system would resume its search finding an object producing sound. If association succeeds, B's position in world coordinates is calculated by using motor information and the position in humanoid coordinates obtained by vision.

Experimental results show that position estimation by audition and vision can create consistent association even under the condition that the robot is moving with generating motor noises. It should be noted that sound source localization by audition in the experiment uses epipolar geometry for audition, and do not use HRTF. Thus, we can simplified the robot in unknown acoustic environment and localize sound sources.

	Direction	Fig. 12 a)		Fig. 12 b)		Fig. 14 a)		Fig. 14 b)		Average
	obtained by Fig.9	Result	Error	Result	Error	Result	Error	Result	Error	error
Speaker A (initial state)	74.0	98.0±3.4	*28.4*	98.0±3.4	*28.4*	84.5±1.5	*12.0*	84.5±1.5	*12.0*	*20.2*
Speaker B (initial state)	5.0	8.5±1.7	*5.2*	6.7±9.8	*11.5*	1.0±6.6	*10.6*	0.0±6.0	*11.0*	*9.6*
Speaker A (final state)	-4.0	-12.9±6.0	*14.9*	-12.7±2.5	*11.2*	-10.2±1.5	*7.7*	-10.2±1.5	*7.7*	*10.4*
Speaker B (final state)	-73.0	-80.0±4.2	*11.2*	-82.0±2.5	*11.5*	-53.0±2.5	*22.5*	-63.7±5.0	*14.3*	*14.9*
Average error	—	—	*14.9*	—	*15.7*	—	*13.2*	—	*11.3*	*13.7*

Table 1. Errors in localization (degree)

5 Discussion and Future Work

1. The experiment demonstrates the feasibility of the proposed humanoid audition in real-world environments. Since there are a lot of non-desired sounds, caused by traffic, people outside the test-room, and of course internal sounds, the CASA assumption that input sounds consist of a mixture of sounds is essential in real-world environments. Similar work by [19] was done in a simulated acoustic environment, but it may fail in localization and sound stream separation in real-world environments. Most robots capable of auditory localization developed so far assume a single sound source.
2. Epipolar geometry gives a way to unify visual and auditory processing, in particular localization and sound stream separation. This approach can dispense with HRTFs. As far as we know, no other systems can do it. Most robots capable of auditory localization developed so far use HRTFs explicitly or implicitly, and may fail in identifying some spatial directions or tracking moving sound sources.
3. The cover of the humanoid is very important to separate its internal and external worlds. However, we've realized that resonance within a cover is not negligible. Therefore, its inside material design is important.
4. Social interaction realized by utilizing body movements extensively makes auditory processing more difficult. The Cog Project focuses on social interaction, but this influence on auditory processing has not been mentioned [4]. A cover of the humanoid will play an important role in reducing sounds caused by motor movements emitted toward outside the body as well as in giving a friendly outlook to human.

Future Work Active perception needs self recognition. The problem of acquiring the concept of self recognition in robotics has been pointed out by many people. For audition, handling of internal sounds made by itself is a research area of modeling of self. Other future work includes more tests for feasibility and robustness, real-time processing of vision and auditory processing, internal sound cancellation by independent component analysis, addition of more sensor information, and applications.

6 Conclusions

In this paper, we present active audition for humanoid which includes internal sound cancellation, a new method for auditory localization, and a new method for separating sound sources from a mixture of sounds. The key idea is to use epipolar geometry to calculate the sound source direction and to integrate vision and audition in localization and sound stream separation. This method does not use HRTF (Head-Related Transfer Function) which is a main

obstacle in applying auditory processing to real-world environments. We demonstrate the feasibility of motion tracking by integrating vision, audition and motion information. The important research topic now is to explore possible interaction of multiple sensory inputs which affects quality (accuracy, computational costs, etc) of the process, and to identify fundamental principles for intelligence.

Acknowledgments

We thank our colleagues of Symbiotic Intelligence Group, Kitano Symbiotic Systems Project; Yukiko Nakagawa, Takahiro Miyashita, Dr. Iris Fermin, and Dr. Theo Sabish for their discussion. We thank Prof. Hiroshi Ishiguruo of Wakayama University for his help in active vision and integration of visual and auditory processing.

References

1. Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1987.
2. S. F. Boll. A spectral subtraction algorithm for suppression of acoustic noise in speech. In *Proceedings of 1979 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79)*, pages 200–203. IEEE, 1979.
3. R. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, and M. Williamson. The cog project: Building a humanoid robot. Technical report, MIT, 1999.
4. R. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, and M. Williamson. The cog project: Building a humanoid robot. In *Lecture Notes in Computer Science, LNAI 1562*, pages 52–87. Spriver-Verlag, 1999.
5. G. J. Brown. *Computational auditory scene analysis: A representational approach*. University of Sheffield, 1992.
6. J. Cavaco, S. ad Hallam. A biologically plausible acoustic azimuth estimation system. In *Proceedings of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99)*, pages 78–87. IJCAI, 1999.
7. M. P. Cooke, G. J. Brown, M. Crawford, and P. Green. Computational auditory scene analysis: Listening to several things at once. *Endeavour*, 17(4):186–190, 1993.
8. S.J. Elliott *et al*. A multiple error lms algorithm and application to the active noise control of sound and vibration. *IEEE Trans. Acoust.*, ASSP-15(10):1423–1434, 1987.
9. O. D. Faugeras. *Three Dimensional Computer Vision: A Geometric Viewpoint*. The MIT Press, MA., 1993.
10. J. Huang, N. Ohnishi, and N. Sugie. Separation of multiple sound sources by using directional information of sound source. *Artificial Life and Robotics*, 1(4):157–163, 1997.
11. R. E. Irie. Multimodal sensory integration for localization in a humanoid robot. In *Proceedings of the Second IJCAI Workshop on Computational Auditory Scene Analysis (CASA'97)*, pages 54–58. IJCAI, 1997.
12. H. Kitano, H. G. Okuno, K. Nakadai, I. Fermin, T. Sabish, Y. Nakagawa, and T. Matsui. Designing a humanoid head for robocup challenge. In *Proceedings of Agent 2000 (Agent 2000)*, page *to appear*, 2000.
13. T. Lourens, K. Nakadai, H. G. Okuno, and H. Kitano. Selective attention by integration of vision and audition. In *Humanoids2000*, in this volume.
14. Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi. Multi-person conversation via multi-modal interface — a robot who communicates with multi-user. In *Proceedings of Eurospeech*, pages 1723–1726. ESCA, 1999.
15. 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)*, pages 832–839. AAAI, 2000.
16. 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. (*accepted*).
17. K. Nakadai, H. G. Okuno, and H. Kitano. A method of peak extraction and its evaluation for humanoid. In *SIG-Challenge-99-7*, pages 53–60. JSAI, 1999.
18. K. Nakadai, H. G. Okuno, and H. Kitano. Humanoid active audition system improved by the cover acoustics. In *Proceedings of 6th Pacific Rim International Conferences on Artificial Intelligence(PRICAI 2000)*, 2000. (*accepted*).
19. 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)*, pages 768–775. AAAI, 1999.
20. T. Nakatani, H. G. Okuno, and T. Kawabata. Auditory stream segregation in auditory scene analysis with a multi-agent system. In *Proceedings of 12th National Conference on Artificial Intelligence (AAAI-94)*, pages 100–107. AAAI, 1994.
21. T. Nakatani, H. G. Okuno, and T. Kawabata. Residue-driven architecture for computational auditory scene analysis. In *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, volume 1, pages 165–172. AAAI, 1995.
22. P.A. Nelson and S.J. Elliott. *Active Contorol of Sound*. ACADEMIC PREESS, London, 1992.
23. H. G. Okuno, T. Nakatani, and T. Kawabata. Listening to two simultaneous speeches. *Speech Communication*, 27(3-4):281–298, 1999.
24. D. Rosenthal and H. G. Okuno, editors. *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

25. M Slaney, D. Naar, and R. F. Lyon. Auditory model inversion for sound separation. In *Proceedings of 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 77–80, 1994.
26. A. Takanishi, S. Masukawa, Y. Mori, and T. Ogawa. Development of an anthropomorphic auditory robot that localizes a sound direction (*in japanese*). *Bulletin of the Centre for Informatics*, 20:24–32, 1995.