Real-Time Multiple Speaker Tracking by Multi-Modal Integration for Mobile Robots

†*Kazuhiro Nakadai*, †*Ken-ichi Hidai*, * *Hiroshi G. Okuno*, *and* ‡*Hiroaki Kitano*

†Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.
* Graduate School of Infomatics, Kyoto University
‡Sony Computer Science Laboratories, Inc.
{nakadai, hidai}@symbio.jst.go.jp, okuno@nue.org, kitano@csl.sony.co.jp

Abstract

In this paper, real-time multiple speaker tracking is addressed, because it is essential in robot perception and human-robot social interaction. The difficulty lies in treating a mixture of sounds, occlusion (some talkers are hidden) and real-time processing. Our approach consists of three components; (1) the extraction of the direction of each speaker by using interaural phase difference and interaural intensity difference, (2) the resolution of each speaker's direction by multi-modal integration of audition, vision and motion with canceling inevitable motor noises in motion in case of an unseen or silent speaker, and (3) the distributed implementation to three PCs connected by TCP/IP network to attain real-time processing.

As a result, we attain robust real-time speaker tracking with 200 ms delay in a non-anechoic room, even when multiple speakers exist and the tracking person is visually occluded.

1. Introduction

Recently, robots such as a pet at a living room and a service robot for entertainment and welfare attract a great deal of attention. They are expected to be peers of human in the future. For such robots, a function of tracking speakers is important and fundamental because the robots 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. Speaker tracking can induce fertile human behaviors as well because of achievement of more friendly interaction.

In addition, it naturally requires active behaviors of a robot. The active behaviors are essential in robot perception. Perception with active behaviors, i.e. active perception, is natural and common in human and animals for better perception. A robot controlled by motors should be active to understand the surrounded environments. A robot can see and hear better by active behaviors such as focusing attention on a specific direction. Therefore, speaker tracking is minimally necessary for robots from the viewpoints of social interaction and active perception.

We consider four issues which are necessary for robust speaker tracking; these are active audition, general sound understanding, multi-modal integration and real-time processing.

Active audition is an expansion of active perception to auditory processing. 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. However, in the auditory research field, audition with behaviors has not been studied extensively even though people hear sounds while in motion. Indeed, some of robotics researches mention auditory processing with motion, but they assume that the maximum number of sound sources is at most one and the input sound is loud enough to ignore motor noises[1]. These assumptions are not enough to understand high-level auditory functions. Then, we presented the *active audition* for humanoids to improve sound source tracking by integrating audition, vision and motor controls [2]. Although 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.

Usually we hear a mixture of sounds, not a sound of single source. It is necessary for understanding general sounds to understand a mixture of sounds. In speaker tracking, multiple persons can speak simultaneously. In such case, a robot have to separate and localize sound sources. *Computational Auditory Scene Analysis* (CASA) is a framework of understanding a mixture of sounds, and a lot of works have been studied in this area[3][6]. Therfore, the concepts of CASA should be introduced to be applied in the real environment.

Multi-modal integration is also required. The error in direction determined by a CASA application is about ±10°[3], which is similar to that of a human, i.e. ±8°[4]. However, this is too coarse to separate sound streams from a mixture of sounds. And in visual processing, there are other problems such as narrow visual field of an ordinary camera and visual occlusion on overlapping persons. It is difficult to solve these problems by only visual or auditory processing. Therefore integration of vision and audition is necessary. Some works succeed in the integration, but they are studied in the simulated environment[5]. Audition and vision should be integrated in a real environment for robust speaker tracking.

Finally, real-time processing is important to take appropriate actions in daily environments where many people, robots and objects exist. However, one of the main problems in applying CASA to real-world applications is real-time processing[6]. In addition, a crucial problem in active audition is a lack of real-time processing[2].

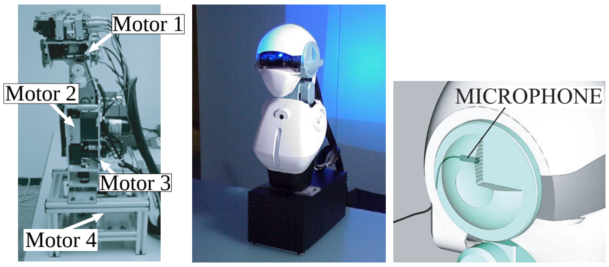
We implement a speaker tracking system to solve these problems, and demonstrate robust tracking with integration audition and vision in a non-anechoic room.

The paper is organized as follows: Section 2 presents humanoid *SIG*, our testbed robot. Section 3 presents the speaker tracking system in detail. Section 4 shows evaluation of the system, and Section 5 gives conclusion.

2. Humanoid *SIG*

As a testbed of speaker tracking in the real world, we designed a humanoid robot (hereafter, referred to as *SIG*) [7].

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



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

Figure 1: *SIG* the humanoid

3. Real-Time Speaker Tracking System

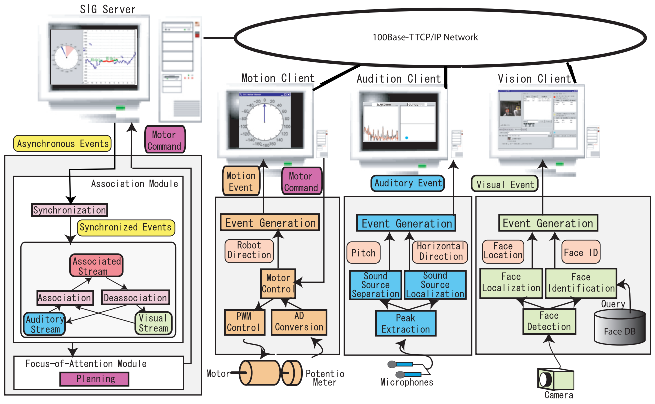


Figure 2: Real-time Speaker Tracking System

Fig. 2 depicts the logical structure of the system based on client/server model. Each server or client executes the following modules:

1. Audition — extracts auditory events by pitch extraction, sound source separation and localization, and send those events to Association,
2. Vision — extracts visual events by face extraction, identification and localization, and then send visual events to Association,
3. Motor Control — generates PWM (Pulse Width Modulation) signals to DC motors and sends motor events to Association,
4. Association — integrates various events to create streams,
5. Focus-of-Attention — makes a plan of motor control, and
6. Viewer — instruments various streams and data.

Since the system should run in real-time, the above clients are physically distributed to three Pentium-III based Linux nodes connected by TCP/IP over 100Base-TX network and run asynchronously.

3.1. Audition: Real-Time Sound Source Tracking

Audition module can analyze a mixture of sound and process sounds in real-time using two microphones. And Auditory epipolar geometry and Dempster-Shafer theory are introduced for robust localization of harmonic sounds. As a result, this module sends an auditory event consisting of pitch ($F0$) and a list of 20-best direction (θ) with reliability for each harmonics.

3.1.1. *Pitch Extraction*

First, to extract pitches, a power spectrum is filtered by a band-pass filter that the pass band is from 90 to 3000 Hz and frequencies where the power is more than a threshold are permitted to be passed. The threshold is determined experimentally. Then a filtered spectrum is clustered by a process that when two clusters include neighboring sub-bands each other, they are unified into one. This processing is repeated until no neighboring sub-bands between any two clusters is found. As a result, peaks are extracted a set of samples with the strongest power in each cluster. Finally, harmonic sounds with strong power such as vowel are extracted using harmonic relationships against the extracted peaks. Our system does not need to obtain a whole speech because we do not use speech recognition.

3.1.2. *Auditory Epipolar Geometry*

Head Related Transfer Function (HRTF) is often used for binaural sound source localization. However, it is easy to change if surrounded environment is changed. So, HRTF is hard to use for sound source localization in real environments. *Auditory Epipolar Geometry* is proposed to extract directional information of sound sources without using HRTF [2]. In stereo vision research, epipolar geometry is one of the most common localization method [8]. Auditory epipolar geometry is an extension of epipolar geometry in vision to audition as shown in Fig 3. Since auditory epipolar geometry extracts directional information by using the geometrical relation, it can dispense with HRTF. And we take the cover form of *SIG* into account because a sound can reach at most an ear directly due to the cover form. For example, in Fig. 4, a sound has to travel along the cover because left path is not direct to the left ear from a sound source. Then, the sound direction θ is given by Eq. (1) using the *Interaural Phase Difference* (IPD) as the difference of phases between right and left signals.

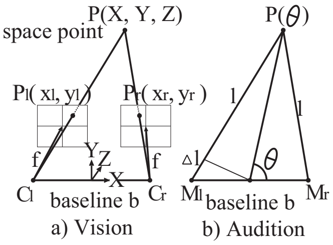
$$\theta = F^{-1}\left(\frac{v}{2\pi f}\Delta\varphi\right) \qquad (1)$$

where where v is the velocity of sound, $\Delta\varphi$ is IPD and f is the frequency of sound and F^{-1} is the inverse function of F which represents the difference of distance between left and right ears from an infinite sound source given by Eq. (2).

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

3.1.3. *Matching using Auditory Epipolar Geometry*

The sound source direction θ is assumed within $\pm 90^\circ$ by 10° from the median plane of *SIG*. The system uses IPD and *Inter-*



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

Figure 3: Epipolar geometry for localization

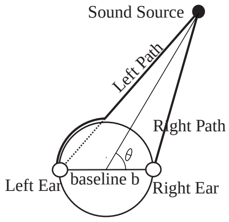


Figure 4: Cover Influence

aural Intensity Difference (IID) for sound source localization. As for sound source localization by IPD, the system creates hypotheses of IPD by Eq. (1) for each θ. Then it calculates the distance between each hypothesis and IPD of input harmonics at frequencies of lower than 1200 Hz because IPD is efficient in the frequencies according to the baseline of *SIG* (about 18cm). The cost function is as follows:

$$d(\theta) = \frac{1}{n_{f<1200\mathrm{Hz}}} \sum_{f=\mathrm{F0}}^{1200\mathrm{Hz}} (P_h(\theta, f) - P_s(f))^2 / f \tag{3}$$

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

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

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

3.1.4. *Integration of IPD and IID by Dempster Shafer theory*

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

$$p(\theta) \quad = \quad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{d(\theta)-m}{\sqrt{s/n}}} \exp\left(-\frac{x^2}{2}\right) dx \tag{4}$$

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

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

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

Table 1: IID belief factor

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

$$\begin{aligned} P_{IPD+IID}(\theta) \quad = \quad & P_{IPD}(\theta)P_{IID}(\theta) + (1 - P_{IPD}(\theta))P_{IID}(\theta) \\ & + P_{IPD}(\theta)(1 - P_{IID}(\theta)) \end{aligned} \tag{5}$$

As a result of the integration, 20-best θs are sent per a sound according to $P_{IPD+IID}$.

3.2. Vision: Face Identification Module

Since the visual processing detects several faces simultaneously, extracts, identifies and tracks each face, the size, direction, and brightness of each face changes frequently. The key idea is the combination of skin-color extraction, correlation based matching, and multiple scale images generation [9].

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.

$$P_v = \int_{\frac{d^2}{2}}^{\infty} e^{-t} t^{\frac{L}{2}-1} dt \tag{6}$$

The discrimination matrix is created in advance or on demand using a set of variation of the face with an ID (name). This analysis is done by Online Linear Discriminant Analysis [10].

The face localization module converts a face position in 2-D image plane into 3-D world space. 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. Then the face position in the world space is obtained as a set of azimuth θ, elevation ϕ, and distance r as follows:

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

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.

3.3. Stream Formation and Association

Association synchronizes the results (events) given by other modules. It forms an auditory, visual or associated stream by their proximity. Events are stored in the short-term memory only for 2 seconds. Synchronization process runs with the delay of 200 msec, which is the largest delay of the system, that is, vision module.

An auditory event is connected to the nearest auditory stream within ±10° and with common or harmonic pitch. A visual event is connected to the nearest visual stream within 40 cm and with common face ID. In either case, if there are plural 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. After 500 msec of keep-alive state, the stream terminates.

An auditory and a visual streams are associated if their direction difference is within ±10° and this situation continues for more than 50% of the 1 sec period.

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.

3.4. Focus-of-Attention

Focus-of-Attention Control is based on continuity and triggering. By continuity, the system tries to keep the same status, while by triggering, the system tries to track the most interesting object. This time, we fix that an associated stream has higher priority of attention than auditory and visual streams.

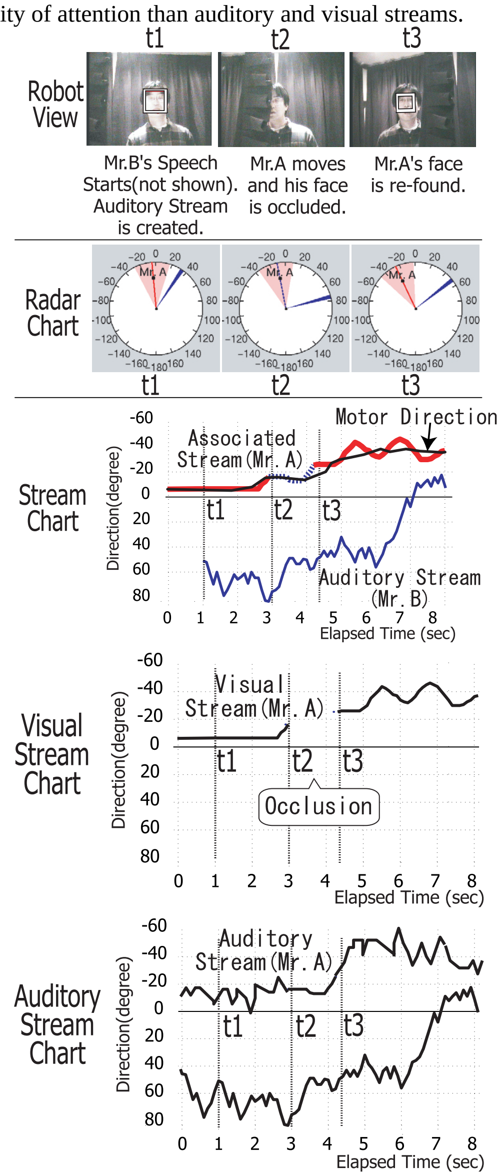


Figure 5: Multiple Speaker Tracking using *SIG*

4. Evaluation

The system is evaluated through tracking people when they are speaking and moving. The time sequence of the experiment is shown in Fig. 5. Mr. A and B stand at 15° left and 60° right to *SIG*at the non-anechoic room of $10m^2$, respectively. First, *SIG* is looking at Mr. A. Because he is talking and facing to *SIG*, *SIG* can detect his face and voice. And an associated stream on Mr. A is created. Then Mr. B starts talking at $t1$. They continue to talk until $t = 8(s)$. Mr. A starts moving at $t = 2.5(s)$, while Mr. B does not change his position through the experiment. *SIG* is configured to track the first speaker, Mr. A. So, *SIG* turns its body according to Mr. A's direction. Mr. A is occluded by a curtain at $t2$, and appears again at $t3$. He continues to speak during the occlusion, so *SIG* can track him by only audition. Two speeches are separated well from "Auditory Stream Chart" which is the localization result by only audition. The error of sound source localization of Mr. A is about $\pm 15°$, judging from difference between motor and sound direction. The error is within the same range even when Mr. A is moving, so motor control works well because *SIG* is tracking Mr. A smoothly. This means that sound source localization with motion works well. The associated stream in "Stream Chart", which is the localization result by the integration, shows the accuracy of localization is improved by using vision information. And, due to visual occlusion between $t2$ and $t3$, *SIG* cannot continue tracking of Mr. A by only visual information from "Visual Stream Chart" which is the localization result by only audition. In such case, *SIG* can continue tracking using auditory information in "Stream Chart".

5. Conclusion

We presented real-time multiple speaker tracking system by integration of audition and vision. It can localize simultaneous two speeches in a non-anechoic room using active audition and CASA concepts. In addition, we attained robust tracking by the multi-modal integration. Higher level integration using speech recognition and speaker identification to make the system robuster is a future work.

6. References

[1] 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*. 1999, pp. 1723–1726, ESCA.

[2] 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)*. 2000, pp. 832–839, AAAI.

[3] 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)*. 1995, vol. 1, pp. 165–172, AAAI.

[4] S. Cavaco and J. Hallam, "A biologically plausible acoustic azimuth estimation system," in *Proceedings of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99)*. 1999, pp. 78–87, IJCAI.

[5] 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)*. 1999, pp. 768–775, AAAI.

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

[7] H. Kitano, H. G. Okuno, K. Nakadai, T. Sabish, and T. Matsui, "Design and architecture of sig the humanoid," in *Proceedings of International Conference on Robotics and Systems 2 000 (IROS 2000)*, 2000, pp. 181–190.

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

[9] 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)*. 2000, pp. 1397–1384, IEEE.

[10] K. Hiraoka, M. Hamahira, K. Hidai, H. Mizoguchi, T. Mishima, and S. Yoshizawa, "Fast algorithm for online linear discriminant analysis," in *Proceedings of ITC-2000*. 2000, pp. 274–277, IEEK/IEICE.