REAL-TIME SOUND SOURCE LOCALIZATION AND SEPARATION FOR ROBOT AUDITION

Kazuhiro Nakadai [*], *Hiroshi G. Okuno* [*,†], *and Hiroaki Kitano* [*,‡]

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

ABSTRACT

Robot audition in the real world should cope with environment noises and reverberation and motor noises caused by the robot's own movements. This paper presents the *active direction-pass filter* (ADPF) to separate sounds originating from the specified direction with a pair of microphones. The ADPF is implemented by hierarchical integration of visual and auditory processing with hypothetical reasoning on interaural phase difference (IPD) and interaural intensity difference (IID) for each subband. In creating hypotheses, the reference data of IPD and IID is calculated by the *auditory epipolar geometry* on demand. Since the performance of the ADPF depends on the direction, the ADPF controls the direction by motor movement. The human tracking and sound source separation based on the ADPF is implemented on an upper-torso humanoid and runs in real-time with 4 PCs connected over Gigabit ethernet. The signal-to-noise ratio (SNR) of each sound separated by the ADPF from a mixture of two speeches with the same loudness is improved to about 10 dB from 0 dB.

1. INTRODUCTION

Robot audition can be improved by active motion and multi-modal integration as well as *active vision* [1]. As a technique for robust robot audition, *active audition* is proposed to control microphone parameters to attain better auditory perception [2]. The difficulty in active audition lies in sound source separation under real world environments. The technical issues include as follows:

1. cancellation of motor noise in motion,
2. auditory processing without using *Head Related Transfer Function (HRTF)* and measuring acoustic environments in advance, and
3. sound source separation techniques.

Robot's active motion makes inevitable motor noise, and it makes auditory processing more difficult. Therefore, robots with audition adopt the "stop-hear-act" principle, that is, a robot stops to hear. Otherwise it uses a microphone attached near the mouth of each speaker for automatic speech recognition [3, 4].

HRTFs are often used for sound source localization. Since, HRTFs are measured in an anechoic room, room characteristics should be measured in advance to use HRTFs. In addition HRTFs are available only at discrete points due to discrete measurement. Therefore, such sound source localization methods cannot be applied to a robot that changes its position by moving or rotation.

Common current sound source separation techniques, such as beamformers with a microphone array [5, 6], independent component analysis as blind source separation [7], and computational auditory scene analysis (CASA) techniques based on human auditory system [8], have not met the above requirements for robot audition yet. A real-time real-time sound localization and separation system based on the minimum-variance beamformer with eight microphones placed on a circle is developed for robot audition [5], which needs the database of location vector for the near-field measured in advance.

To cope with these technical issues, a new sound source separation method called the *active direction-pass filter (ADPF)* is proposed. The ADPF does sound source localization by *auditory epipolar geometry* without using HRTFs or the measurement of acoustic environments in advance [9]. The ADPF is implemented as a part of the real-time multiple speaker tracking system installed on an upper-torso humanoid called *SIG*. Motor noises and other noises caused by its movement are canceled by using cover acoustics. The tracking system attains accurate localization of multiple speakers by integrating face detection by stereo vision, multiple face recognition, and active motion control of *SIG* [10].

The rest of this paper is organized as follows: Section 2 describes our humanoid *SIG* and the real-time human tracking system. Section 3 explains sound source separation by the ADPF. Section 4 evaluates the performance of the ADPF. The last section provides discussion and conclusion.

2. THE REAL-TIME HUMAN TRACKING SYSTEM

We use the upper torso humanoid *SIG* shown in Fig. 1 as a platform 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 potentio-meters.

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

Sound, Face and Stereo Vision generate an *event* by feature extraction. Motor Control also generates an event of motion. Sound separates sound sources and localizes them. In addition, by feedback of streams in Association, it can extract a specific sound source by the ADPF. Face detects multiple faces by com-

Fig. 1. Humanoid *SIG*

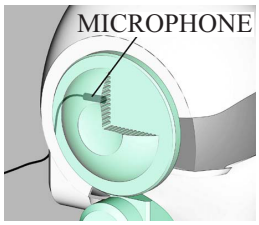

Fig. 2. *SIG* microphone

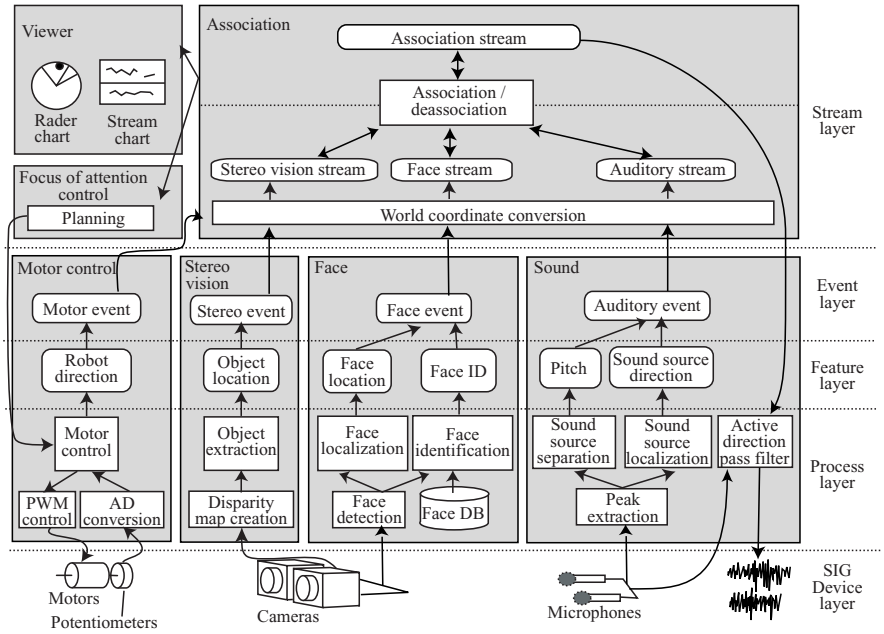

Fig. 3. Hierarchical Architecture of Real-Time Tracking System

bining skin-color extraction, correlation based matching, and multiple scale image generation [11]. 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 [12]. In addition, the faces are localized in 3-D world coordinates by assuming average face size. Stereo Vision localizes lengthwise objects such as people precisely by using a disparity map. Association forms *streams* as temporal sequences of these events and associates these streams into a higher level representation, that is, an *association* stream according to the proximity in location. Focus-of-Attention plans *SIG*'s movement based on the status of streams. The precedence is An association stream, a visual stream, and an auditory stream in decreasing order. 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 4 PCs and combination of Gigabit and Fast Ethernet.

3. SOUND SOURCE LOCALIZATION

Sound source localization in the system uses auditory epipolar geometry without using HRTFs. In addition, several sound clues are integrated to make it more robust. The rest of this section describes the flow of the sound source localization in detail.

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 subband selection, which a subband with a frequency between 90 Hz and 3 KHz is selected 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 reconstructed, and thus separated from a mixture of sounds.

Sound Source Localization by Integration: Sound source localization in the real world consists of 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.

For localization by IPD, a hypothesis of the IPD for each 5° candidate is generated by auditory epipolar geometry. 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 directions are estimated 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, an auditory event consisting of pitch ($F0$) and a list of 20-best directions (θ) with reliability factors and observation times for each harmonics is generated.

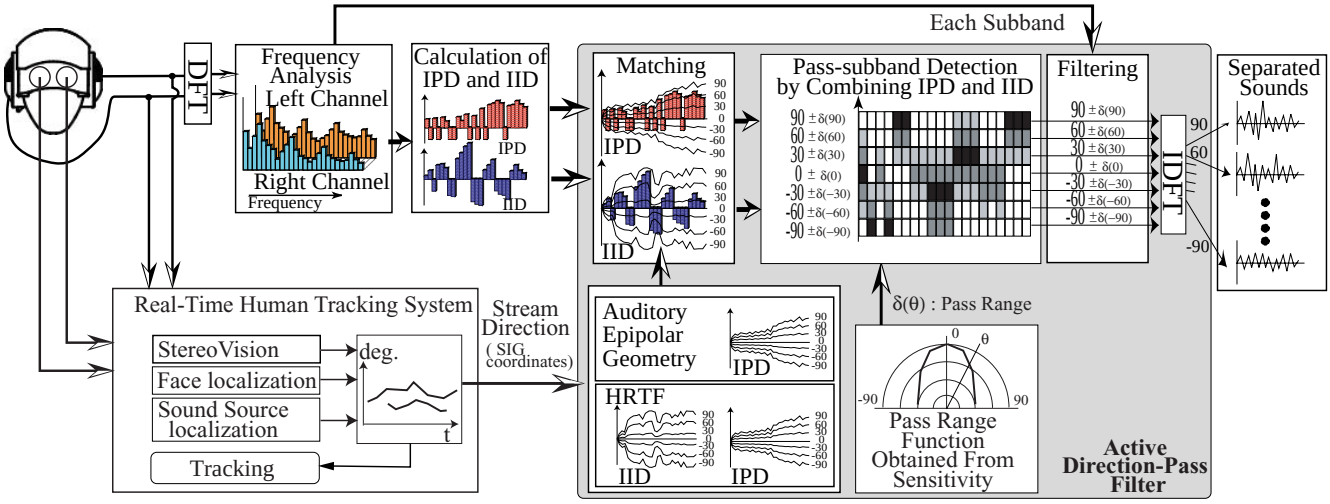

Fig. 4. Active Direction-Pass Filter

4. ACTIVE DIRECTION-PASS FILTER

To improve auditory processing in the real world, the ADPF based on auditory epipolar geometry, which is shown in Fig. 4, uses techniques such as calculation of IPD and IID without using HRTFs and control of pass range taken the sensitivity of the direction-pass filter into account. 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_s in the *SIG* coordinate system 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 side directions because of lower sensitivity. Let us $\theta_l = \theta_s - \delta(\theta_s)$ and $\theta_h = \theta_s + \delta(\theta_s)$.
4. The IPD $\Delta\varphi_E(\theta)$ and IID $\Delta\rho_E(\theta)$ are calculated for each sub-band by auditory epipolar geometry. Likewise, the IPD $\Delta\varphi_H(\theta)$ and IID $\Delta\rho_H(\theta)$ are obtained from HRTFs.
5. Peaks are extracted from the input, and IPD $\Delta\varphi'$ and IID $\Delta\rho'$ is calculated.
6. The sub-bands are collected if the IPD and IID satisfy the specified condition. Four kinds of conditions are used as follows:

   A: $f < f_{th}$: $\Delta\varphi_E(\theta_l) \le \Delta\varphi' \le \Delta\varphi_E(\theta_h)$,

   B: $f < f_{th}$: $\Delta\varphi_H(\theta_l) \le \Delta\varphi' \le \Delta\varphi_H(\theta_h)$, and
   $f \ge f_{th}$: $\Delta\rho_H(\theta_l) \le \Delta\rho' \le \Delta\rho_H(\theta_h)$

   C: $\Delta\varphi_E(\theta_l) \le \Delta\varphi' \le \Delta\varphi_E(\theta_h)$ for every frequency,

   D: $f < f_{th}$: $\Delta\varphi_E(\theta_l) \le \Delta\varphi' \le \Delta\varphi_E(\theta_h)$, and
   $f \ge f_{th}$: $\Delta\rho_H(\theta_l) \le \Delta\rho' \le \Delta\rho_H(\theta_h)$.

   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.
7. 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.

5. EXPERIMENTS

The performance of the ADPF is evaluated by two experiments. In these experiments, *SIG* and loud speakers are located in a room of 10 square meters. The distance and elevation between *SIG* and the speakers is about 80cm and -40°, respectively. The direction of a loud speaker is represented as 0° for *SIG* front direction.

A metric for evaluation is difference of SNR (signal-noise ratio) between input and separated speech defined by Eq. 2. For speech data, 20 sentences read by men and women from the Mainichi Newspapers are used.

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

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 ADPF, respectively. β is the attenuation ratio of amplitude between original and observed signals.

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

Experiment 2: The first loud speaker is fixed at 0°, the second one is located in 30°, 60° and 90° of *SIG*. Two loud speakers make different speeches with the same loudness simultaneously. Speech from the first loud speaker is extracted by the ADPF. The filter pass range function $\delta(\theta)$ is defined by localization errors obtained from *Experiment 1* is used. Fig. 6 shows the improvement of SNR by the ADPF by each filtering condition **A** to **D**.

Observation: Sound source localization by Stereo Vision is the most accurate shown in Fig. 5. 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,

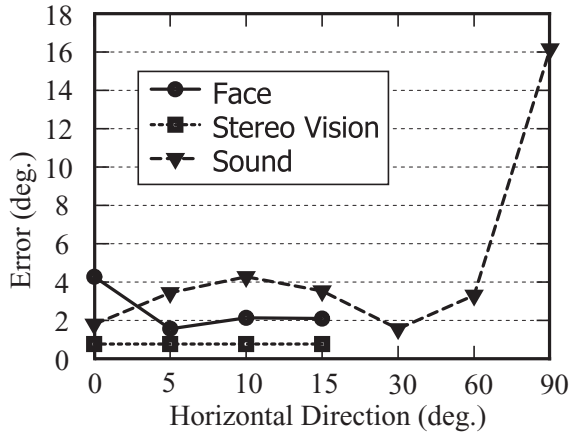

Fig. 5. Error of localization by Face, Stereo Vision and Sound

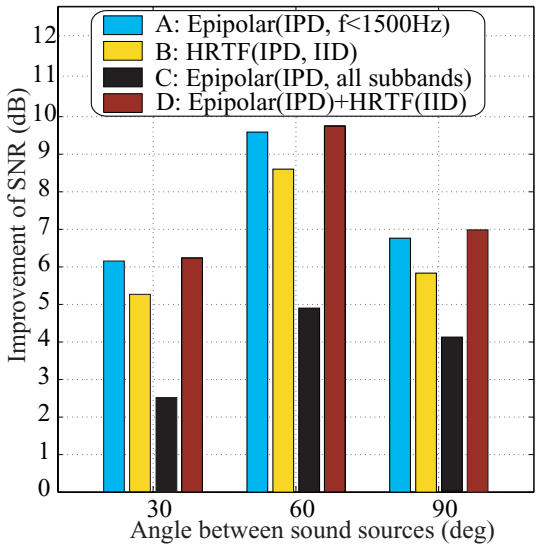

Fig. 6. Front speech extraction

Sound can estimate the direction of sound from more than $\pm 15^\circ$ 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 $\pm 5^\circ$ 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.

The filtering condition **D** has the best improvement of SNR regardless of direction between two sound sources in Fig. 6. This shows that the efficiency of the ADPF is $6 - 10dB$ in case of two speakers. But separation of two speakers closer together than 30° would be more difficult. The improvement by the filtering condition **B** based on HRTFs is lower than by the filtering condition **A** and **D** based on auditory epipolar geometry. This proves the efficiency of auditory epipolar geometry in sound source separation under real environments. The difference of improvement between the filtering condition **A** and **D** is small, because subbands with frequencies of more than 1500 Hz collected by IID have lower power. But, it is expected that the difference of recognition rate becomes bigger in case of automatic speech recognition (ASR), because ASR uses information from subbands with higher frequencies. In case of the filtering condition **C**, most subbands with more than 1500 Hz are collected because of a limitation of the baseline between the *SIG*'s ears. Therefore, the improvement of SNR is not so big.

6. CONCLUSION

This paper reports sound source separation by an ADPF connected with a real-time multiple speaker tracking system. The ADPF with adaptive sensitivity control is shown to be effective in improving sound source separation. The sensitivity of the ADPF has not been reported so far in the literature and the idea of the ADPF resides in active motion to face a sound source to make the best use of the sensitivity. The combination of most up-to-date robust automatic speech recognition with the ADPF filter is one of exciting future work.

7. REFERENCES

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

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

[3] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," *Proc. of IJCAI-99*, 1999, pp. 1146–1151.

[4] Y. Matsusaka, *et al.*,and T. Kobayashi, "Multi-person conversation via multi-modal interface — a robot who communicates with multi-user," *Proc. of EUROSPEECH-99*, 1999, pp. 1723–1726, ESCA.

[5] F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition.," *Proc. of EUROSPEECH 2001*, 2001, pp. 1013–1016, ESCA.

[6] H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Speech enhancement using nonlinear microphone array based on complementary beamforming.," *IEICE Trans. Fundamentals*, vol. E82-E, no. 8, pp. 1501–1510, Aug. 1999.

[7] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution.," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, Jun 1995.

[8] M. Mizumachi and M. Akagi, "Noise reduction by paired-microphones using spectral subtraction.," *Proc. of ICASSP-98*, 1998, pp. 1113–1116, IEEE.

[9] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," *Proc. of AAAI-2000*, 2000, pp. 832–839, AAAI.

[10] K. Nakadai, H.G. Okuno, and H. Kitano, "Real-time multiple speaker tracking by multi-modal integration for mobile robots.," *Proc. of EUROSPEECH 2001*, 2001, pp. 1193–1196, ESCA.

[11] K. Hidai, H. Mizoguchi, K. Hiraoka, M. Tanaka, T. Shigehara, and T. Mishima, "Robust face detection against brightness fluctuation and size variation," *Proc. of IROS-2000*, 2000, pp. 1397–1384, IEEE.

[12] K. Hiraoka, S. Yoshizawa, K. Hidai, M. Hamahira, H. Mizoguchi, and T. Mishima, "Convergence analysis of online linear discriminant analysis." *Proc. of IJCNN 2000*, III–387–391, 2000, IEEE.