Sound and Visual Tracking for Humanoid Robot

Hiroshi G. Okuno[1,2], Kazuhiro Nakadai[1], Tino Lourens[1], and Hiroaki Kitano[1,3]
[1] Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp., Tokyo, Japan
[2] Department of Information Sciences, Science University of Tokyo, Chiba, Japan
[3] Sony Computer Science Laboratories, Inc., Tokyo, Japan
E-mail: okuno@nue.org, {nakadai, tino}@symbio.jst.go.jp, kitano@csl.sony.co.jp
Keywords: Robotics, Sensor Fusion, Active Audition

Abstract— **Mobile robots with auditory perception usually adopt "*stop-perceive-act*" principle to avoid sounds made during moving due to motor noises or a bumpy roads. Although this principle reduces the complexity of the problems involved auditory processing for mobile robots, it restricts their capabilities of auditory processing. In this paper, sound and visual tracking is investigated to compensate each drawbacks in tracking objects and to attain robust object tracking. Visual tracking may be difficult in case of occlusion, while sound tracking may be ambiguous in localization due to the nature of auditory processing. For this purpose, we present an active audition system for humanoid robot. The audition system of the highly intelligent 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 improved sound source tracking by integrating audition, vision, and motor movements. Given the multiple sound sources in the auditory scene, *SIG the 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. The system adaptively cancels motor noise using motor control signals. The experimental result demonstrates the effectiveness of sound and visual tracking.**

I. Introduction

Mobile robots with auditory perception usually adopt "*stop-perceive-act*" principle to avoid sounds made during moving due to motor noises or a bumpy road. Although this principle reduces the complexity of the problems involved auditory processing for mobile robots, it restricts their capabilities of auditory processing. In this paper, sound and visual tracking is investigated to compensate each drawbacks in tracking objects and to attain robust object tracking. Visual tracking may be difficult in case of occlusion, while sound tracking may be ambiguous in localization due to the nature of auditory processing.

The goal of the research reported in this paper is to establish a technique of multi-modal integration for improving perception capabilities. We use an upper-torso humanoid robot 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. 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 signifies coupling 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 coupling 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, changing focus, zooming in or out, changing camera resolution, widening or narrowing iris and so on. Therefore, active vision system is always coupled with servo-motor system, which means that active vision system is in general associated with motor noise.

The concept of active perception can be extended to audition, too. Human audition is 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 [2], [3], [4], [5]. Its goal is to understand an arbitrary sound mixture including speech, non-speech sounds, and music 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 [6].

A. Audition for Humanoid in Daily Environments

For audition, the following issues should be resolved in order to deploy our robot in daily environments:

1. Ability to localize sound sources in unknown acoustic environment.
2. Ability to move its body to obtain further information from audition, vision, and other perceptions.
3. Ability to continuously perform auditory scene analysis under noisy environment, where noise comes from both environment and motor noise 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

Head-Related Transfer Function (*HRTF*) was measured in the specific room environment, and measurement has to be repeated 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 or at least less dependent on HRTF is essential for practical systems. The system reported in this paper implements epipolar geometry-based sound source localization to eliminate the need for HRTF. The use of epipolar geometry for audition is advantageous when combined with the vision system because many vision systems uses epipolar geometry for visual object localization.

Second, active audition that couples 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:

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

Therefore, *gaze stabilization* for microphones by sound tracking 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 create considerable noise. In a complex robotic system, such as humanoid, motor noise is complex and often irregular because numbers 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 noise. To avoid using such an implementation, we implemented an adaptive noise canceling scheme that uses motor control signal to anticipate and cancel 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 suppression algorithms. 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 the problems of active perception and proposes new sound source separation. Last two sections give discussion and conclusion.

II. 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 [7], [8], [9]. 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 includes understanding general sounds, sensor fusion, active audition, and internal sound suppression.

A. 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 [2], [10]. 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 [2].

Nakatani et al [4] used harmonic structures as the clue of separation and developed a monaural-based harmonic stream separation system, called HBSS. 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 [11]. Bi-HBSS uses harmonic structures and the direction of sound sources as clues of separation. Okuno et al. [12] extended Bi-HBSS to separate speech streams, and uses the resulting system as a front end for automatic speech recognition.

B. 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 [13]. 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 *interaural time/phase difference* (*ITD* or *IPD*), and/or the *interaural intensity/amplitude difference* (*IID* or *IAD*) 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. $\pm 8^\circ$ [14]. However, this is too coarse to separate sound streams from a mixture of sounds.

Nakagawa et al. [13] improved the accuracy of the sound source direction by using the direction extracted by image processing, because the direction by vision is more accurate. 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.
2. HRTF is needed for creating a direction-pass filter.

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.

C. Sound Source Localization

Some robots developed so far had a capability of sound source localization. Huang et al. [7] developed a robot that had three microphones. Three microphones were 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 backward. 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.

Humanoids of Waseda University can localize a sound source by using two microphones [8], [9]. 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 [15], [16]. In the Cog, auditory localization is trained by visual information.

These approaches do not use HRTF, but assumes a single sound source and "*stop-perceive-act*" strategy. In some cases, human voices are obtained by a microphone attached to a speaker. To summarize, both approaches lack for the CASA viewpoints.

D. Active Perception

A humanoid is active in the sense that it tries to do some activity to improve perceptual processing. Such activity includes to change the position of cameras (active vision) and microphones (active audition) by motor control, and object tracking.

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. 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.

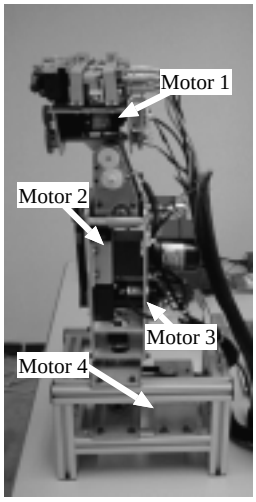
However, another problem occurs because gaze stabilization is attained by visual servo or auditory servo. Sounds are generated by motor rotation, gears, belts and ball bearings. Since these internal sound sources are much closer than other external sources, even if the absolute power of sounds is much 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.

E. Internal Sound Suppression

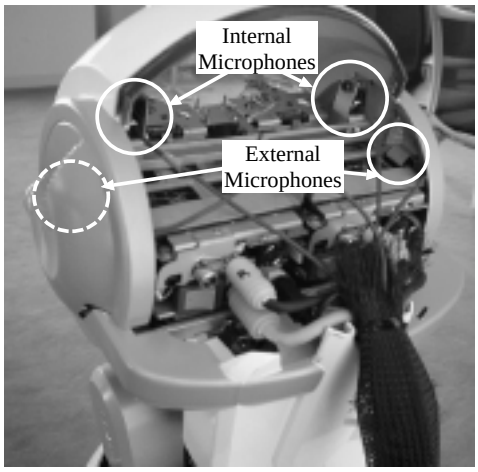
A cover affects the spectrum of sounds like a dummy head in which a pair of dummy headphones are embedded. This spectral effect is known as HRTF (Head-Related Transfer Function). HRTF plays an important function in localizing (calculating the position of) a sound source. To obtain HRTF by measuring impulse response for each spatial position is timeconsuming. In addition, HRTF depends on environments such as room, wall, objects and so on. Therefore, a new localization method without using HRTF or by adjusting HRTF to a real-world should be invented.

Since active perception causes sounds by the movement of various movable parts, internal sound suppression is critical to enhance external sounds. A cover of humanoid body reduces sounds of motors emitted to the external world by separating internal and external world of the robot. 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.

a) Cover



b) Mechanical structure



c) Internal microphones (top) and cameras

Fig. 1. *SIG* the Humanoid

III. Active Audition System

An active audition system consisting of two components; internal sound suppression, and sound stream separation is developed for an upper-torso humanoid.

A. SIG the humanoid

As a testbed of integration of perceptual information to control motor of high degree of freedom (DOF), we designed a humanoid robot (hereafter, referred as *SIG*) with the following components [17]:

- 4 DOFs of body driven by 4 DC motors — Its mechanical structure is shown in Figure 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 (Figure 1b).
- Two pairs of nondirectional microphones (Sony ECM-77S) (Figure 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 (Figure 1a) reduces sounds to be emitted to external environments, which is expected to reduce the complexity of sound processing.

B. Internal Sound Suppression 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 about 3.7 dB.
- Body motors — sounds of standby and movement are about 5.6 dB and 23 dB, respectively.

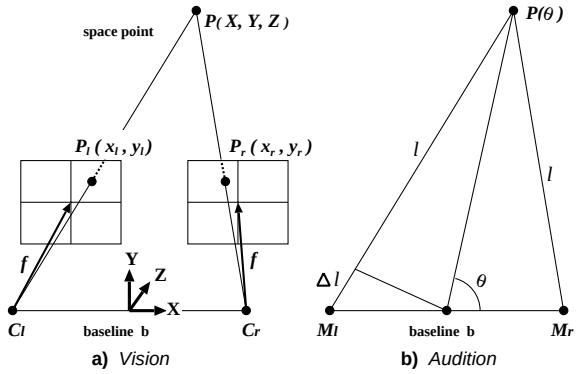
Comparison of noise cancellation by adaptive filtering, ICA, case-based suppression and model-based suppression, we concluded that only adaptive filters work well. Four microphones are not enough for ICA to separate internal sounds. Case-based and model-based suppression affect the phase of original inputs, which causes errors of IPD.

Our adaptive filter uses *heuristics about internal microphones*, which specifies the condition to cut off burst noise mainly caused by motors. Such burst noises include sounds at stoppers, by friction between cable and body, creaks at joints of cover parts may occur.

First, we store the data of the acoustic measurement in the system. The noise data of each motor is stored as a power spectrum of the averaged measured noises. Next, we use the stored data as templates to judge burst noises. When the motor makes a burst noise, the intensity of the noise is quite stronger because microphones location is relatively near the motor. Therefore if the spectrum and intensity of captured noise is similar to those of a noise template, the captured noise is regarded as a burst noise. Specifically, the subband is canceled if the following conditions are satisfied:

1. Intensity difference between external and internal microphones is similar to measured motor noise intensity differences.
2. Intensity and pattern of the spectrum are similar to measured motor noise frequency responses
3. A motor command is being processed.

We tried to make as adaptive filter an FIR (Finite Impulse Response) filter of order 100, because this filter is a linear phase filter. This property is essential to localize the sound source by IID (Interaural Intensity Difference) or ITD/IPD (Interaural Time/Phase Difference). The parameters of the FIR filter is calculated by least-mean-square method as adaptive algorithm. Noise cancellation by the FIR filter suppresses internal sounds but some errors occur. These errors make poor localization compared to results of localization without internal sound suppression. Case-based or model-based cancellation is not adopted, because



C_l, C_r: camera center, M_l, M_r: microphone center

Fig. 2. Epipolar geometry for localization

the same movement generates a lot of different sounds and thus it is difficult to construct case or model-based cancellation.

C. Sound Stream Separation by Localization

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

C.1 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 Figure 2a). 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 [18]:

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

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 [19]. 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 timeconsuming. 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.

C.2 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

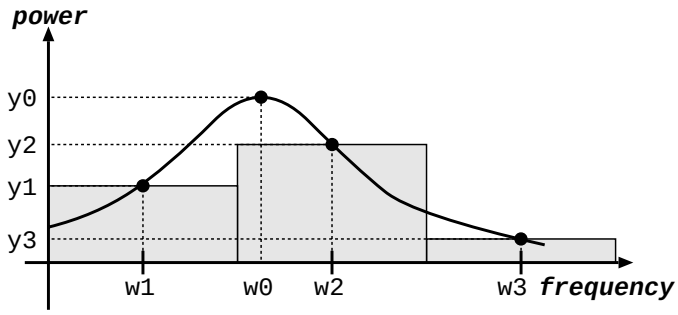


Fig. 3. A spectral peak by Fourier Transformation

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 $Sp^{(r)}$. 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.

C.3 Pitch Extraction

Pitches are extracted by a kind of spectral subtraction [20]. It uses peak approximation method based on characteristics of FFT and window function. When a peak $[\omega_2, y_2]$ is detected by FFT, usually it is not the true peak (see Fig. 3). Let $[\omega_1, y_1]$ and $[\omega_3, y_3]$ be values of both neighbors. The true peak $[\omega_0, y_0]$ is estimated as follows:

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

$$\begin{aligned} Arg(y_0) &= \tan^{-1}\left(\frac{\Im[y_0]}{\Re[y_0]}\right) \\ &= \tan^{-1}\left(\frac{\Im[y_2]}{\Re[y_2]}\right) + \frac{T}{2}(\omega_2 - \omega_0) \quad (2) \\ |y_0| &= \frac{\Delta\omega\left(-T^2\Delta\omega^2 + 4\pi^2\right)}{2\pi^2 \sin\frac{T}{2}\Delta\omega}|y_2|, \\ \Delta\omega &= \omega_2 - \omega_0 \quad (3) \end{aligned}$$

ω_0 is estimated as the following Equation (1). And the phase and amplitude of the true peak y_0 are estimated as Equations (2) and (3), respectively. $\Re[x]$ and $\Im[x]$ are the real and imaginary part of a complex number x.

Because the above equations require relatively the small number of calculation, our method can run faster and extract more accurate pitches. For example, in comparison with Bi-HBSS [11], which is known as a sound source separation system using a pitch extraction method by spectral

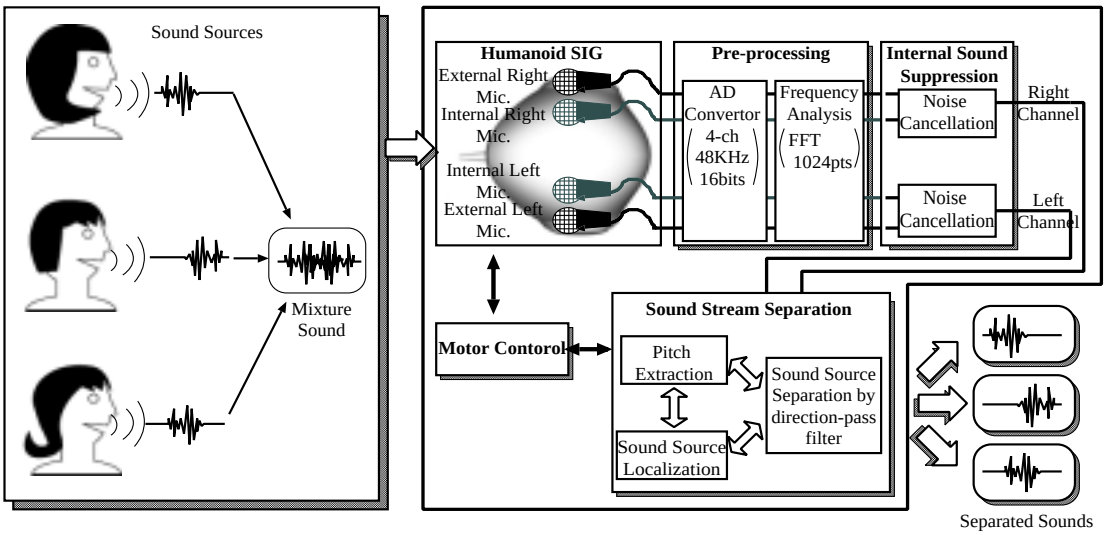


Fig. 4. Active Audition System

subtraction, our method needs only 1/200 of amount of calculation per a peak [21].

C.4 Direction-Pass Filter by Epipolar Geometry

As mentioned earlier, HRTF is 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, HRTF should be interpolated for auditory localization of a moving sound source, because HRTF is 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 $\triangle\varphi$ for each subband (47 Hz).
2. Extract peaks and calculated IPD, $\triangle\varphi'$.
3. If IPD satisfies the specified condition, namely, $\triangle\varphi' = \triangle\varphi$, 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 Figure 2b). 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:

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

C.5 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.

C.6 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 follows according to the above observation:

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

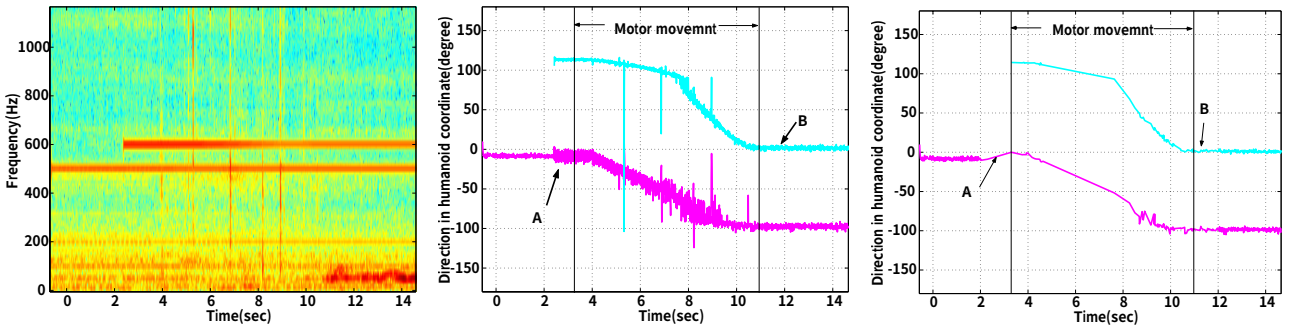
C.7 Sensor Integrated System

The system contains a perception system that integrates sound, vision, and motor control (Figure 4). 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 [19].

IV. Experiments of Sound and Visual Tracking

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.

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 noise. The sound environment is not controlled at all for experiments to ensure feasibility of the approach in daily life.



a) Spectrogram of input sound b) Localization without suppression c) localization by suppression method

Fig. 6. Localization Experiments of sound sources

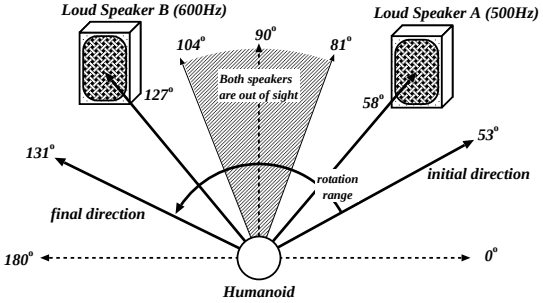


Fig. 5. Experiment: sound and visual tracking while *SIG* moves.

One sound source A (Speaker A) plays a monotone sound of 500 Hz, while the other one 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 distance from *SIG* to each sound source is about 210cm (the maximum distance in the room). 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. Figure 5 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.

Results: Results of the experiment were very promising. First, accurate sound source localization was accomplished without using the HRTF. The use of epipolar geometry for audition was proven to be very effective. Epipolar geometry based non-HRTF method locate approximate direction of sound sources (see localization data for initial 5 seconds in Figure 6c). In this figure, time series data for estimated sound source direction using only audition 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.

The effect of adaptive noise canceling is clearly shown. The effects of internal sound suppression by heuristics are shown in Figure 6c. Figure 6b shows estimated sound source directions without motor noise suppression. Sound direction estimation is seriously hampered when the head is moving (around time 5 - 6 seconds). The spectrogram (Figure 6a) clearly indicates extensive motor noise. When the robot 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.

Such accurate localization by audition makes association between audition and vision possible. 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 indicate that position estimation by audition and vision is accurate enough to create consistent association even under the condition that the robot is constantly moving and generating motor noise. It should be refined that sound source localization by audition in the experiment uses epipolar geometry for audition, and do not use HRTF. Thus, we can simply field the robot in unknown acoustic environment and localize sound sources.

V. Discussion and Future Work

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 Nakagawa et al. [13] 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.

Epipolar geometry gives a way to unify visual and auditory processing, in particular localization and sound stream separation. This approach can dispense with HRTF. As far as we know, no other systems can do it. Most robots capable of auditory localization developed so far use HRTF explicitly or implicitly, and may fail in identifying some spatial directions or tracking moving sound sources.

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.

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 [22]. 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.

As 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 suppression by independent component analysis, addition of more sensor information, and applications.

VI. Conclusion

In this paper, we present active audition for humanoid which includes internal sound suppression, 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, Dr. Iris Fermin, and Dr. Theo Sabish for their discussions. 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*, vol. 1, no. 4, pp. 333–356, 1987.

[2] G. J. Brown, *Computational auditory scene analysis: A representational approach*. University of Sheffield, 1992.

[3] M. P. Cooke, G. J. Brown, M. Crawford, and P. Green, "Computational auditory scene analysis: Listening to several things at once," *Endeavour*, vol. 17, no. 4, pp. 186–190, 1993.

[4] 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)*, pp. 100–107, AAAI, 1994.

[5] D. Rosenthal and H. G. Okuno, eds., *Computational Auditory Scene Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1998.

[6] 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)*, pp. 832–839, AAAI, 2000.

[7] 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.

[8] 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*, pp. 1723–1726, ESCA, 1999.

[9] 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*, vol. 20, pp. 24–32, 1995.

[10] 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*, vol. 2, pp. 77–80, 1994.

[11] 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)*, vol. 1, pp. 165–172, AAAI, 1995.

[12] H. G. Okuno, T. Nakatani, and T. Kawabata, "Listening to two simultaneous speeches," *Speech Communication*, vol. 27, no. 3-4, pp. 281–298, 1999.

[13] 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.

[14] 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)*, pp. 78–87, IJCAI, 1999.

[15] R. Brooks, C. Breazeal, M. Marjanovie, B. Scassellati, and M. Williamson, "The cog project: Building a humanoid robot," tech. rep., MIT, 1999.

[16] 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)*, pp. 54–58, IJCAI, 1997.

[17] 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)*, p. *to appear*, 2000.

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

[19] T. Lourens, K. Nakadai, H. G. Okuno, and H. Kitano, "Selective attention by integration of vision and audition," in *Proceedings of First IEEE-RAS International Conference on Humanoid Robot (Humanoid-2000)*, 2000.

[20] 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)*, pp. 200–203, IEEE, 1979.

[21] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Humanoid active audition system improved by the cover acoustics," in *PRICAI-2000 Topics in Artificial Intelligence (Sixth Pacific Rim International Conference on Artificial Intelligence)*, vol. Lecture Notes in Computer Science, No.1886, pp. 544–554, Springer Verlag, 2000.

[22] R. Brooks, C. Breazeal, M. Marjanovie, B. Scassellati, and M. Williamson, "The cog project: Building a humanoid robot," in *Lecture Notes in Computer Science*, p. *to appear*, Spriver-Verlag, 1999.