Real-Time Speaker Localization and Speech Separation by Audio-Visual Integration

Kazuhiro Nakadai*, Ken-ichi Hidai*, Hiroshi G. Okuno*,†, Hiroaki Kitano*,‡
* Kitano Symbiotic Systems Project, ERATO, Japan Science and Tech. Corp., Tokyo, Japan
† Graduate School of Informatics, Kyoto University, Kyoto, Japan
‡ Sony Computer Science Laboratories, Inc., Tokyo, Japan
okuno@nue.org, nakadai@symbio.jst.go.jp, kitano@csl.sony.co.jp

Abstract— **Robot audition in real-world should cope with motor and other noises caused by the robot's own movements in addition to environmental noises and reverberation. This paper reports how auditory processing is improved by audio-visual integration with active movements. The key idea resides in hierarchical integration of auditory and visual streams to disambiguate auditory or visual processing. The system runs in real-time by using distributed processing on 4 PCs connected by Gigabit Ethernet. The system implemented in a upper-torso humanoid tracks multiple talkers and extracts speech from a mixture of sounds. The performance of epipolar geometry based sound source localization and sound source separation by active and adaptive direction-pass filtering is also reported.**

Keywords— **robot audition, audio-visual integration, multiple speaker tracking, sound source localization, sound source separation**

I. Introduction

Robust perception is essential to robots for rich and intelligent social interaction. This robustness should be attained by integration of multi-modal sensory input, because a single sensory input carries inevitable ambiguities. Among various perception channels, *active perception* is one of promising techniques to improve perception. In vision, *active vision* is proposed to control camera parameters to attain better visual perception, and a lot of research on active vision has been performed [1]. The concept of "*active*" should be extended to other media.

Active audition is also proposed to control microphone parameters to attain better auditory perception [2]. Although sound is the most important medium for human communication and life, only a little attention is paid to it in robotics. This is partially because the research on social interaction of robots has started only recently [3]. IROS 2001 is the first major robotics conference that has a session on "Sound and Speech". Most work reported so far, however, has not used robot's ears (microphones) for social interaction with humans.

The difficulties in robot audition, in particular, active audition, reside in sound source separation under real world environments. Active perception, audition or vision, involves motor movements, which make auditory processing more difficult. Therefore, one approach to avoid this problem is to adopt the "stop-hear-act" principle; that is, a robot stops to hear. Another approach is to use a microphone attached near the mouth of each speaker for automatic speech recognition. The latter examples include *Kismet* of the MIT AI Lab [4] and *ROBITA* of Waseda University [5].

The technical issues in sound source separation during movement include active noise cancellation, adaptation to dynamic environment, and sound source separation itself. Since the current technology of beam forming for microphone arrays assumes that the microphone array should be fixed, mobile robots equipped with a microphone array on them cannot meet the above requirements. Independent Component Analysis (ICA) has recently been a popular technique for sound source separation [6]. It can handle reverberation of a room to some extent, but in ICA, the maximum number of sound sources is limited to the number of microphones. This assumption usually does not hold in the real world. In addition, motor noise cancellation in motion as well as dynamic environmental change by active motion makes the performance of ICA poorer.

Computational auditory scene analysis (CASA) studies a general framework of sound processing and understanding [7], [8], [9], [10]. Its goal is to understand an arbitrary sound mixture including speech, non-speech sounds, and music in various acoustic environments. However, most of the sound source separation systems work only in off-line and simulation environments. For example, Bi-HBSS [9] uses *Head Related Transfer Function (HRTF)* for sound source separation by binaural processing. HRTFs are measured in an anechoic room, and are usually not available in real-world environments, because these are prone to environmental changes. In addition, it takes a lot of time to measure HRTFs. Therefore, sound source separation without HRTFs should be developed for robot

Fig. 1. Humanoid *SIG*

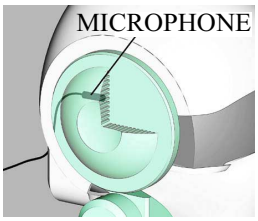

Fig. 2. *SIG* microphone

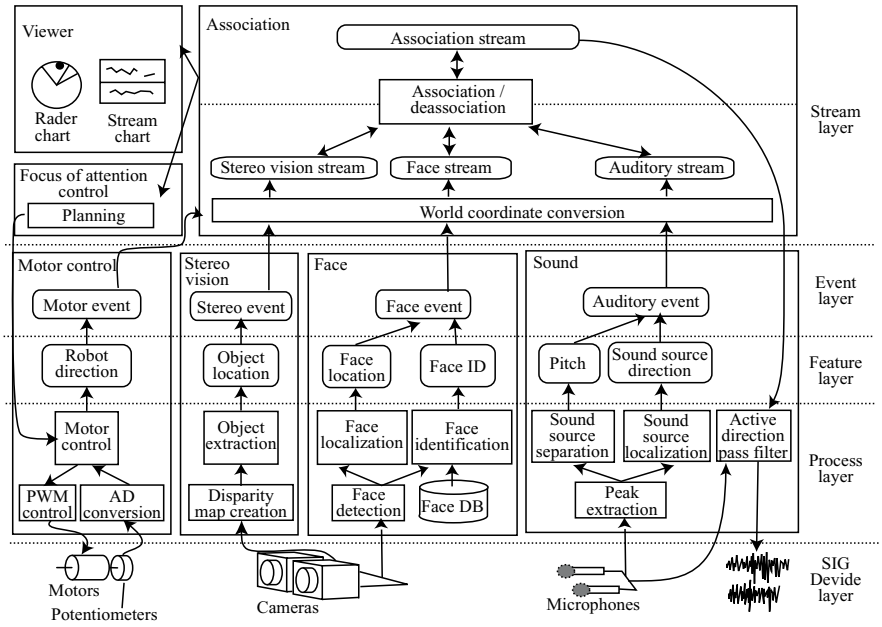

Fig. 3. Hierarchical Architecture of Real-Time Tracking System

audition.

A real-time multiple speaker tracking system has been developed by integrating audition and vision [11]. For auditory processing, the system uses active audition, which can perform sound source localization in a residential room by a new localization method without HRTFs and motor noise cancellation in motion by using cover acoustics. For visual processing, multiple face detection and recognition are used. By integrating auditory and visual processing with distributed processing on PCs, the system can track several people in real-time even when occlusion and two simultaneous speeches occur.

This system, however, has the following limitations:

1. Face recognition fails in the case of a partial face such as a profile.
2. No sound source separation is possible.
3. The communication load is almost 100% on Fast Ethernet (100Mbps).
4. The implementation cannot be scaled, using more processing nodes, to attain real-time processing.

In this paper, these limitations will be overcome by the following improvements:

1. Stereo vision is introduced for robust face recognition
2. Sound source separation is performed by an *active direction-pass filter* which takes sensitivity of direction into account.
3. Gigabit Ethernet is used and load distribution is introduced.
4. A more general implementation is adopted.

This paper reports the first three functionalities in detail and mentions the last one briefly.

The rest of this paper is organized as follows: Section II describes our humanoid *SIG* and the real-time human tracking system. Section III explains sound source separation by active direction-pass filter. Section IV shows evaluation of the system. The last section provides discussion and conclusion.

II. The Real-Time Human Tracking System

We use the upper torso humanoid *SIG* shown in Fig. 1 as a testbed for multi-modal integration. *SIG* has a cover made of FRP (fiber reinforced plastic). It is designed to separate the *SIG* inner world from the external world acoustically. A pair of CCD camera (Sony EVI-G20) is used for stereo vision. Two pairs of microphones are used for auditory processing. One pair is located in the left and right ear position for sound source localization (Fig. 2). The other is installed inside the cover mainly for canceling self-motor noise in motion. *SIG* has 4 DC motors (4 DOFs) with functions of position and velocity control by using potentiometers.

Fig. 3 shows the architecture of the real-time human tracking system using *SIG*. The system consists of seven modules, i.e., Sound, Face, Stereo Vision, Association, Focus-of-Attention, Motor Control and Viewer.

Sound, Face and a new module Stereo Vision generate an *event* by feature extraction. Motor Control also generates an event of motion. Association forms *streams* as temporal sequences of these events and associates these streams into a higher level representation, an *association* stream. Focus-of-Attention plans *SIG*'s movement

based on the status of streams, associated or not. Motor Control is activated by the Focus-of-Attention module and generates PWM (Pulse Width Modulation) signals to DC motors. Viewer shows the status of auditory, visual and associated streams in the radar and scrolling windows.

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*. The *SIG Device Layer* includes sensor equipment such as cameras, microphones and the motor system. They send images from cameras and acoustic signals from microphones to the *Process Layer*. In *Process Layer*, various features are extracted from raw data such as images and signals, and they are sent to the *Feature Layer*. Features are transformed to events with observed time for communication, then they are sent from the *Event Layer* to the *Stream Layer*. In the *Stream Layer*, event coordinates are converted into world coordinates. They are connected by taking their time series into account to make a stream. When two streams are close enough to be regarded as originating from a single source, they are associated into an association stream. Such an association stream gives *SIG* strong attention.

A. Real-Time Processing

Modules are distributed to four PCs of Pentium III 1GHz running RedHat Linux 7.1J. Although our previous system realized real-time processing with three PCs, one more PC is added to the system because of the introduction of Stereo Vision, which requires a lot of CPU power. This addition of one PC increases load average of communication. To reduce the communication load, each node in our current system has two network interfaces of Fast Ethernet and Gigabit Ethernet. Because Sound, Face, Stereo Vision and Motor create a lot of events for asynchronous communication, Gigabit Ethernet is used for event communication. Fast Ethernet is used for light communication such as synchronization by NTP (network time protocol). The system can work in real-time with a small latency of 500ms and synchronize modules with time difference within 100 μs, because the system is designed to select a suitable interface according to the properties of communication.

B. Sound Module

Generally, humans often use sounds for understanding the surroundings. However, it is difficult for a computer because of reverberation, environmental noises and their dynamic change. Sound module can cope with a mixture of sounds, i.e, it can separate sound sources and localize them in the real world. Robust localization is not achieved by only one sound clue, but by integration of several sound clues. The rest of this section describes the flow of auditory processing.

Peak Extraction and Sound Source Separation: First, a STFT (Short-Time Fourier Transform) is applied to the input sound. A peak on spectrum is extracted by a band-pass filter, which lets a frequency between 90 Hz and 3 KHz pass if its power is a local maximum and more than the threshold. This threshold is automatically determined by 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 $|\frac{Fn}{F0} - \lfloor \frac{Fn}{F0} \rfloor| \leq 0.06$ holds. The constant, 0.06, is determined by trial and error. By applying an Inverse FFT to a set of peaks in harmonicity, a harmonic sound is separated from a mixture of sounds.

Sound Source Localization: Robust sound source localization in the real world is achieved by four stages of processing, i.e., 1.localization by interaural phase difference (IPD) and auditory epipolar geometry, 2.localization by interaural intensity difference (IID), 3.integration of overtones, and 4.integration of 2. and 3. by Dempster-Shafer theory.

HRTF is of less use in the real world because HRTF depends on the shape of head and it also changes as environments change. Therefore, instead of HRTF, we use auditory epipolar geometry[12], which is an extension of epipolar geometry in stereo vision to audition, for sound source localization by IPD. Auditory epipolar geometry generates a hypothesis of the IPD for each 5° candidate. The distance between each hypothesis and the IPD of the input sound is calculated. IPDs of all overtones are summed up by using a weighted function. It is converted into belief factor B_{P} by using a probability density function (PDF).

For localization by IID, by calculating summation of IID of all overtones, belief factors supported by the left, front, and right direction are estimated.

Thus, Sound estimates sound directions by IPD and by IID with belief factors. Then, the belief factors of B_{P} and B_{I} are integrated into a new belief factor of $B_{\mathrm{P+I}}$ supported by both of them using Dempster-Shafer theory defined by

$$B_{\mathrm{P+I}}(\theta) = B_{\mathrm{P}}(\theta)B_{\mathrm{I}}(\theta) + \big(1 - B_{\mathrm{P}}(\theta)\big)B_{\mathrm{I}}(\theta) + B_{\mathrm{P}}(\theta)\big(1 - B_{\mathrm{I}}(\theta)\big). \quad (1)$$

Finally, Sound sends an auditory event consisting of pitch ($F0$) and a list of 20-best directions (θ) with reliability factors and observation times for each harmonics.

C. Face Identification Module

Face detects, recognizes and localizes multiple faces, and sends face events. To implement on a robot and apply to a real world, this module employs fast and robust processing for frequent changes in the size, direction and brightness of a face.

The face detection submodule detects multiple faces robustly by combining skin-color extraction, correlation based matching, and multiple scale image generation [13].

Then, the face recognition submodule can identify each detected face by Linear Discriminant Analysis (LDA), which can create an optimal subspace to distinguish classes and continuously update a subspace on demand with a small amount of computation [14].

The face localization submodule converts a face position in the 2-D image plane into 3-D world coordinates by assuming average face size.

Finally, Face sends a face event consisting of a list of 5-best Face ID (Name) with reliabilities, observation time and position (distance r, azimuth θ and elevation ϕ) for each face.

D. Stereo Vision Module

Stereo Vision is introduced to improve the robustness of the system. It can do what our previous system could not: track a person who looks away and does not talk. Stereo Vision enables tracking of such a person. In addition, accurate localization of lengthwise objects such as people is achieved by using a disparity map.

First, a disparity map is generated by an intensity based area-correlation technique. This is processed in real-time on a PC by a recursive correlation technique and an optimization peculiar to Intel architecture [15].

In addition, left and right images are calibrated by affine transformations in advance. An object is extracted from a 2-D disparity map by assuming that a human body is lengthwise. A 2-D disparity map is defined by

$$DM_{2D} = \{D(i,j)|i = 1,2,\cdots W, j = 1,2,\cdots H\} \quad (2)$$

where W and H are width and height, respectively and D is a disparity value.

As a first step to extract lengthwise objects, the median of DM_{2D} along the direction of height shown as Eq. (3) is extracted.

$$D_l(i) = Median(D(i,j)) \quad (3)$$

A 1-D disparity map DM_{1D} as a sequence of $D_l(i)$ is created.

$$DM_{1D} = \{D_l(i)|i = 1,2,\cdots W\} \quad (4)$$

Next, a lengthwise object such as a human body is extracted by segmentation of a region with similar disparity in DM_{1D}. This achieves robust body extraction so that only the torso can be extracted when the human extends his arm. Then, for object localization, epipolar geometry is applied to the center of gravity of the extracted region.

Finally, Stereo Vision creates stereo vision events which consist of distance, azimuth and observation time.

E. Association Module

Association forms a stream by connecting events to a time course, and associates the streams to create a higher level stream, which is called an *association stream*.

Stream Formation: Since location information in sound, face, stereo vision events is observed in a *SIG* coordinate system, event coordinates should be converted into world coordinates by comparing a motor event observed at the same time.

The converted events are connected to a stream with some error corrections according to the following algorithm, and a non-connected event generates a new stream.

- **Sound Event:** A sound event is connected to a sound stream when it satisfies two conditions that they have harmonic relationship, and their direction difference is within $\pm 10^\circ$. The value of $\pm 10^\circ$ is determined according to the accuracy of auditory epipolar geometry.
- **Face and Stereo Vision Event:** A face or a stereo vision event is connected to a face or a stereo vision stream when they have the same event ID and their distance is within 40 cm. The value of 40 cm is defined by assuming that human motion speed is less than 4 m/sec.

A stream is terminated if there is no event to be connected for more than 500 ms.

The advantages of stream formation are detection of object (human body) tracks and disambiguation of temporary errors of pitch detection and face recognition.

Association: When the system judges that multiple streams originate from the identical person, they are associated into an association stream, higher level stream representation. When one of the streams forming an association stream is terminated, the terminated stream is removed from the association stream, and the association stream is deassociated to some separated streams.

The advantage of association is an improvement of robustness by disambiguation of missing information, e.g., temporary occlusion can be compensated by

sound stream and sound direction can be compensated by more accurate visual information.

F. Focus-of-Atenttion

Focus-of-Attention selects a *SIG* action by audio-visual servo to keep the direction of a stream with attention and sends motor events to Motor. The principle of focus-of-attention control is as follows:

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

III. Active Direction Pass Filter

The direction-pass filter extracts sound originating from a specific direction by hypothetical reasoning about the IPD and IID of each sub-band [16]. Hypothetical reasoning compares actual IPD and IID with ideal ones which are calculated based on HRTF. This filter can extract not only harmonic sounds but also non-harmonic sound such as unvoiced consonants. The direction may be given by vision or by audition itself. Since the direction obtained by vision is much more accurate, that obtained by audition is used only in case when visual direction is not available due to occlusion. The filter improves the accuracy of sound source separation and is shown effective in automatic speech recognition of three simultaneous speeches in a clean environment. It, however, has some severe problems as follows:

- It is not robust in the real world, because IPD and IID are calculated by HRTF.
- It does not take into account the sensitivity of the direction-pass filter, although the accuracy of direction-pass filter depends on the direction, that is, higher sensitivity in the front while lower by deviating from it.
- HRTF is available only at discrete points.

To cope with these problems in the real world, we propose an *active direction-pass filter* based on auditory epipolar geometry, which is shown in Fig. 4. The algorithm is described as follows:

1. Direction of a stream with current attention is obtained from Association.
2. Because the stream direction is obtained in world coordinates, it is converted into azimuth θ in the *SIG* coordinate system by considering latency of processing.
3. The IPD $\triangle\varphi$ of θ is calculated for each sub-band by auditory epipolar geometry.
4. Peaks are extracted from the input and IPD $\triangle\varphi'$ is calculated.
5. If the IPD satisfies the specified condition, namely, $|\triangle\varphi' - \triangle\varphi| \leq \delta(\theta)$, then the sub-band is collected. $\delta(\theta)$ is determined by measurement. Because the *SIG* front direction has maximum sensitivity, δ has a minimum value. δ has a larger value at the side directions because of lower sensitivity.
6. A wave consisting of collected sub-bands is constructed.

The active direction-pass filter can improve sound source separation in the real world by supporting active motion of *SIG* and controlling adaptive sensitivity according to direction. In addition, sound source separation can work properly even when a sound source and/or *SIG* itself may be moving, because it obtains an accurate direction from the stream representation in Association module. Note that the direction of an associated stream is specified by visual information not by auditory one.

IV. Evaluation

The performance of the active direction-pass filter is evaluated by four kinds of experiments. In these experiments, *SIG* and loud speakers are located in a room of 10 square meters. The distance between *SIG* and the speakers is 50cm. The direction of a loud speaker is represented as 0° for *SIG* front direction.

Two metrics are used for evaluation; difference of SNR (signal-noise ratio) defined by Eq. 5 between input and separated speech, and word recognition rate of automatic speech recognition (ASR). As ASR, the Japanese dictation software, "Julius", is used, and as speech data, 20 sentences from the Mainichi Newspapers are used.

$$SNR = 10 \log_{10} \frac{\sum_n (s(n) - \beta s_o(n))^2}{\sum_n (s(n) - \beta s_s(n))^2} \qquad (5)$$

where, $s(n)$, $s_o(n)$, and $s_s(n)$ are the original signal, the signal observed by robot microphones and the signal separated by the active direction-pass filter, respectively. β is the attenuation ratio of amplitude between original and observed signals.

Experiment 1: The error of sound source localization of Sound, Face and Stereo Vision is measured. The results are shown in Fig. 5 when sound source direction is from 0° to 90°.

Experiment 2: Speeches from a loud speaker located of 0°, 30°, 60° and 90° are extracted by the active direction-pass filter. In this case, the direction of a loud speaker is given. When the pass range of the filter δ varies from ±5° to ±90°, Fig. 6 shows a comparison of the word recognition rate between observed signal and separated signal.

Experiment 3: The first loud speaker is fixed at 0°, the second one is located in 30°, 60° and 90° of *SIG*. Two speakers make sounds simultaneously. Speech from the first loud speaker is extracted

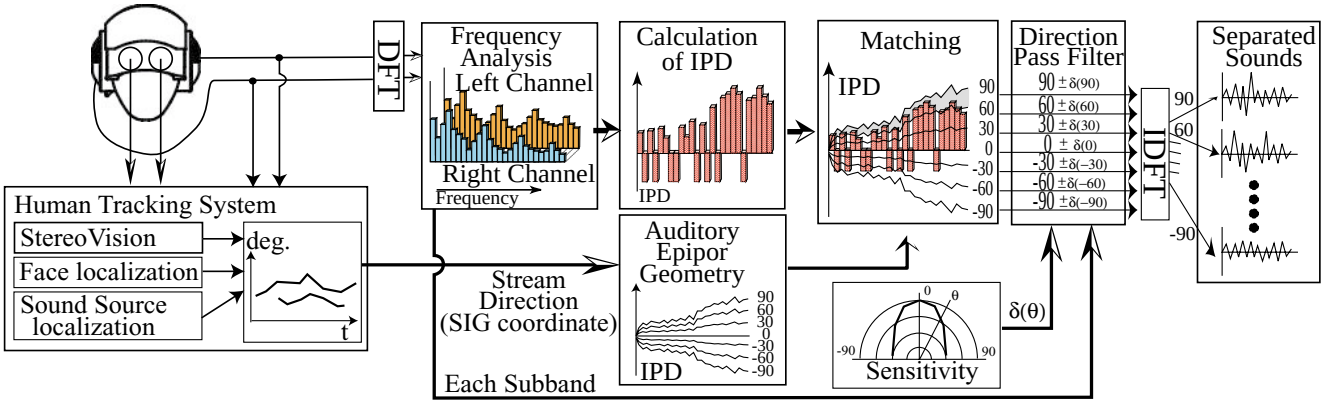

Fig. 4. Active Direction-Pass Filter

by the active direction-pass filter. The filter pass range function $\delta(\theta)$ obtained from *Experiment 1* is used. Fig. 7 shows the improvement of SNR by the active direction-pass filter.

Experiment 4: Two loud speakers are used. One is fixed in the direction of 60°. The other is moving from left to right repeatedly within the visual field of *SIG*. Speeches from the second loud speaker are extracted by the active direction-pass filter. Fig. 8 shows the improvement of SNR by using of stereo vision information.

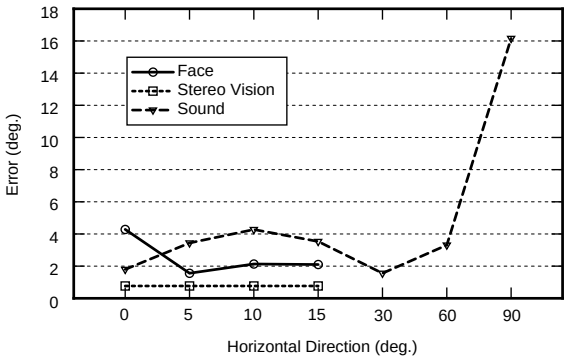

Fig. 5. Error of sound source localization

Fig. 5 shows that sound source localization by Stereo Vision is the most accurate. The error is within 1°. Generally, localization by vision is more accurate than by audition. However, Sound has the advantage of an omni-directional sensor. That is, Sound can estimate the direction of sound from more than ±15° of azimuth. The sensitivity of localization by Sound depends on sound source direction. It is the best in the front direction. The error is within ±5° from 0° to 30°, and it is getting worse at more than 30°. This proves that active motion such as turning to face a sound source improves sound source localization.

Fig. 6 shows that the front direction has a high sensitivity in sound source localization. For example, when δ is 20°, the difference of speech recognition rate between the front and the side direction is 50%. When a sound source is located at 60° and 90° from the front direction of *SIG*, the recognition rate is not good even if an optimal δ is used. This is caused by the *SIG* cover, i.e, the cover gives omni-directional microphones a directivity of the front direction. Facing the sound source improves sensitivity and SNR. The word recognition rate of separated sound increases 5−−10% in the direction of 0° and 30° in comparison with non-separated sound. This proves that the active direction-pass filter reduce environmental noise and improves the SNR.

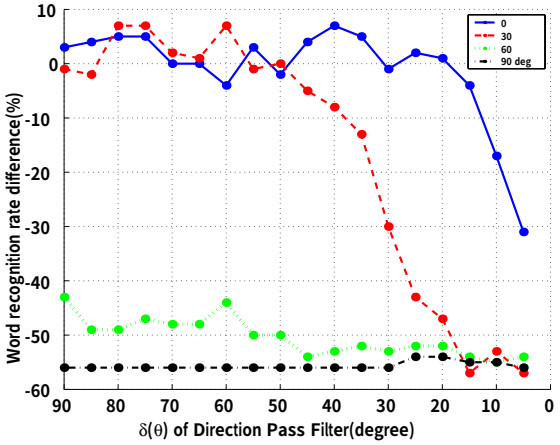

Fig. 6. Difference of speech recognition rate by direction

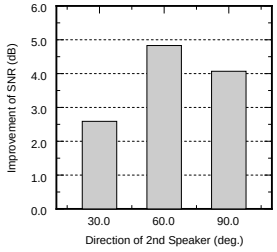

Fig. 7. Static speaker extraction

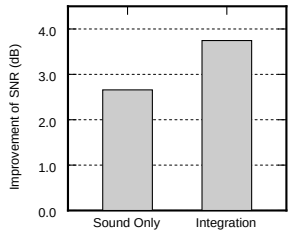

Fig. 8. Moving speaker extraction

Fig. 7 shows the sound source separation of two static speakers. It proves that the efficiency of the active direction-pass filter is $4 - -5dB$ when the angle between two speakers is more than $60°$, but separation of two speakers closer together than that is more difficult. For speech recognition, better sound source separation should be required because the result of the ASR is not good.

Fig. 8 shows that integration with visual information is not so effective, about 1dB improvement. This is because the sound stream is *manually* created. A "sound stream" consists of so many fragments that automatic stream formation failed. On the contrary, a stream by "integration" is *automatically* created by compensating such a gap in the sound stream with the aid of visual information.

V. Conclusion

This paper reports real-time sound source seperation by an active direction-pass filter as well as some improvements of our previous real-time multiple speaker tracking system. Robustness of sound source localization is improved by incorporating stereo vision, because it achieves more accurate localization even when only a partial face is available. By distributing communication load to Gigabit Ethernet and Fast Ethernet, computationl costs of *Stereo Vision*, which requires a lot of CPU power, does not affect the real-time processing.

The active direction-pass filter with adaptive sensitivity control is shown to be effective in improving sound source separation. The sensitivity of the direction-pass filter has not been reported so far in the literature and the idea of the active direction-pass filter resides in active motion to face a sound source to make the best use of the sensitivity. Since we use a conventional automatic speech recognition as it is, the recognition rate is not so good. However, we believe that the results reported in this paper should be used as the baseline performance for robust speech recognition. The combination of most up-to-date robust automatic speech recognition with the active direction-pass filter is one of exciting future work.

For the improvement of sound source separation, a more accurate direction-pass filter, integrated with other clues such as IID, is another future work. For a robust ASR, missing data such as masking signals by reverberation and environmental noise should be taken into account. A switch of acoustic and linguistic models by context extraction also would be necessary. Disambiguation of sound source localization and separation by hierarchical multi-modal integration, as humans do, would lead to a robust total perception system.

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

[3] R. Brooks, C. Breazeal, M. Marjanovie, B. Scassellati, and M. Williamson, "The cog project: Building a humanoid robot," in *Computation for metaphors, analogy, and agents*, C.L. Nehaniv, Ed. 1999, pp. 52–87, Spriver-Verlag.

[4] 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)*, 1999, pp. 1146–1151.

[5] 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)*. 1999, pp. 1723–1726, ESCA.

[6] M. Z. Ikram and D. R. Morgan, "A multiresolution approach to blind separation of speech signals in a reverberant environment," in *Proceedings of 2001 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2001)*. 2001, pp. 2757–2760, IEEE.

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

[8] 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.

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

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

[11] H. G. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano, "Human-robot interaction through real-time auditory and visual multiple-talker tracking," in *Proceedings of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2001)*. 2001, IEEE.

[12] 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.

[13] 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.

[14] 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*. 2000, pp. III–387–391, IEEE.

[15] Okada K. Inaba M. Inoue H. Kagami, S., "Real-time 3d optical flow generation system," in *Proc. of International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI'99)*, 1999, pp. 237–242.

[16] H.G. Okuno, K. Nakadai, T. Lourens, and H. Kitano, "Separating three simultaneous speeches with two microphones by integrating auditory and visual processing," in *Proceedings of European Conference on Speech Processing(Eurospeech 2001)*. 2001, ESCA.