Proceedings of the 2003 IEEE
International Conference on Robotics & Automation
Taipei, Taiwan, September 14-19, 2003

Robot Recognizes Three Simultaneous Speech By Active Audition

Kazuhiro Nakadai*, 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
nakadai@nakadai.com, okuno@nue.org, kitano@csl.sony.co.jp

Abstract—

Robots should listen to and recognize speeches with their own ears under noisy environments and simultaneous speeches to attain smooth communications with people in a real world. This paper presents three simultaneous speech recognition based on active audition which integrates audition with motion.

Our robot audition system consists of three modules – a real-time human tracking system, an active direction-pass filter (ADPF) and a speech recognition system using multiple acoustic models. The real-time human tracking system realizes robust and accurate sound source localization and tracking by audio-visual integration. The performance of localization shows that the resolution of the center of the robot is much higher than that of the peripheral. We call this phenomena "auditory fovea" because it is similar to visual fovea (high resolution in the center of human eye). Active motions such as being directed at the sound source improve localization because of making the best use of the auditory fovea. The ADPF realizes accurate and fast sound separation by using a pair of microphones. The ADPF separates sounds originating from the specified direction obtained by the real-time human tracking system. Because the performance of separation depends on the accuracy of localization, the extraction of sound from the front direction is more accurate than that of sound from the periphery. This means that the pass range of the ADPF should be narrower in the front direction than in the periphery. In other words, such active pass range control improves sound separation. The separated speech is recognized by the speech recognition using multiple acoustic models that integrates multiple results to output the result with the maximum likelihood. Active motions such as being directed at a sound source improve speech recognition because it realizes not only improvement of sound extraction but also easier integration of the results using face ID by face recognition.

The robot audition system improved by active audition is implemented on an upper-torso humanoid. The system attains localization, separation and recognition of three simultaneous speeches and the results proves the efficiency of active audition.

I. Introduction

Robots that interact with human should separate and recognize various kinds of sounds. This means that robot audition is important social interaction as well as trigger of an event. To realize such robots, four issues should be considered as follows: 1) noise cancellation while in motion, 2) information integration of audition, vision and other sensory information, 3) sound source separation under noisy environment, and 4) speech recognition of each sound source if it is speech. Because most robots consider them partially, robust and accurate auditory processing in robots has been difficult so far.

The difficulties in robot audition lie in sound source separation under real world environments. For example, *Kismet* of the MIT AI Lab [1] and *ROBITA* of Waseda University [2] can interact with people by automatic speech recognition and gestures, but they use a microphone attached near the mouth of each speaker to avoid motor noise in motion. Therefore, it does not have sound source separation function. *WA-2* of Waseda University[3] can localize a sound source by using a pair of microphones in the robot, but they do not take motor noises in motion into account. Therefore they adopt the "stop-hear-act" principle; that is, a robot stops to hear. They also assume a single sound source, so the robot does not have sound source separation function. *SmartHead* which can localize and track multiple sound sources by using four microphones and stereo cameras[4]. However, they use only low level information, and it is difficult to resolve ambiguity which is solved by higher level information such as face ID. Since they do not assume sound distortion by a robot's head shape, it is difficult to apply their method to a robot head with sound distortion such as human-like head. In addition, the maximum number of sound sources is limited theoretically.

To solve these problems, we proposed *active audition* to control microphone parameters to perceive auditory information better with cancellation of self motor noise[5]. The active audition is integrated with face localization and recognition and stereo vision by using streams, and real-time multiple human tracking system has been reported [6]. Furthermore, an active direction-pass filter (hereafter, ADPF) to separate sound sources by using accurate sound direction obtained from the real-time multiple human tracking system also has been reported[7]. The ADPF uses a pair of microphones to separate sound sources. It calculates interaural phase difference (IPD) and interaural intensity difference (IID) for each sub-band and then determines the sound source direction by performing hypothetical reasoning with a set of IPD's and IID's. Finally the ADPF collects sub-bands of which IPD and IID match those of the specified direction. The performance evaluation of the ADPF reveals that the sensitivity of localization depends on the direction of the sound source. In other

0-7803-7736-2/03/$17.00 ©2003 IEEE

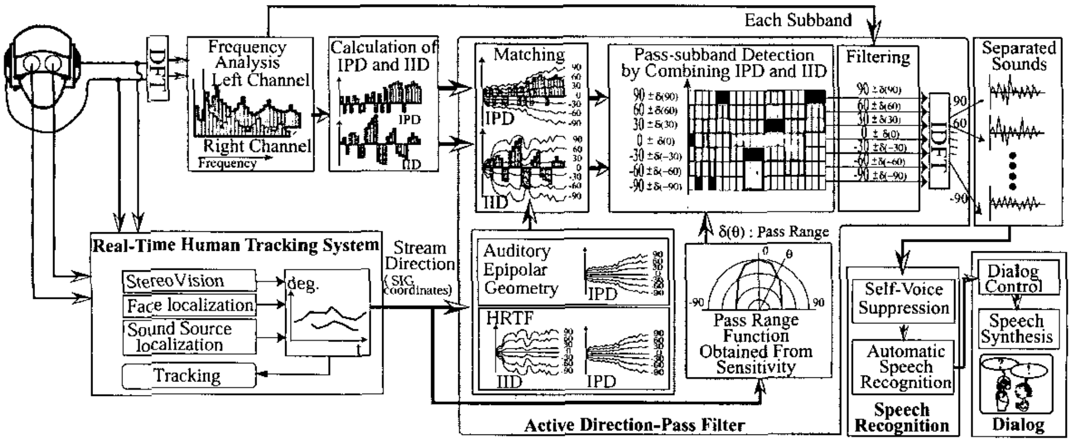


Fig. 1. The Robot Audition System for Simultaneous Speech Recognition

words, the ADPF separates sound streams very precisely when the sound source is just in front of the robot, while it separates sound streams very poorly when the sound source is sideways. Although this phenomenon occurs concerning a pair of microphone, it is quite similar to *fovea*, which means the difference of resolution in vision, that is, higher resolution in the center of eye while lower in the periphery. We call the auditory equivalent of fovea "*auditory fovea*".

As an application of the ADPF, we also reported automatic recognition of simultaneous speech by specific two persons[8]. However, the integration method of recognition results by multiple acoustic models is based on a simple majority rule.

In this paper, we propose an integration method based on recognition rate of each acoustic model. The method also provides audio-visual integration in speech recognition. We present the robot audition system that can localize, separate and recognize *three* simultaneous speech by using the real-time multiple human tracking system, the auditory fovea based ADPF, and the speech recognition using multiple acoustic models[1].

The rest of this paper is organized as follows: Section 2 describes robot audition system for simultaneous speech recognition. Section 3, 4 and 5 describe the real-time multiple human tracking system, the active direction-pass filter, and the speech recognition using multiple acoustic models, respectively. Section 6 evaluates the performance by the robot audition system. The last section provides discussion and conclusion.

II. Robot Audition System for Simultaneous Speech Recognition

The architecture of the robot audition system for simultaneous speech recognition is shown in Fig. 1. It consists of three modules – the real-time human tracking system, the active direction-pass filter and speech recognition by using multiple acoustic models. Sounds captured by robot's microphones and images captured by robot's cameras are sent to the real-time human tracking system described in later section. The sound source directions are obtained from auditory and visual streams generated in the real-time human tracking system. The sound source directions are sent to the ADPF. The ADPF extracts sound sources from the directions by hypothesis matching of interaural intensity difference (IID) and interaural phase difference (IPD) which are calculated from input spectra of left and right channels. The speech recognition module recognizes the extracted speeches by using multiple acoustic models.

We use the upper torso humanoid *SIG* as a testbed of the research. *SIG* has a cover by 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. 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 potentio-meters.

The following sections describe three modules in detail.

III. Real-Time Human Tracking System

The real-time human tracking system extracts accurate sound source directions by integration of audition and vision, and gives them to the ADPF. The architecture of the

[1] The series of our works on robot audition and audio-visual integration by Humanoid *SIG* are described in *http://www.symbio.jst.go.jp/SIG/*

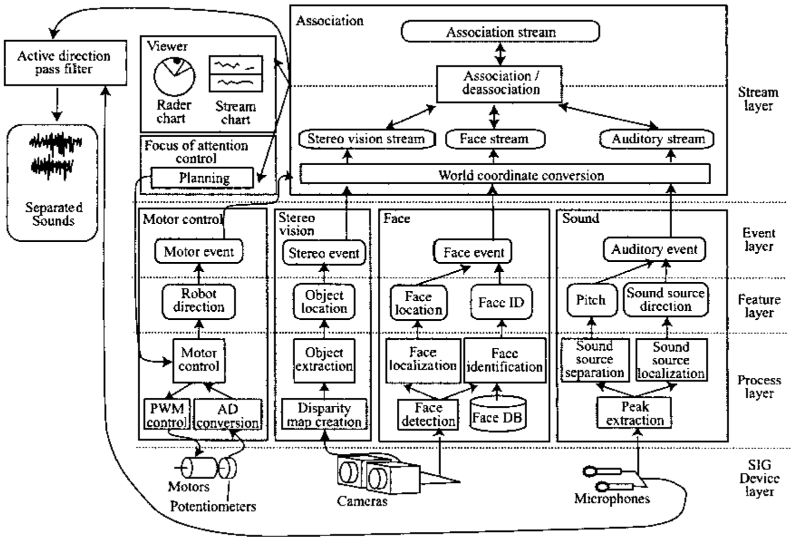


Fig. 2. Hierarchical Architecture of Real-Time Tracking System

real-time human tracking system using *SIG* shown in Figure 2 consists of seven modules, i.e., Sound, Face, Stereo Vision, Association, Focus-of-Attention, Motor Control and Viewer.

Sound localizes sound sources. Face detects multiple faces by combining skin-color extraction, correlation based matching and multiple scale image generation [9]. It identifies each face by Linear Discriminant Analysis (LDA), which creates an optimal subspace to distinguish classes and continuously update a subspace on demand with a small amount of computation [10]. In addition, the faces are localized in 3-D world coordinates by assuming average face size. Finally, the 10 best face IDs with probability P_v and their locations are sent to Association. Stereo Vision localizes lengthwise objects such as people precisely by using fast disparity map generation[11]. It improves the robustness of the system in point of tracking a person who looks away and does not talk. Association forms *streams* and associates them into a higher level representation, that is, an *association* stream according to the proximity in location. The directions of streams are sent to the ADPF with captured sounds. Focus-of-Attention plans *SIG*'s movement based on the status of streams. 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 association streams in the radar and scrolling windows. The whole system works in real-time with a small latency of 500 ms by distributed processing with 5 PCs and combination of Gigabit and Fast Ethernet.

Stream Formation and Association: Streams are formed in Association by connecting events from Sound, Face and Stereo Vision to a time course. First, since location information in sound, face, stereo vision events is observed in a *SIG* coordinate system, the coordinates is converted into world coordinates by comparing a motor event observed at the same time. The converted events are connected to a stream by using a Kalman filter based algorithm described in [12] in detail. Kalman filter is efficient to reduce the influence of process and measurement noise in localization, especially in auditory processing with bigger ambiguities. In sound stream formation, when a sound stream and an event have a harmonic relationship, and the difference in azimuth between the stream direction predicted by the Kalman filter and a sound event is less than $\pm 10°$, they are connected. In face and stereo vision stream formation, a face or a stereo stream event is connected to a face or a stereo vision stream when the distance difference between the predicted location of the stream and the location of the event is within 40 cm, and they have the same event ID. An event ID is a face name or an object ID generated in face or stereo vision module.

When the system judges that multiple streams originate from the identical person, they are associated into an association stream, higher level stream representation[6]. 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 de-associated to one or some separated streams.

Control of Tracking: The tracking is controlled by Focus-of-Attention to keep the direction of a stream with attention and sends motor events to Motor. By selecting a stream with attention and tracking it, the ADPF can continue to make the best use of foveal processing. The selection of streams, that is, focus-of-attention control is programmable accord-

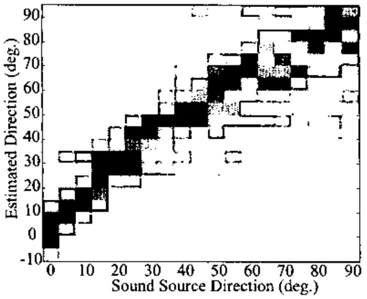


Fig. 3. Distribution of Sound localization

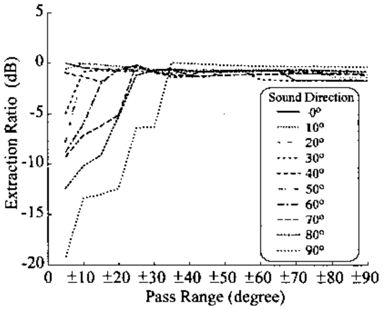


Fig. 4. Extraction of Single Sound Source

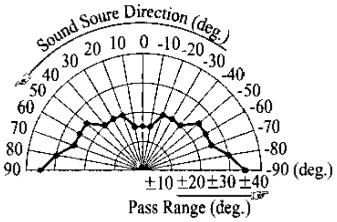


Fig. 5. Pass Range Function

ing to the surrounding situations. In this paper, the precedence of focus-of-attention control for the ADPF is used – an associated stream including a sound stream has the highest priority, a sound stream has the second priority, and other visual streams have the third priority.

IV. Active Direction Pass Filter

The architecture of the ADPF is shown in a dark area in Fig. 1. The APDF uses two key techniques, auditory epipolar geometry and the auditory fovea. The auditory epipolar geometry is a localization method by IPD and IID without using HRTFs. The auditory epipolar geometry is described in the next section in detail. In this paper, the ADPF is implemented to be able to use both of HRTFs and auditory epipolar geometry for evaluation. The auditory fovea uses to control pass range of the ADPF, that is, the pass range is narrow angles in front direction, and wider angles in the periphery. The detail algorithm of the ADPF is described as follows:

1. IPD $\Delta\varphi'$ and $IID\Delta\rho'$ in each sub-band are obtained by the difference between left and right channels.
2. Let θ_s azimuth of a stream with current attention in the robot coordinate system in the real-time human tracking system. The θ_s is sent to the ADPF through Gigabit Ether network by considering latency of processing.
3. The pass range $\delta(\theta_s)$ of the ADPF is selected according to θ_s. The pass range function δ has a minimum value in the *SIG* front direction, because it has maximum sensitivity. δ has a larger value at the peripheral because of lower sensitivity. Let us $\theta_l = \theta_s - \delta(\theta_s)$ and $\theta_h = \theta_s + \delta(\theta_s)$.
4. From a stream direction, the IPD $\Delta\varphi_E(\theta)$ is estimated for each sub-band by auditory epipolar geometry. The IID $\Delta\rho_H(\theta)$ is obtained from HRTFs.
5. The sub-bands are collected if the IPD and IID satisfy the following condition.
   $f < f_{th}: \quad \Delta\varphi_E(\theta_l) \leq \Delta\varphi' \leq \Delta\varphi_E(\theta_h)$, and
   $f \geq f_{th}: \quad \Delta\rho_H(\theta_l) \leq \Delta\rho' \leq \Delta\rho_H(\theta_h)$.
   The f_{th} is the upper boundary of frequency which is efficient for localization by IPD. It depends on the baseline of the ears. In *SIG*'s case, the f_{th} is 1500 Hz.
6. A wave consisting of collected sub-bands is constructed.

Note that the direction of an association stream is specified by visual information not by auditory one to obtain more accurate direction.

A. Auditory Fovea and Pass Range Control

In the retina, the resolution of images is high in the fovea, which is located at the center of the retina, and much poorer towards the periphery which serves to capture information from a much larger area. Because the visual fovea gives a good compromise between the resolution and field-of-view without the cost of processing a large amount of data, it is useful for robots [13], [14]. The visual fovea must face the target object to obtain good resolution, so it is a kind of *active vision* [15]. It is well-known that sound localization in human is the most accurate at the front direction, and is getting worse in the periphery[16]. It is true of a robot with two microphones. Fig. 3 shows distribution map of sound source localization in auditory module in real-time human tracking system.

The x axis means input sound directions from 0° to 90° at intervals of 10°. The 200 trials of localization are performed against each input sound direction. The result of localization is represented as a histgram in each input direction. The darker square means more concentration on localization. Figure 3 proves that localization of sounds from front direction is concentrated on a correct direction, and is widely distributed in the periphery. Thus, sound localization in a robot is accurate in the front direction. We call this phenomena auditory fovea.

Akin to the visual fovea, the auditory fovea also needs to be directed at the target object, such as a speaker. Therefore, it too relies on active motion. Such integration of sound and active motion, termed *active audition* [5], can be used to attain improved auditory perception. The active motion is essential in audition and vision not only for friendly humanoid-human interaction, but also for better perception.

The accuracy of sound source localization affects the performance of sound source separation by the ADPF. Because the accuracy of sound source localization depends on sound direction, pass range of the ADPF should be controlled according to the sound direction. Figure 4 shows results of single sound source extraction by the ADPF. The x and y axes are pass range of the ADPF and signal-to-noise ratio, respectively. When signal-to-noise ratio is 0dB, it is regarded that the sound source is extracted completely. Each line in Fig. 4 differs the direction of a speaker, and it is changed from 0° to 90° by 10°.

In case of sound from the front direction of the robot, the pass range of $\pm 10°$ is necessary to extract the sound properly. But, in case of 90° sound from the front direction of the robot, at least, the pass range of $\pm 35°$ is necessary. On single sound source, the wider pass range realizes the higher signal-to-noise ratio of sound extraction. However, background noise and other sound sources should be considered in real environment, so the narrower pass range is the better in a sense of noise cancellation. In Fig. 4, we select the narrowest pass ranges which extract a sound source properly and define the pass range function shown in Fig. 5.

B. Auditory Epipolar Geometry

Auditory Epipolar Geometry is proposed to extract directional information of sound sources without using HRTF [17]. The epipolar geometry is the fundamental geometric relationship between two perspective cameras in stereo vision research [18]. Auditory epipolar geometry is an extension of the epipolar geometry in vision (hereafter, *visual epipolar geometry*) to audition. Since auditory epipolar geometry extracts directional information by using the geometrical relation, it can dispense with HRTF. When the distance between a sound source and a robot is more than 50cm, the influence of the distance can be ignored [12]. Then, when the influence by a head shape is considered, the auditory epipolar geometry is defined by

$$\Delta\varphi = \frac{2\pi f}{v} \times r\,(\theta + \sin\theta) \qquad (1)$$

where f, v, r and θ are the frequency of sound, the velocity of sound, radius of a robot head and the sound direction, respectively. $\Delta\varphi$ is an estimated *IPD* corresponding to θ.

V. Speech Recognition for Separated Sound

Robust speech recognition against noises is one of the hottest topics in speech community. Some approaches such as multi-condition training and missing data [19], [20] shows efficiency in speech recognition with noise to some extent. However, these methods are of less use when signal-noise ratio is as low as 0dB. In this case, speech enhancement by a front-end processing is necessary. This kind of speech enhancement is efficient for speech recognition in higher signal-noise ratio, though such approach has not been studied so much. Then, we propose speech recognition using multiple acoustic models to use the sound source separation by the ADPF as front-end processing.

A. Acoustic Model

The Japanese automatic speech recognition software "Julian" is used for automatic speech recognition (ASR). For speech data, 150 words such as numbers, colors and fruits by 2 men (Mr. *A* and Mr. *C*) and 1 woman (Ms. *B*) are used.

For acoustic models, the words played by loud speakers of B&W Nautilus 805 are recorded by a pair of *SIG* microphones. They are installed in a 3m×3m room, the distance between *SIG* and a speaker is 1 m. The training datasets are created as follows:

1. 150 words by three persons are recorded by using robot microphones. The sound direction is -60°, 0°, or 60°. Two and three simultaneous speeches by combination of 0° and ±60° are also recorded. All patterns of combination of sound source directions and persons are recorded.
2. Speech of each direction is extracted from recorded data by the ADPF.
3. Separated speeches are clustered by person and direction to be a training dataset.

As a result, 9 training datasets are obtained as follows:

(a) a dataset of Mr.*A* from 0°
(b) a dataset of Mr.*A* from 60°
(c) a dataset of Mr.*A* from -60°
(d) a dataset of Ms.*B* from 0°
(e) a dataset of Ms.*B* from 60°
(f) a dataset of Ms.*B* from -60°
(g) a dataset of Mr.*C* from 0°
(h) a dataset of Mr.*C* from 60°
(i) a dataset of Mr.*C* from -60°

Nine acoustic models by Hidden Markov Model (HMM) are trained by the above training sets. Each HMM is an acoustic model by triphone and is trained 10 times by using Hidden Markov Model Toolkit (HTK).

B. Speech Recognition using Multiple Acoustic Models

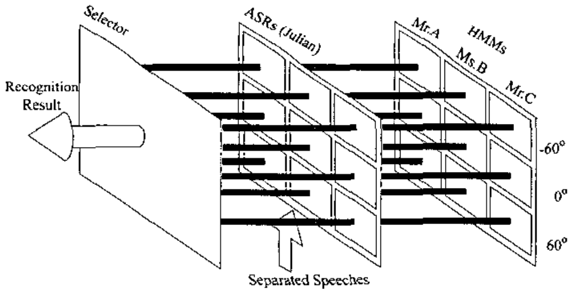
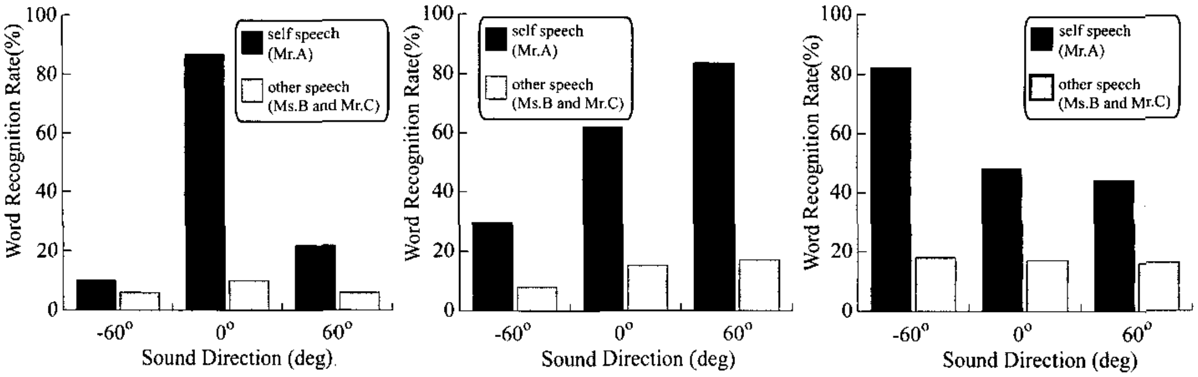


Fig. 7. Speech Recognition using Multiple Acoustic Models



a) acoustic model of Mr.*A* and 0° b) acoustic model of Mr.*A* and 60° c) acoustic model of Mr.*A* and -60°

Fig. 6. Recognition Results by Acoustic Models of Mr.*A*

In speech recognition, nine ASRs are processed against an input in parallel, and each ASR uses different acoustic model shown in Fig. 7. The selector integrates all results of ASRs, and outputs the most reliable result.

To define the algorithm of the integration, we checked word recognition rates by using Mr. *A*'s acoustic models. Figures 6a), b) and c) are the results by using Mr. *A*'s acoustic models of 0°, 60° and -60°, respectively. In Figs, the x axes are input sound directions, and the y axes are word recognition rates. The dark and light dark bars mean word recognition rates for self speech (Mr.*A*) and other speeches (Ms.*B* and Mr.*C*).

These results show that the influence by direction is less than by person, and the word recognition rate is more than 80% if both of the person and the direction are correct.

Because the system localizes sound sources by the real-time human tracking system before the separation by the ADPF, the directions of the sound sources are known. The selector uses a cost function by Eq. 2 to integrate the results.

$$V(p_e) = \left(\sum_d r(p_e,d)\cdot v(p_e,d) + \sum_p r(p,d_e)\cdot v(p,d_e) - r(p_e,d_e)\right)\cdot P_v(p_e). \quad (2)$$

$$v(p,d) = \begin{cases} 1 & if \quad Res(p,d) = Res(p_e,d_e), \\ 0 & if \quad Res(p,d) \neq Res(p_e,d_e). \end{cases}$$

where $r(p,d)$ and $Res(p,d)$ are recognition rate shown in Fig. 6 and recognition result against input speech when an acoustic model of person p and sound direction d is used. The d_e is the sound source direction estimated by the real-time tracking system, and the p_e is a person to be evaluated.

If an associated stream on a person exists, a probability $P_v(p_e)$ in the face recognition module is available. When a sound stream exists instead of the associated stream, $P_v(p_e)$ is set to 0.5 because face recognition is unavailable. Finally, the selector selects person p_e and result $Res(p_e,d_e)$ with the largest $V(p_e)$.

If the largest $V(p_e)$ is too small (less than 1) or close to the second largest one, *SIG* turns to the sound source and asks the person corresponding to the sound source again to make sure what he/she said.

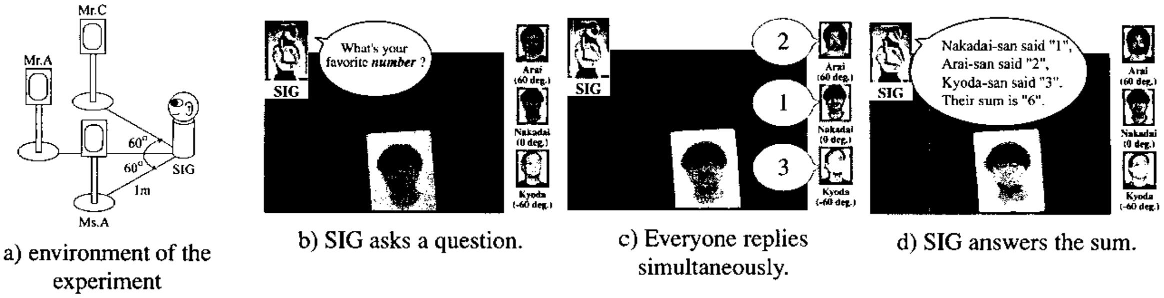
Thus, the system can recognize speeches and the name of the person by using multiple acoustic models. In addition, face recognition can improve the robustness of speech recognition if it is available.

VI. Evaluation

A whole system for dialog is constructed by combination of the ADPF, the real-time human tracking and the speech recognition shown in Figure 1. A simple function of self-voice suppression between the speech recognition and the ADPF is implemented by using information whether *SIG* is talking or not. A commercial speech synthesis is used for dialog. The system is evaluated through a scenario. The scenario is as follows:

1. *SIG* asks three persons(Mr.*A*, Ms.*B* and Mr.*C*) about the favorite number.
2. Each speaker selects any of 1 to 10, and they speak simultaneously.
3. *SIG* separates sound sources, and recognizes the separated speeches. *SIG* also identifies who spoke the number.
4. When *SIG* fails speech recognition or speaker identification, it turns to the speaker and asks the question again.
5. Finally, *SIG* replies the numbers with person names and their summation.

In this paper, we use three loud speakers of the B&W Nautilus 805 attached a photograph of a person instead of real persons. Figure 8a) shows the initial situation of the scenario. The speaker *A* attached Mr.*A*'s photograph is located


Mr.C
Mr.A
60°
60°
SIG
1m
Ms.A
What's your favorite *number* ?
SIG
Arai (60 deg.)
Nakadai (0 deg.)
Kyoda (-60 deg.)
2
1
3
Nakadai-san said "1", Arai-san said "2", Kyoda-san said "3". Their sum is "6".

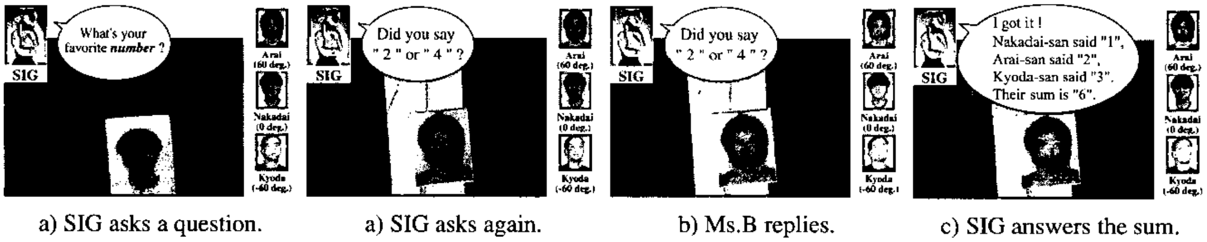
a) environment of the experiment

b) SIG asks a question.

c) Everyone replies simultaneously.

d) SIG answers the sum.

Fig. 8. Snapshots of Three Simultaneous Speech Recognition – I


What's your favorite *number* ?
SIG
Arai (60 deg.)
Nakadai (0 deg.)
Kyoda (-60 deg.)
Did you say " 2 " or " 4 " ?
Did you say " 2 " or " 4 " ?
I got it ! Nakadai-san said "1", Arai-san said "2", Kyoda-san said "3". Their sum is "6".

a) SIG asks a question.

a) SIG asks again.

b) Ms.B replies.

c) SIG answers the sum.

Fig. 9. Snapshots of Three Simultaneous Speech Recognition – II

at the front direction of *SIG*. The speaker *B* attached Ms.*B*'s photograph and the speaker *C* attached Mr.*C*'s are located at the 60° and -60°, respectively. Typical two results are shown as follows:

I. *SIG* succeeded speech recognition at a time.

1. *SIG* asks a question about the favorite number in Fig. 8b).
2. The speakers play three words simultaneously. Note that the combination of words is included in the training set in Fig. 8c).
3. *SIG* localizes the speakers by using the real-time tracking system. The ADPF uses the sound direction to separate sound sources. Each separated sound is recognized by the nine ASRs. The selector in multiple acoustic model recognizers takes a majority and decides the best result of each separated sound.
4. *SIG* answers the numbers with the names of persons, and summation of the numbers in Fig. 8d).

II. *SIG* fails speech recognition of the first time.

1. The initial situation is the same as Fig. 8a).
2. *SIG* asks a question about the favorite number in Fig. 9a). The speakers play numbers simultaneously. *SIG* separates mixture of speeches and recognizes them, but in this case, it cannot decide whether the speaker *B* played (Ms.*B* said) "2" or "4".
3. *SIG* turns to the speaker *B*, and asks her whether she said "2" or "4" in Fig. 9b).
4. She (speaker *B*) replies "2" in Fig. 9c). Because the speaker *B* is now at the front direction of *SIG*,
5. It turns to the speaker *A* again, and answers the numbers with the person names and their summation in Fig. 9d).

The observed results show the efficiency of sound source separation by the ADPF, speech recognition using multiple acoustic models. The sound source separation is efficient as a front-end of the speech recognition even when three simultaneous speeches occur. The integration by using multiple acoustic models realizes recognition of speech separated by the ADPF. The extraction of sounds from side direction is more difficult than that of the front direction. So, *SIG* sometimes fails speech recognition against input sounds from side direction. In such cases, *SIG* solve the problem by turning to the sound source and asks the question again such as the second observation. By turning to the sound source, it is easy to extract the sound because of the front direction. In addition, the system can use face recognition to improve speech recognition.

This proves that the auditory fovea based active audition improves speech recognition as well as sound source separation. Therefore active audition is essential to improve robot perception.

VII. Conclusion

The paper shows the robot audition system that localizes, separates and recognizes simultaneous speech. The system attains recognition of three simultaneous speeches by making the best use of auditory fovea based sound source separation by the ADPF and speech recognition by using multiple acoustic models. This proves that active audition using the auditory fovea and motor movement of the body is essential to improve robot audition. An expansion of the system in more natural environment such as use of speeches besides the training dataset and unknown speakers and evaluations of speech recognition of moving speakers are the future works.

ACKNOWLEDGEMENT

We thank to Prof. Tatsuya Kawahara, Kyoto Univ. and Mr. Hiroaki Nanjyo, Kyoto Univ. for their discussion on speech recognition. We also thank to Mr. Ken-ichi Hidai, Sony Corp., Mr. Daisuke Matsuura and Mr. Chihaya Okuno for their help.

REFERENCES

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

[2] 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 *Proc. of 6th European Conference on Speech Communication Technology(EUROSPEECH-99)*. 1999, pp. 1723–1726, ESCA.

[3] A. Takanishi, S. Masukawa, Y. Mori, and T. Ogawa, "Development of an anthropomorphic auditory robot that localizes a sound direction (*in japanese*)," *Bulletin of the Centre for Informatics*, vol. 20, pp. 24–32, 1995.

[4] S. Ando, "An autonomous three–dimensional vision sensor with ears," *IEICE Transactions on Information and Systems*, vol. E78–D, no. 12, pp. 1621–1629, 1995.

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

[6] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano, "Real-time active human tracking by hierarchical integration of audition and vision," in *Proc. of First IEEE-RAS International Conference on Humanoid Robots (Humanoids2001)*. 2001, pp. 91–98, IEEE.

[7] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano, "Real-time speaker localization and speech separation by audio-visual integration," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA 2002)*. 2002, pp. 1043–1049, IEEE.

[8] K. Nakadai, H. G. Okuno, and H. Kitano, "Auditory fovea based speech separation and its application to dialog system," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002)*. 2002, pp. 1314–1319, IEEE.

[9] K. Hidai, H. Mizoguchi, K. Hiraoka, M. Tanaka, T. Shigehara, and T. Mishima, "Robust face detection against brightness fluctuation and size variation," in *Proc. of IEEE/RAS Int. Conf. on Intelligent Robots and Systems (IROS-2000)*. 2000, pp. 1397–1384, IEEE.

[10] K. Hiraoka, S. Yoshizawa, K. Hidai, M. Hamahira, H. Mizoguchi, and T. Mishima, "Convergence analysis of online linear discriminant analysis," in *Proc. of IEEE/INNS/ENNS Int. Joint Conference on Neural Networks*. 2000, pp. III-387–391, IEEE.

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

[12] K. Nakadai, H. G. Okuno, and H. Kitano, "Exploiting auditory fovea in humanoid-human interaction," in *Proceedings of 18th National Conference on Artificial Intelligence (AAAI-2002)*. 2002, pp. 431–438, AAAI.

[13] W.N. Klarquist and A.C. Bovik, "Fovea: A foveated vergent active stereo vision system for dynamic 3-dimensional scene recovery," *RA*, vol. 14, no. 5, pp. 755–770, October 1998.

[14] S. Rougeaux and Y. Kuniyoshi, "Robust real-time tracking on an active vision head," in *Proc. of IEEE/RAS Int. Conf. on Intelligent Robots and Systems (IROS-97)*. 1997, pp. 873–879, IEEE.

[15] Y. Aloimonos, I. Weiss, and A. Bandyopadhyay., "Active vision," *International Journal of Computer Vision*, 1987.

[16] J. Blauert, *Spatial Hearing*, The MIT Press, 1999.

[17] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000)*. 2000, pp. 832–839, AAAI.

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

[19] J. Barker, M.Cooke, and P.Green, "Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. of 7th European Conference on Speech Communication Technology (EUROSPEECH-01)*. 2001, vol. 1, pp. 213–216, ESCA.

[20] Philippe Renevey, Rolf Vetter, and Jens Kraus, "Robust speech recognition using missing feature theory and vector quantization," in *Proc. of 7th European Conference on Speech Communication Technology (EUROSPEECH-01)*. 2001, vol. 2, pp. 1107–1110, ESCA.