Real-Time Auditory and Visual Multiple-Object Tracking for Humanoids

Kazuhiro Nakadai[†], Ken-ichi Hidai[†], Hiroshi Mizoguchi[‡], Hiroshi G. Okuno[†§], and Hiroaki Kitano [†¶]

† Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.

Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya-ku, Tokyo 150-0001, Japan

‡ Department of Information and Computer Science, Saitama University, Saitama 338-8570, Japan

§ Department of Intelligence Science and Technology, Kyoto University, Kyoto 606-8501, Japan

¶ Sony Computer Science Laboratories, Inc., Tokyo 141-0022, Japan

{nakadai, hidai, okuno, kitano}@symbio.jst.go.jp, hm@me.ics.saitama-u.ac.jp

Abstract

This paper presents a real-time auditory and visual tracking of multiple objects for humanoid under real-world environments. Real-time processing is crucial for sensorimotor tasks in tracking, and multiple-object tracking is crucial for real-world applications. Multiple sound source tracking needs perception of a mixture of sounds and cancellation of motor noises caused by body movements. However its real-time processing has not been reported yet. Real-time tracking is attained by fusing information obtained by sound source localization, multiple face recognition, speaker tracking, focus of attention control, and motor control. Auditory streams with sound source direction are extracted by active audition system with motor noise cancellation capability from 48 KHz sampling sounds. Visual streams with face ID and 3D-position are extracted by combining skin-color extraction, correlation-based matching, and multiple-scale image generation from a single camera. These auditory and visual streams are associated by comparing the spatial location, and associated streams are used to control focus of attention. Auditory, visual, and association processing are performed asynchronously on different PC's connected by TCP/IP network. The resulting system implemented on an upper-torso humanoid can track multiple objects with the delay of 200 msec, which is forced by visual tracking and network latency.

1 Introduction

Humanoids and entertainment robots or at least mobile robots have attracted a lot of attention last year, e.g., at IEEE/RSJ First Humanoids-2000 conference, and are expected to play a role of human partners in the 21st century. Let us imagine the situation autonomous robots are used in social and home environment, such as a pet robot at living room, a service robot for office, or a robot serving people at a party. The robot shall identify people in the room, pay attention to their voice and look at them to identify visually, and associate voice and visual images, so that highly robust event identification can be accomplished. These are minimum requirements for social interaction [Brooks *et al.*, 1998].

Some robots are equipped with improved robot-human interface. *Jijo-2* [Asoh *et al.*, 1997] can recognize a phrase command by speech-recognition system; *AMELLA* [Waldherr *et al.*, 1998] can recognize pose and motion gestures. *Kismet* of MIT AI Lab [Breazeal and Scassellati, 1999] can recognize speeches by speech-recognition system and express various kinds of sensation. *Hadaly* of Waseda University [Matsusaka *et al.*, 1999] can localize the speaker as well as recognize speeches by speech-recognition system.

However, the technologies developed so far are still immature; in particular, auditory processing and integrated perception among vision, audition, and motor control. At robotic conferences such as IROS, SMC, and ICRA as well as AI-related conferences, there were at most one or two papers related to auditory processing[1], and most papers on robot perception is limited to vision-only and vision with ultrasonic, infra-red or laser range finders. This is unfortunate because integrated processing of auditory and visual processing combined with appropriate motor control is essential in social interaction of robot systems.

For auditory and visual tracking, Nakadai *et al.* presented the *active audition* for humanoids to improve sound source tracking by integrating audition, vision, and motor controls [Nakadai *et al.*, 2000]. An active audition system is implemented in a upper-torso humanoid to demonstrate that 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. Although such an active head movement inevitably creates motor noise, the system adaptively cancels motor noise 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 variety of conditions. One of crucial problems is a lack of real-time processing.

Matsuyama *et al.* presented an architecture for asynchronous coordination of sensorimotor control [Matsuyama *et al.*, 2000] so that the camera moves smoothly to track the

[1] Auditory processing shall be distinguished from speech recognition. Auditory processing (auditory scene analysis) aims at separation and understanding of multiple sound streams, such as human speech, environmental noise, music, etc.

object in real-time. This architecture, called *dynamic memory*, is general, but they use it only for vision-based motor control. Shafer *et al.* presented the software architecture of sensor fusion for an autonomous mobile robot [Shafer *et al.*, 1986]. The architecture is based on a parallel blackboard system, and the sensors include vision, range finder, but not microphones. It exploits global consistency regarding position and orientation of the vehicle and sensors. Murphy presented the sensor fusion system for mobile robots called the Sensor Fusion Effects (SFX) architecture, which is based on the uncertainty management system by Dempster-Shafer theory [Murphy, 1998].

Other robots with microphones as ears for sound source localization or sound source separation have attained little in auditory tracking. *Kismet* has a pair of omni-directional microphones outside the simplified pinnae [Breazeal and Scassellati, 1999]. Since it is designed for one-to-one communication and its research focuses on social interaction based on visual attention, the auditory tracking has not been implemented so far. *Hadaly* uses a microphone array to perform sound source localization, but the microphone array is mounted in the body and its absolute position is fixed during head movements [Matsusaka *et al.*, 1999]. In the both cases, sound source separation is not exploited and a microphone for speech recognition is attached to the speaker.

In the research of computational auditory scene analysis (CASA) to understand a mixture of sounds, real-time processing is one of the main problems in applying CASA to real-world applications [Rosenthal and Okuno, 1998]. Real-time processing is important to take appropriate actions in daily environments where many people, robots and objects exist. Auditory and visual tracking by Nakadai *et al.* accepts sounds in daily environment, but does not run in real-time [Nakadai *et al.*, 2000]. Another auditory and visual tracking by Nakagawa *et al.* does not run in real-time [Nakagawa *et al.*, 1999].

Two major issues that have not been done in the past are attacked in this paper, that is, association of multiple auditory and visual streams, and real-time processing of integrated auditory and visual scene analysis.

The rest of the paper is organized as follows: Section 2 explains the robot hardware which is used as a testbed and presents the issues in real-time tracking. Section 3 describes the design of the system and the details of each module are described in Section 4. Section 5 demonstrates and evaluates the performance of the system. Section 6 discusses the observations of the experiments and future work and concludes the paper.

2 Issues in Real-Time Tracking

2.1 Robot Hardware

As a testbed of real-time multiple-object tracking, we use a upper-torso humanoid called *SIG* shown in Fig. 1 [Nakadai *et al.*, 2000]. The cover of the mechanics is made of FRP and discriminates internal and external world acoustically. *SIG* has two microphones at the left and right ear positions to capture external sounds from outside of the body, and two microphones within the body to capture internal sounds mainly caused by motor movements. All the microphones are omni-directional microphones of Sony ECM-77S. *SIG*'s body has four DOFs (degree of freedom), each of which is a DC motor controlled by a potentiometer. *SIG* is equipped with a pair of CCD cameras of Sony EVI-G20, but the current vision module uses only one camera.

Figure 1: Humanoid, *SIG*

2.2 Task and Issues

The task in this paper is to track multiple objects in real-time with two kinds of sensors, a camera and two microphones, and one actuator to rotate the body. Some important issues in this task are:

- Robustness in visual stream extraction (temporal sequences of face localization and face identification) against non-uniform environments due to lighting conditions or moving humans.
- Robustness in auditory stream extraction (temporal sequence of sound source localization and sound source separation) against dynamic environments because objects (people) move and the humanoid also moves.
- Association of visual and auditory streams by face identification and sound source localization to compensate missing or ambiguous data. Common representation for both auditory and visual feature extractions is needed.
- Focus of attention control based on association to control actuators.
- Trade-off of processing speed vs quality of feature extractions for real-time processing.
- Method of synchronization between asynchronous auditory and visual processing that have different processing speeds. The frame rate of vision is 30 Hz, while the sampling rate of sound is 48 KHz. Therefore, asynchronous processing is essential to exploit the full range of concurrency.

3 Design of the System

From the viewpoint of functionality, the whole system can be decomposed into five layers — *SIG* Device Layer, Process Layer, Feature Layer, Event Layer, and Stream Layer

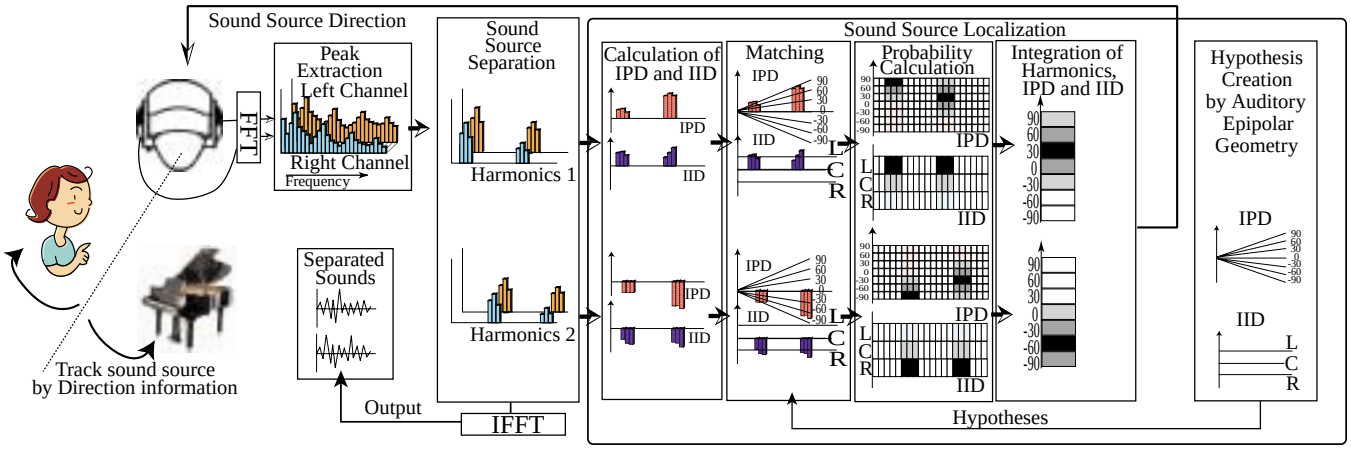

Figure 3: Audition Module extracts harmonic structures to localize and separate sound sources.

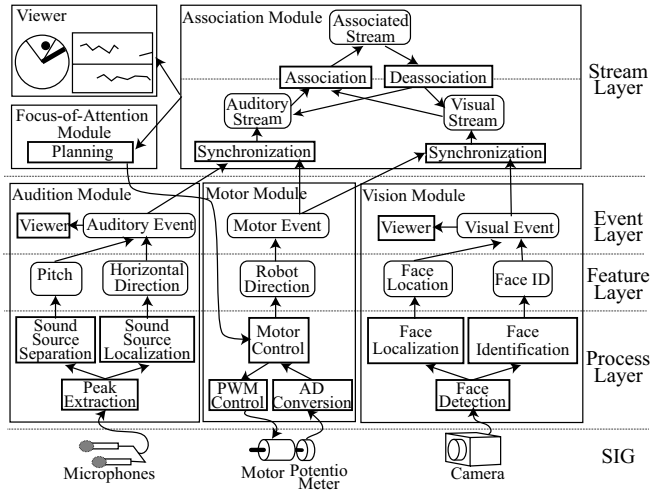

Figure 2: Modules and Layers of the System

(see Fig. 3) . From the viewpoint of implementation, the whole system consists of six asynchronous modules — Audition, Vision, Association, Focus-of-Attention, Motor Control, and Viewer. This relation between two viewpoints is depicted in Fig. 2. Since Vision module utilizes the full power of Pentium-III 733 MHz CPU, the whole system is organized in the form of distributed processing with three Linux nodes based on Pentium-III with RedHat Linux 6.2J. The first node with 600 MHz is for Audition, the second node with 733 MHz for Vision, and the third with 450 MHz for the rest. They are connected by TCP/IP over Fast Ethernet 100Base-TX.

The estimated processing time of each module executed on a node is summarized below:

- Vision – 200 msec for face localization and identification,
- Audition — 40 msec for sound source localization,
- Motor Control — 100 msec
- Network latency — up to 200 msec

Therefore, we set the goal that the response time of the system should be 200 msec of delay.

Audition and Vision generate an *event* by feature extraction and organize a *stream* as a temporal sequence of events. Motor Control also generates an event of motion. Association fuses these events to make a higher level representation. This fusion *associates* auditory and visual streams to make an *associated* stream. Focus-of-Attention makes a planning of *SIG*'s movement based on the status of streams, that is, whether they are associated or not.

Motor Control is activated by Focus-of-Attention module and generates PWM (Pulse Width Modulation) signals to DC motors. It also sends a motor event consisting of motor direction (azimuth of the midsagittal plain) to Association module. Viewer shows the status of auditory, visual and associated streams in the radar and scrolling windows (see screen shots shown in Fig. 5). Some modules are explained in details in the next section.

4 Details of Each Module

4.1 Active Audition Module

Sound localization for a robot or an embedded system is usually solved by using interaural phase difference (IPD) and interaural intensity difference (IID). These values are calculated by using Head-Related Transfer Function (HRTF). However, HRTF depends on the shape of head and it also changes as environments change. For real-world applications, sound localization without HRTF is preferable. Nakadai *et al.* proposed the method based on the auditory epipolar geometry, an extension of epipolar geometry in stereo vision to audition [Nakadai *et al.*, 2000]. They also proposed *active audition* for sensorimotor task with canceling motor and mechanical noises. However, they failed in doing the jobs in real-time, because they stuck to pure-tone processing. In this paper, we extend their approach (1) by exploiting the harmonic structure to extract peaks precisely and (2) by solving the uncertainty in sound source localization by Dempster-Shafer theory.

Audition module equipped with active audition is depicted in Fig. 3. The input signal, a mixture of sounds originating from different directions, is sampled with sampling frequency

Table 1: Belief Factor of IID, $BF_{\mathrm{IID}}(\theta)$

	θ	$90^\circ \sim 35^\circ$	$30^\circ \sim -30^\circ$	$-35^\circ \sim -90^\circ$
I	+	0.35	0.5	0.65
	-	0.65	0.5	0.35

of 48 KHz and 16-bit quantization, and its spectrogram is calculated by Fast Fourier Transforms (FFT). Audition extracts pitches (fundamental frequency, $F0$), separates and localizes sound sources.

Peak Extraction and Sound Source Separation: First a peak is extracted by a band-pass filter, which passes a frequency between 90 Hz and 3 KHz if its power is a local maximum and more than the threshold. This threshold is automatically determined by the stable auditory conditions of the room. Then, extracted peaks are clustered according to *harmonicity*. A frequency of Fn is grouped as an overtone (integer multiple) of $F0$ if the relation $\left|\frac{Fn}{F0} - \lfloor\frac{Fn}{F0}\rfloor\right| \leq 0.06$ holds. The constant, 0.06, is determined by trial and error. By applying Inverse FFT to a set of peaks in harmonicity, a harmonic sound is separated from a mixture of sounds.

Sound Source Localization: Once a harmonic structure is obtained, the direction of sound source is calculated by hypothetical reasoning for IPD (Interaural Phase Difference) and IID (Interaural Intensity Difference). The azimuth (horizontal direction) is quantized and represented by every 5° discrete value in the range of $\pm 90^\circ$. The front direction of *SIG* is 0°.

From the extracted harmonic structure of left and right channels, a pair of harmonic structures is obtained. Then the IPD, P_s, is calculated. Auditory Epipolar Geometry generates a hypothesis of IPD P_h for each 5° candidate, θ [Nakadai *et al.*, 2001]. Since the IPD is ambiguous for frequencies of more than 1200 Hz, the distance, $d(\theta)$, in IPD between the data and a hypothesis is defined as follows:

$$d(\theta) = \frac{1}{n_{f<1200\mathrm{Hz}}} \sum_{f=F0}^{1200\mathrm{Hz}} \frac{\left(P_h(\theta, f) - P_s(f)\right)^2}{f} \qquad (1)$$

Where $n_{f<1200Hz}$ is the number of overtones of which frequency is less than 1200 Hz.

The similar relation may hold for IID, but our experience with IID proves that it can discriminate at most the side, that is, left or right. Suppose that $I_s(f)$ is the IID for peak frequency f. If the value of $I = \sum_{f=1200\mathrm{Hz}}^{3000\mathrm{Hz}} I_s(f)$ is non-negative, the direction is decides as left, otherwise as right:

Integration of IPD and IID by Dempster-Shafer theory To determine the sound source direction, the belief factors of IPD and IID are calculated and then integrated by Dempster-Shafer theory. The belief factor of IPD, BF_{IPD}, is calculated by using probability density function defined by Eq. (2).

$$BF_{\mathrm{IPD}}(\theta) = \int_{-\infty}^{\frac{d(\theta)-m}{\sqrt{\frac{s}{n}}}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \qquad (2)$$

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

The belief factor of IID, $BF_{\mathrm{IID}}(\theta)$ is defined by Table 1.

Then, belief factors of IPD and IID, BF_{IPD} and BF_{IID}, are integrated using Dempster-Shafer theory as defined in Eq. (3).

$$BF_{\mathrm{IPD+IID}}(\theta) = BF_{\mathrm{IPD}}(\theta) BF_{\mathrm{IID}}(\theta) + \left(1 - BF_{\mathrm{IPD}}(\theta)\right) BF_{\mathrm{IID}}(\theta) + BF_{\mathrm{IPD}}(\theta)\left(1 - BF_{\mathrm{IID}}(\theta)\right) \qquad (3)$$

θ for the maximum $BF_{\mathrm{IPD+IID}}$ is treated as the sound source direction of the harmonics. Finally, Audition sends an auditory event consisting of pitch ($F0$) and a list of 20-best directions (θ) with reliability factor for each harmonics.

4.2 Real-Time Multiple Face Tracking

Multiple face detection and identification suffers more severely from frequent changes in the size, direction and brightness of face. To cope with this problem, Hidai *et al.* combines skin-color extraction, correlation based matching, and multiple scale images generation [Hidai *et al.*, 2000].

The requirements on multiple face tracking are the capability of discriminating face data of the same face ID from others and on-line learning. The first requirement is a class concept. Turk *et al.* proposed the eigenface matching technique as a kind of subspace method [Turk and Pentland, 1991]. A subspace for discrimination is created by Principal Component Analysis (PCA). PCA, however, does not provide the means to group a data according to its face ID, since such an ID cannot be generated by PCA. Thus, the subspace obtained by PCA is not always suitable to distinguish such classes.

On the other hand, Liner Discriminant Analysis (LDA) can create an optimal subspace to distinguish classes. Therefore, we use Online LDA [Hiraoka *et al.*, 2000]. In addition, this method continuously updates a subspace on demand with a small amount of computation.

The face identification module (see Fig. 2) projects each extracted face into the discrimination space, and calculates its distance d to each registered face. Since this distance depends on the degree (L, the number of registered faces) of discrimination space, it is converted to a parameter-independent probability P_v as follows.

$$P_v = \Gamma\left(\frac{1}{2}, \frac{d^2}{2}\right) = \int_{\frac{d^2}{2}}^{\infty} e^{-t}\, t^{\frac{L}{2}-1} dt \qquad (4)$$

The face localization module converts a face position in 2-D image plane into 3-D world coordinate. Suppose that a face is $w \times w$ pixels located in (x, y) in the image plane, whose width and height are X and Y, respectively (see screen shots shown in Fig. 5). Then the face position in the world coordinate is obtained in terms of distance r, azimuth θ and elevation ϕ by the following equations.

$$r = \frac{C_1}{w},\ \theta = \sin^{-1}\left(\frac{x - \frac{X}{2}}{C_2\, r}\right),\ \phi = \sin^{-1}\left(\frac{\frac{Y}{2} - y}{C_2\, r}\right)$$

where C_1 and C_2 are constants defined by the size of the image plane and the image angle of the camera.

Finally, Vision module sends a visual event consisting of a list of 5-best Face ID (Name) with its reliability and position (distance r, azimuth θ and elevation ϕ) for each face.

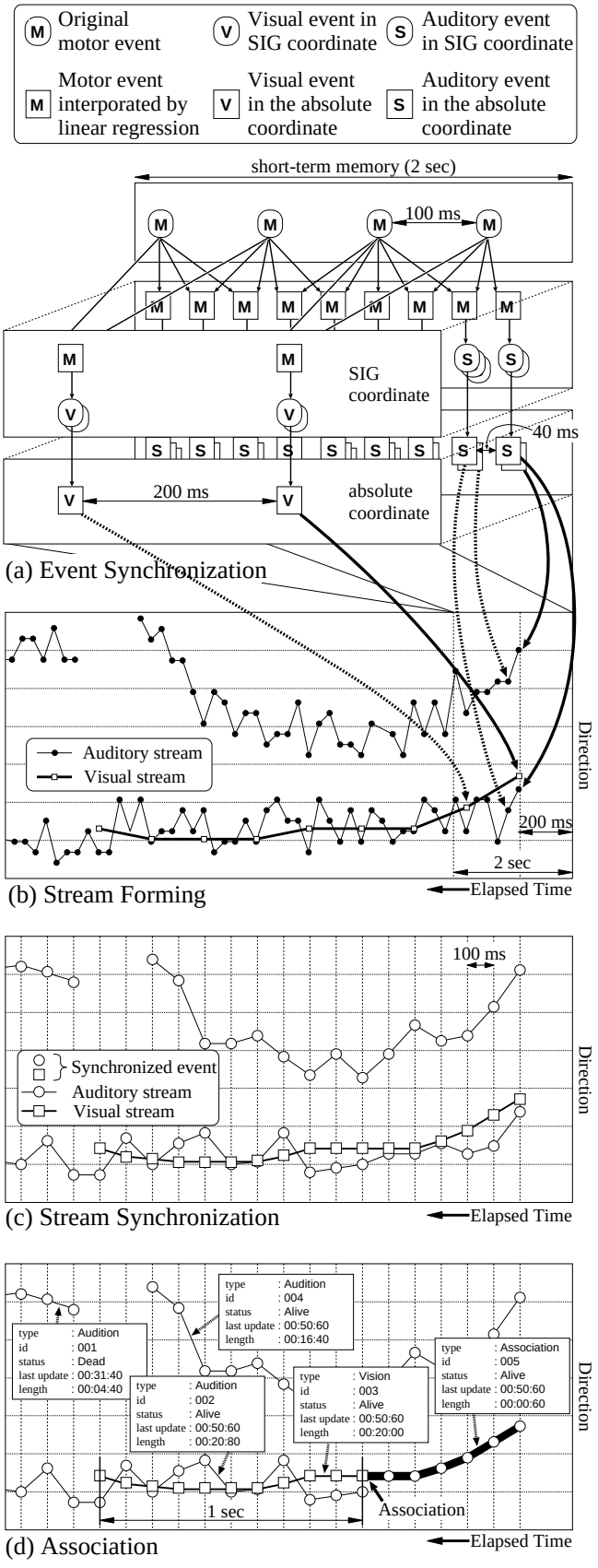

Figure 4: Association Module Forms Streams

4.3 Stream Formation and Association

Association module forms auditory streams from auditory events and visual streams from visual events, and associates a pair of auditory and visual streams to create a higher level stream, which is called an *associated stream* (see Fig. 2). The flow of processing in stream formation and association is summarized as follows (Figs. 4(a)-(d)):

1. Events from Audition, Vision and Motor modules are stored in the short-term memory.
2. Direction information of events is converted into the absolute coordinate to treat them in the common coordinate.
3. Events are grouped into an auditory or a visual stream according to a temporal sequence of events.
4. Streams are synchronized by every 100 msec to calculate the distance between streams.
5. An auditory and a visual stream which are close for more than a constant time are associated as an associated stream.

First, events are stored in the short-term memory and kept only for 2 seconds to attain incremental and real-time processing. In Fig. 4(a), where Ⓢ, Ⓥ and Ⓜ represent events created by Audition, Vision and Motor modules, respectively. Each module creates events at its own cycle, e.g. 40 msec for audition, 200 msec for vision and 100 msec for motion. Then, motor events are synchronized with auditory and visual events. To put it concretely, a motor direction when an auditory or a visual event appeared is estimated from motor events in the short-term memory. A motor event with the estimated motor direction is shown as [M]. Because visual events from Vision module and auditory events from Audition module are represented in robot coordinate, the directions of these events are converted to ones in the absolute coordinate by using estimated motor direction. These are represented as [S] or [V] in Fig. 4(a). This synchronization process runs with a delay of 200 msec as mentioned in Sec. 3.

Auditory and visual streams are formed in Fig. 4(b). X-axis indicates elapsed time from right to left, and Y-axis indicates azimuth in the absolute coordinate. Thin lines with small filled circles and a thick line with small rectangles represent auditory streams and a single visual stream. An auditory event is connected to the nearest auditory stream within the range of $\pm 10^\circ$ and with common $F0$. A visual event is connected to the nearest visual stream within 40 cm and with a common face ID. In either case, if there are multiple candidates, the most reliable one is selected. If any appropriate stream is found, such an event becomes a new stream. In case that no event is connected to an existing stream, such a stream remains alive for up to 500 msec. The system cannot detect an auditory or a visual event when a person stops talking and looks away for a moment. This margin of 500 msec is prepared to continue streams in case of the missing event extraction. After 500 msec of keep-alive state, the stream terminates.

When the distance between an auditory and a visual stream is close for more than a constant time, they are regarded as

streams originating from the same object and integrated into an associated stream, which is a higher layer representation of a stream shown in Fig. 2. Because auditory and visual streams consist of events with 40 msec and 200 msec cycles, respectively, it is difficult to evaluate the distance between these two streams without synchronization. Then, they are synchronized with the same cycle, 100 msec. Fig. 4(c) illustrates synchronized streams as lines with large circles and rectangles. If an event is not available in this case, linear regression is used for interpolation in the same way as synchronization with motor events.

An auditory and a visual streams are associated if their direction difference is within the range of $\pm 10^\circ$ and this situation continues for more than 50% of the 1 sec period shown in Fig. 4(d).

The visual direction is usually used for the direction of the associated stream because visual information is more accurate. However, when a tracking person is occluded, the system cannot use visual information. In this case, auditory information is used for the associated stream. This suggests an advantage of integration of audition and vision, i.e. auditory information is efficient not only for pre-attentive uses such as a trigger of attention but also for compensations of missing or ambiguous information as this case. If either auditory or visual event has not been found for more than 3 sec, such an associated stream is deassociated and only existing auditory or visual stream remains. If the auditory and visual direction difference has been more than 30° for 3 sec, such an associated stream is deassociated to two separate streams.

4.4 Focus of Attention Control

SIG should pay attention for sounds from unseen objects to get further information. When such a sound does not exist, faces with sound, i.e. talking people, should have high priority because they are attractive even for human perception. The principle of focus-of-attention control hereby is as follows:

1. An auditory stream has the highest priority,
2. an associated stream has the second priority, and
3. a visual stream has the third priority.

The algorithm of focus-of-attention control is sketched by using an example shown in Fig. 5, which depicts how auditory and visual streams are generated and associated.

1. Focus of attention changes to a new association stream. [t_1 and t_8 of Fig. 5].
2. If one of the visual and auditory stream of an associated stream terminates due to occlusion, disappearance, or end of speech, association continues [t_4 to t_5 of Fig. 5].
3. If this state continues for a particular time, say 3 seconds, the focus of attention may change.
   (a) Focus of attention changes to one of associated streams.
   (b) If no associated stream is found, focus of attention changes to one of auditory streams. [t_6 of Fig. 5].
   (c) Otherwise, focus of attention changes to one of visual streams.
4. In turning the body to associate the auditory stream to visual one, focus of attention keeps the same even if a new associated stream is generated.

5 Experiments and Evaluation

A 40-second scenario shown in Fig. 5 is used as a benchmark. The performance of integrated auditory and visual tracking is shown in Fig. 5, which shows that focus of attention changes twice. In the first half of the scenario up to $t = 26$ sec, two speakers are apart, while in the second half they are close and viewed in the same camera view field. In both cases, the system can track the speakers well.

The direction of *SIG*'s body is depicted in Fig. 6, which shows that the motor control succeeds in giving correct PWM motor commands. To sum up, sensorimotor task in single- and multi-speaker tracking is well accomplished.

The performance of visual tracking is shown in Fig. 7. This timechart is generated by collecting the first candidate from the internal states of Vision module. Therefore, the motor movement is the same as the above. In the first half of the scenario, occlusion causes a gap of visual streams between t_4 and t_5. From t_6 to t_7, no person can be seen due to the limited range of camera view field. Fig. 5 proves that occlusion and out-of-sight can be easily recovered by associated streams.

The performance of auditory tracking is shown in Fig. 8, which is generated in the same manner as Fig. 7. The Audition module can separate two auditory streams correctly from t_3 to $t = 23$ sec, and t_9 to t_{10}, but generate erroneous streams around t_8 and t_9. In addition, the directions of two speakers are not so correct from $t = 11$ sec (t_5) to $t = 17$ sec, because Mr. A moves and *SIG* tracks him by rotating its body. That is, the reverberation (echo) due to this moving talker and motor noise deteriorates the quality of sound source localization.

5.1 Limitations on the Proposed System

The room used in this experiment is about $3m$ in width and length and $2m$ in height, and sound absorbing materials are attached on walls, ceiling and floor. It is not anechoic, but has reverberation time of 0.1 sec. Because the value in a normal speech studio of the equivalent size is about 0.2 sec, the room has less reverberation. Acoustic conditions, however, depends on objects in the room. When we put a plastic partition of 1.5 m $\times$ 1.5 m with strong reverberation in the room, the correctness of sound source localization is reduced remarkably.

The background noise level of the room is 30 dBA on average. This is measured with the filter by *A weighting*, similar to human auditory characteristic. The value of 30 dBA corresponds to the noise level of a room in a quiet residence. We confirmed that the current system works well up to 40 dBA of background noise level, but we have not checked above 40 dBA.

The room has six halogen lights with adjustment function of the intensity on the ceiling. The range of the light intensity is from 0 to 50 lux. Because the light intensity from 300 to 1500 lux is recommended for a normal office by JIS (Japanese Industrial Standard), even the maximum light intensity of the room is weak. Our face extraction and recognition method works well under the condition of more than 5 lux. Visual

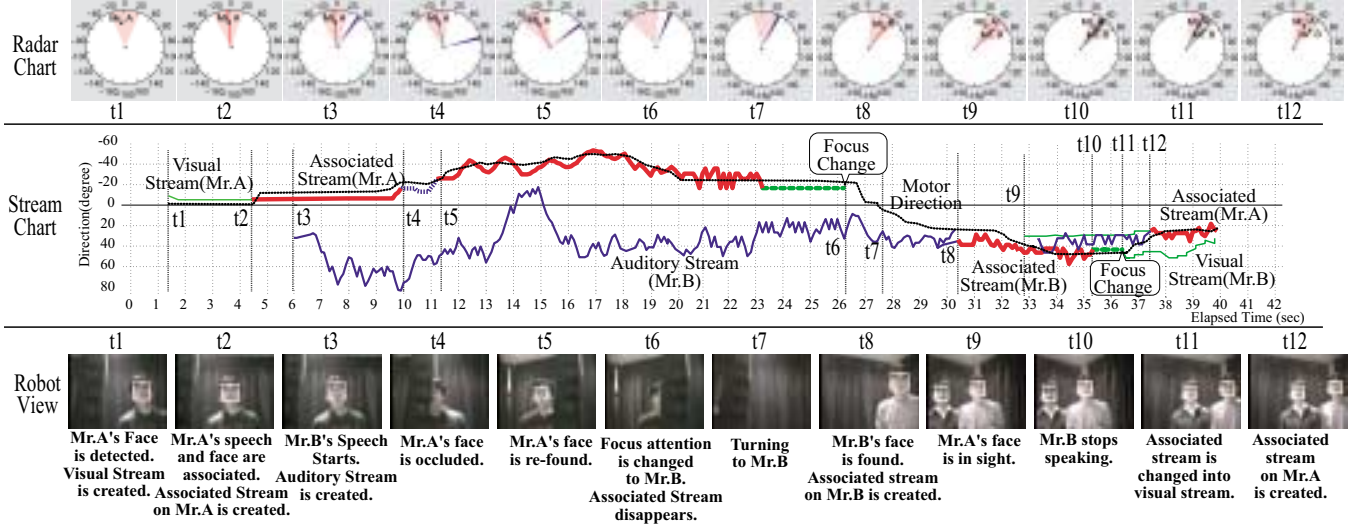

Figure 5: Temporal Sequence of Auditory and Visual Tracking of Two Speakers: **Radar Chart** and **Stream Chart** are screen shots of the viewer. In radar chart, a wide-light and a narrow-dark sector indicate the camera view field and sound source direction, respectively. In stream chart, a thin line indicates auditory or visual stream, while a thick line indicates an associated stream.

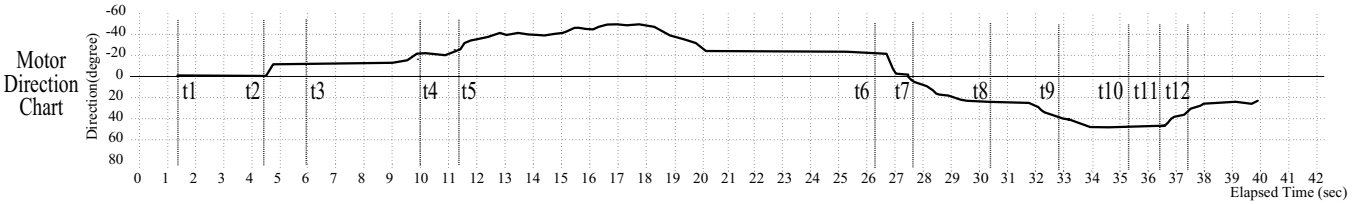

Figure 6: Temporal Sequence of Body Direction controlled by Motor Movement in the same scenario as Fig. 5

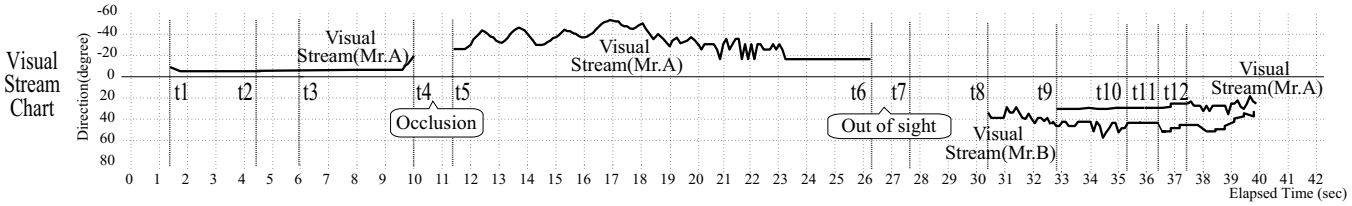

Figure 7: Temporal Sequence of Visual Tracking of Two Speakers in the same scenario as Fig. 5

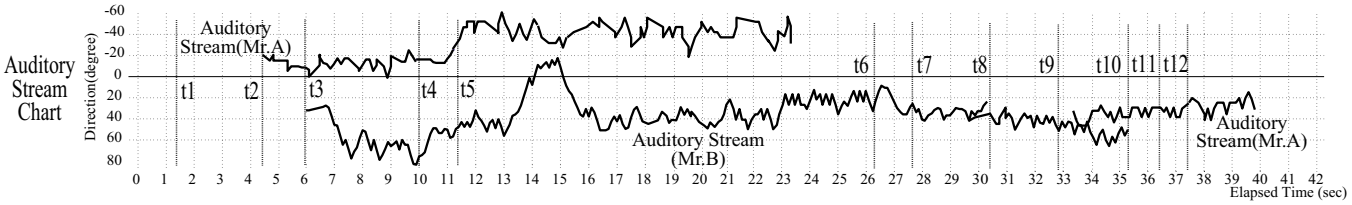

Figure 8: Temporal Sequence of Auditory Tracking of Two Speakers in the same scenario as Fig. 5

processing is robust against the change of light intensity. In this experiment, the light intensity in the room is 50 lux.

Other benchmarks such as crossing of moving talkers, moving talkers without seeing any talkers, and alternative talking of four speakers prove that the resulting system succeeds in real-time sensorimotor tasks of tracking [2].

[2]Since the work is related to the real-time processing, the readers may be suggested to visit the following Web site: http://www.symbio.jst.go.jp/SIG/

6 Conclusion and Future Work

The key idea of real-time tracking is "*For each processing, take it easy, and ambiguities will be resolved with the help of others.*" This idea is obtained by the scrutiny of the behavior of each component of implementations of Nakadai *et al.*'s work [Nakadai *et al.*, 2000]. We do not stick to pure tones, but utilize the collective behavior of harmonic sounds; we prefer frequency resolution over the time resolution by increasing the points of FFT. We give up the precise face localization and identification. Instead, we associate auditory, visual, and motor direction information to localize the sound sources.

Some technical future work includes learning the adaptive association of different or dynamic environments. Since lighting conditions and reverberation (echo) change drastically in such environments, Vision and Audition modules should adjust their parameters on demand. In addition, Association should adapt its parameters for stream forming and association. Baysian algorithm for resolving ambiguities in stream forming and association is a promising technique. Another future work is incorporating stereo vision. Even in a static environment robust auditory processing such as sound source separation and localization would be useful when a room is noisier, has objects with strong reverberation, or has many people.

We believe that our result would open a new era of sound processing, in particular, cocktail party computer or "Shotoku-Taishi" computer that can listen to several things at once.

Auditory and visual tracking should be incorporated in a total system with robot-human interface. We have already built such a system comprising speech recognition, speaker identification, and speech synthesis based on the proposed system. Once the application is fixed, the top-down stream separation may be exploited. Some information that forces top-down stream separation includes speaker identification. The speaker information may reduce the search space of face recognition and speech recognition. For example, let us consider that crossing of two talking persons. In this case, the system may miss judging that they are approaching and then receding because speaker IDs are lacking. Thus, speaker identification can reduce this kind of ambiguity. Its design and implementation will be reported by a separate paper, since this paper focuses on the real-time processing.

Acknowledgments

We thank our colleagues of Symbiotic Intelligence Group, Kitano Symbiotic Systems Project; Mr. Tatsuya Matsui and Dr. Tino Lourens, and Prof. H. Ishiguro of Wakayama University for their discussions.

References

[Asoh *et al.*, 1997] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui. Socially embedded learning of the office-conversant mobile robot *jijo-2*. In *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, volume 1, pages 880–885. AAAI, 1997.

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

[Brooks *et al.*, 1998] R. A. Brooks, C. Breazeal, R. Irie, C. C. Kemp, M. Marjanovic, B. Scassellati, and M. M. Williamson. Alternative essences of intelligence. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 961–968. AAAI, 1998.

[Hidai *et al.*, 2000] K. Hidai, H. Mizoguchi, K. Hiraoka, M. Tanaka, T. Shigehara, and T. Mishima. Robust face detection against brightness fluctuation and size variation. In *Proceedings of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2000)*, pages 1397–1384. IEEE, 2000.

[Hiraoka *et al.*, 2000] K. Hiraoka, S. Yoshizawa, K. Hidai, M. Hamahira, H. Mizoguchi, and T. Mishima. Convergence analysis of online linear discriminant analysis. In *Proceedings of IEEE-INNS-ENNS International Joint Conference on Neural Networks*, volume III, pages 387–391, 2000.

[Matsusaka *et al.*, 1999] 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 6th European Conference on Speech Communication Technology (EUROSPEECH-99)*, pages 1723–1726. ESCA, 1999.

[Matsuyama *et al.*, 2000] T. Matsuyama, S. Hiura, T. Wada, K. Murase, and A. Yoshida. Dynamic memory: Architecture for real time integration of visual perception, camera action, and network communication. In *Proceedings of ICCV*, pages 728–735. IEEE, 2000.

[Murphy, 1998] R.R. Murphy. Dempster-shafer theory for sensor fusion in autonomous mobile robots. *IEEE Transactions on Robotics and Automation*, 14(2):197–206, 1998.

[Nakadai *et al.*, 2000] 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.

[Nakadai *et al.*, 2001] K. Nakadai, H. G. Okuno, and H. Kitano. Epipolar geometry based sound localization and extraction for humanoid audition. *submitted*, 2001.

[Nakagawa *et al.*, 1999] 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.

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

[Shafer *et al.*, 1986] S.A. Shafer, A. Stentz, and C.E. Thorpe. An architecture for sensor fusion in a mobile robot. In *Proceedings of IEEE Conference on Robotics and Automation*, pages 2002–2011. IEEE, 1986.

[Turk and Pentland, 1991] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[Waldherr *et al.*, 1998] S. Waldherr, S. Thrun, R. Romero, and D. Margaritis. Template-based recoginition of pose and motion gestures on a mobile robot. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 977–982. AAAI, 1998.