Humanoid Active Audition System Improved by The Cover Acoustics

Kazuhiro Nakadai[1] and Hiroshi G. Okuno[2] and Hiroaki Kitano[3]

[1] Kitano Symbiotic Systems Project ERATO, Japan Science and Technology Corp.
Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya-ku, Tokyo 150-0001, Japan
Tel: +81-3-5468-1661, Fax: +81-3-5468-1664
[2] Department of Information Sciences, Science University of Tokyo
[3] Sony Computer Science Laboratories, Inc.

Abstract. Perception system for humanoid should be active, e.g., by moving its body or controlling parameters of sensors such as cameras or microphones, to perceive environments better. This paper focuses on active audition, whose main problem is to suppress internal sounds made by humanoid movements. Otherwise, such sounds would deteriorate the performance of auditory processing. Our 4-degree-of-freedom (DOF) humanoid, called *SIG*, has a cover to enclose internal sounds from the outside. *SIG* has a pair of left and right microphones to collect internal sounds and another pair to collect external sounds originating from the outside of *SIG*. A simple strategy of choosing a subband of external sounds if sounds from internal microphones in the same subband is weaker than those from external microphones sometimes fails in internal sound cancellation due to resonance within the cover. In this paper, we report the design of internal sound cancellation system to enhance external sounds. First, the acoustic characteristic of the humanoid cover is measured to make a model of each motor movement. Then, an adaptive filter is designed based on the model by taking movement commands into accounts. Experiments show that this cancellation system enhances external sounds and *SIG* can track and localize sound sources during its movement.

Keywords : robotics, cognitive modelling, active audition

1 Introduction

We have been studying humanoid to understand high-level perceptual functions and their multi-modal integration. We use an upper-torso humanoid called *SIG* as a platform of our research, because we believe that the integration of multi-modal sensory input and high degree-of-freedom (DOF) is essential for intelligence [10].

Recently active perception, i.e. the coupling of perception and behavior, has been studied using high DOF robots [3, 8, 9]. Most of such researches have been carried out as active vision [1]. Although it provides a framework for obtaining necessary information by controlling camera parameters, vision alone is not sufficient for some cases where occluded and/or out-of-sight objects exist.

On the other hand, in audition research, audition with behaviors, i.e. *active audition*, has not been studied yet even though people hear sounds while in motion. Indeed, some robotics researches notice the importance of auditory processing with motion, but they assume that the number of meaningful sound sources is at most 1 and the input sound is loud enough to ignore motor noises [19, 11]. These assumptions are too strong to understand high-level auditory functions. In addition, they also assume that auditory processing is done without motion. Therefore, active vision cannot be integrated with auditory processing.

As traditional auditory research attempts to understand psychological phenomenon such as the *cocktail party effect* [1], Computational Auditory Scene Analysis (CASA) studies a general framework of sound processing and understanding [4, 6, 16, 18]. Its goal is to understand an arbitrary sound mixture including speech, non-speech sounds, and music in various acoustic environment. However, most of these approaches still stay within the realm of audition research.

Therefore, active audition is expected to bring a major breakthrough One of main problems in active audition is to suppress internal sounds made by humanoid movements. Otherwise, such sounds would deteriorate the performance of auditory processing.

SIG is equipped with the cover to enclose internal sounds from the outside. However, a simple strategy of treating external sounds as internal noises if internal sounds are stronger than external sounds sometimes fails in internal sound cancellation due to resonance within the cover. Therefore, in this paper, internal noise cancellation system is designed by taking the acoustic characteristics of the cover into accounts.

The paper is organized as follows: Section 2 presents the active audition system. Section 3 presents acoustics of the humanoid cover. Section 4 proposes new sound source localization method by using acoustic measurements. Section 5 shows evaluation of our new localization method, and last two sections give discussion and conclusion.

2 Active Audition System

Fig. 1 shows the active audition system. The input of the system is assumed mixture sounds which come from different directional sound sources. The system consists of 5 modules; the humanoid *SIG* with 4 DOFs and 2 pairs of microphones, pre-processing, internal sound suppression, sound stream separation and motor control. The output is each separated sound source and tracking the specified sound source.

2.1 The Humanoid *SIG*

The mechanical structure of *SIG*is shown in Fig. 2(a). *SIG* has 4 DOFs of body driven by 4 DC motors, a pair of CCD cameras of Sony EVI-G20 as each eye, and

[1] A capability that people usually can separate sounds from the mixture and focus on a particular voice or sound even in a noisy environment.

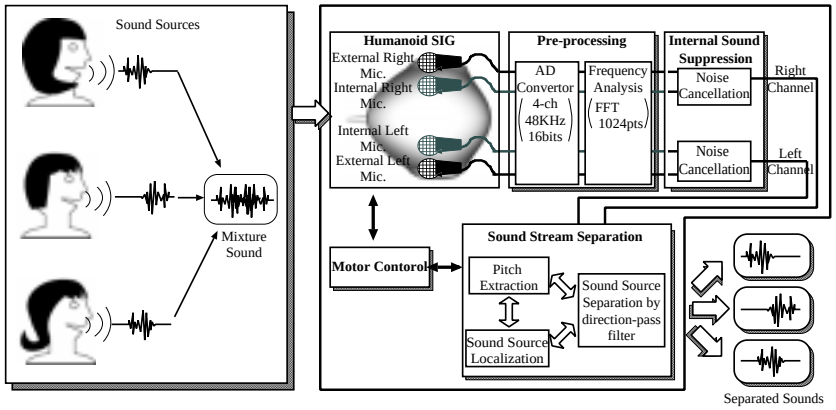
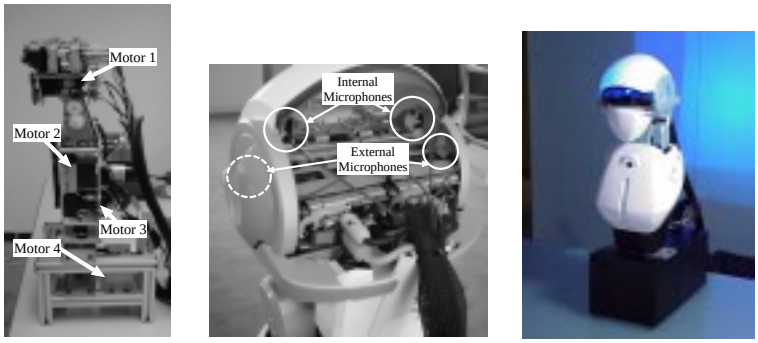


Fig. 1. Active Audition System

two pairs of omnidirectional microphones of Sony electret condenser microphone ECM-77S. Two pairs of microphones are used to separate the outer world from the inner world. One pair of microphones are installed at the ear position of the head to gather sounds from outer world. The other pair of microphones are installed very close to the corresponding microphones to gather sounds from inner world as shown in Fig. 2(b). And the cover is shown in Fig. 2(c). The cover not only separates the inner and outer world of *SIG*, but also has the beauty and functionality for humanoid exterior design [13].



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

Fig. 2. *SIG* the humanoid

2.2 Sound Pre-processing

The system carries out processing of AD conversion and frequency analysis against the input sounds in pre-processing module.

Sonorus AUDI/O is adopted as an AD converter in the system. It can process 48 KHz AD conversion of 4 channels (up to 8 channels) synchronously, i.e. mutual time differences between channels are kept. And it converts sampled 4-channel

sound into ADAT[2] signal. ADAT signal is captured by Sonorus STUDI/O (PCI sound card) through a optical fiber. The card is installed in a PC which has a Pentium III 600MHz CPU and 512M byte memory.

Then, by each channel, frequency analysis transform captured digital sounds into sound spectrograms on the PC. Fast Fourier Transformation (FFT) for 1,024 points is used for frequency analysis.

2.3 Internal Sound Suppression

In this module, motor noises are cancelled by applying a kind of adaptive filter. Because burst noises among motor noises have worse influences on the system, the filter is designed to cut off mainly burst noises by comparing external sounds with the corresponding internal sounds on sound spectrograms. It uses *heuristics*, which orders that localization by sound or direction-pass filter ignore a subband if the following conditions hold:

1. The power of internal sounds is much stronger than that of external sounds.
2. Twenty adjacent subbands have strong power.
3. A motor command is being processed.

The output is two channel; right and left external sounds, which are cancelled burst noises using the corresponding internal sounds.

2.4 Sound Stream Separation

This module consists of three sub-modules; sound source localization, pitch extraction, and sound source separation by a direction-pass filter.

Sound Source Localization Direction information of sound sources is extracted using *auditory epipolar geometry* [12]. Epipolar geometry is a popular localization method for stereo vision [7]. *Auditory epipolar geometry* expands the epipolar geometry in vision to auditory field as shown in Fig 3. This method extracted direction information without using *Head Related Transfer Function (HRTF)*. It is useful to localize sound source without using *HRTF* because *HRTF* is easy to change even if surrounded environments are changed a little, in other words, *HRTF* is hard to use for sound source localization in real environments.

It extracts peaks by using FFT for each subband, and calculates the *interaural phase difference (IPD)* as the difference between phases of right and left peaks. The bandwidth of each subband is 47Hz in our implementation. The sound source direction is estimated by Equation (1):

$$\cos\theta = \frac{v}{2\pi f b}\triangle\varphi \tag{1}$$

where v is the velocity of sound, b is the distance (baseline) between left and right microphones, $\Delta\varphi$ is IPD and f is the frequency of sound. For the moment,

[2] ADAT is a kind of digital format for multi-channel optical digital signals

the velocity of sound is fixed to 340m/sec and is invariant to the temperature and humidity. In *SIG*, the baselines for vision and audition are in parallel. Therefore, whenever sound source is localized by epipolar geometry in vision, it can be converted easily into the angle θ. This can apply to a method of integration visual and auditory information, and we reported the feasibility of such integration based on epipolar geometry [12].

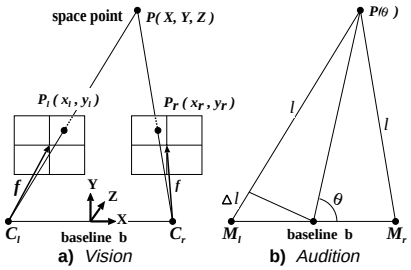


Fig. 3. Epipolar geometry for localization (C_l, C_r: camera center, M_l, M_r: microphone center)

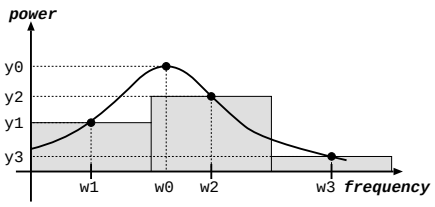


Fig. 4. A spectral peak by Fourier Transformation

Pitch Extraction Pitches are extracted by a kind of spectral subtraction [2]. It uses peak approximation method based on characteristics of FFT and window function. Consider that the peak $[\omega_2, y_2]$ is detected, and the values of both neighbors are $[\omega_1, y_1]$ and $[\omega_3, y_3]$ as shown in Fig. 4. Then, the true peak $[\omega_0, y_0]$ is estimated as follows:

$$\omega_0 = \begin{cases} \omega_2 + \dfrac{2\pi\,(2|y_1| - |y_2|)}{T\,(|y_1| + |y_2|)} & (\omega_1 < \omega_0 \leq \omega_2) \\ \omega_2 - \dfrac{2\pi\,(-|y_2| + 2|y_3|)}{T\,(|y_2| + |y_3|)} & (\omega_2 < \omega_0 < \omega_3) \end{cases} \tag{2}$$

$$\begin{aligned} Arg(y_0) &= \tan^{-1}\left(\frac{\Im[y_0]}{\Re[y_0]}\right) \\ &= \tan^{-1}\left(\frac{\Im[y_2]}{\Re[y_2]}\right) + \frac{T}{2}\,(\omega_2 - \omega_0) \end{aligned} \tag{3}$$

$$|y_0| = \frac{\Delta\omega\left(-T^2\Delta\omega^2 + 4\pi^2\right)}{2\pi^2 \sin\frac{T}{2}\Delta\omega}\,|y_2|\,,$$

$$\Delta\omega = \omega_2 - \omega_0 \tag{4}$$

ω_0 is estimated as the following Equation (2). And the phase and amplitude of the true peak y_0 are estimated as Equations (3) and (4), respectively. $\Re[x]$ and $\Im[x]$ are the real and imaginary part of a complex number x.

Because the above equations require relatively the small number of calculation, our method can run faster and extract more accurate pitches. For example, in comparison with Bi-HBSS [17], which is known as a sound source separation system using a pitch extraction method by spectral subtraction, our method needs only 1/200 of amount of calculation per a peak [14].

Sound Source Separation by Direction-pass Filter The direction-pass filter selects subbands that satisfies the *IPD* of the specified direction. The detailed algorithm is describes as follows:

1. The specified direction θ is converted to $\Delta\varphi$ for each subband (47 Hz).
2. Extract peaks and calculate *IPD*, $\Delta\varphi'$.
3. If *IPD* satisfies the specified condition, namely, $\Delta\varphi' = \Delta\varphi$, then collect the subband.
4. Construct a wave consisting of collected subbands.

2.5 Problem in Active Audition System

The system, however, has a problem that noise cancellation can not be sufficient because the internal microphones can capture louder sounds originating from the outer world than the external microphones.

We considered that the problem was caused by resonance inside the cover. We needed to measure acoustics of the cover to confirm it and to improve noise cancellation. Acoustic measurement is described as the next section.

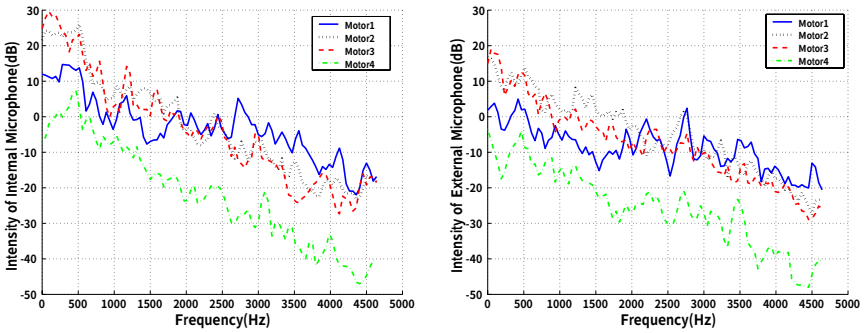
3 Acoustic Analysis of The Cover

The cover acoustics is measured in an anechoic room. The items of acoustic measurements are shown in the following.

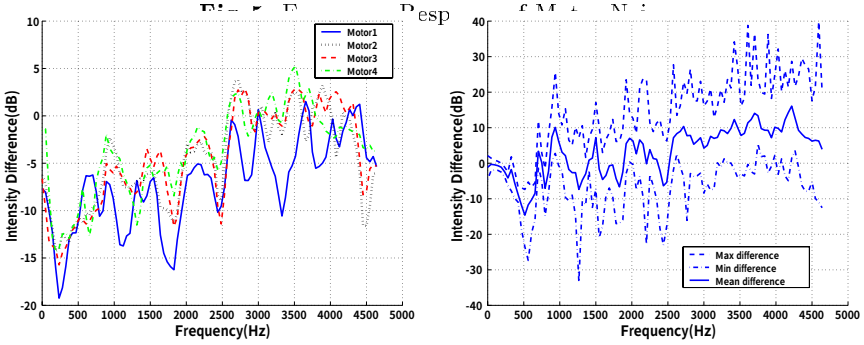
1. Frequency response of each motor noise with both internal and external microphones (Figs. 5(a) and (b)). Each motor moves from $-45°$ to $45°$ ($0°$ is the center of *SIG*) at the constant velocity (14.9 degree/sec). The noises of each motor are captured three times, and the averages are calculated.
2. Intensity difference between internal and external microphones. Fig. 6(a) shows intensity difference of each motor noise. The conditions of motors are the same as 1. The graph is estimated by subtracting internal microphone's frequency response from external one. Fig. 6(b) shows intensity difference of the outer sounds. This is estimated by impulse responses. The impulse responses are measured at 12 points which are elements of a matrix of horizontal and vertical directions; horizontal directions (azimuths) are $0°$, $\pm 45°$, $\pm 90°$ and $180°$ from robot center and vertical directions (elevations) are $0°$ and $30°$.

From the figures, main observations are summarized as follows:

1. Motor noise is broadband and is captured less than 30 dB by internal microphones, is captured less than 20 dB by external ones as shown in Figs. 5(a) and (b).
2. Motor noise is captured louder by external microphones than by internal microphones for frequencies of more than 2.5 KHz as shown in Fig. 6(a). This shows that the cover makes it easier to capture motor noise by internal microphones, because sounds from outer world is cut off by the cover.



(a) internal microphone (b) external microphone



(a) motor noise (b) outer sounds

Fig. 6. Intensity Difference

3. Acoustic signals are often captured louder by internal microphones than by external microphones for frequencies of less than 2 KHz. Especially, the tendency is more remarkable for frequencies of less than 700Hz as shown in Fig. 6(b). This shows resonance within the cover. The diameter of the cover is about 18 cm, which is corresponded to $\lambda/4$ at frequency of 500Hz. This causes resonance which has 500 Hz of center frequency. The similar resonance is occurred in Fig. 6(a).
4. Internal sound is captured about 10 dB louder than external sound on average by comparing Fig. 6(a) and Fig. 6(b). Therefore, the cover efficiency to separate the inner and outer sounds is about 10 dB.

4 New Noise Cancellation Method

We revise a noise canceling method using the acoustics. First, we store the data of the acoustic measurement in the system. The noise data of each motor is stored as a power spectrum of the averaged measured noises. Next, we use the stored data as templates to judge burst noises. When the motor makes a burst noise, the intensity of the noise is quite stronger because microphones location is relatively near the motor. Therefore if the spectrum and intensity of captured noise is similar to those of a noise template, the captured noise is regarded as a

burst noise. Specifically, the subband is cancelled if the following conditions are satisfied:

1. Intensity difference between external and internal microphones is similar to measured motor noise intensity differences.
2. Intensity and pattern of the spectrum are similar to measured motor noise frequency responses
3. A motor command is being processed.

5 Experiments

In this section, we demonstrate the effectiveness of noise cancellation by the new method in sound source localization.

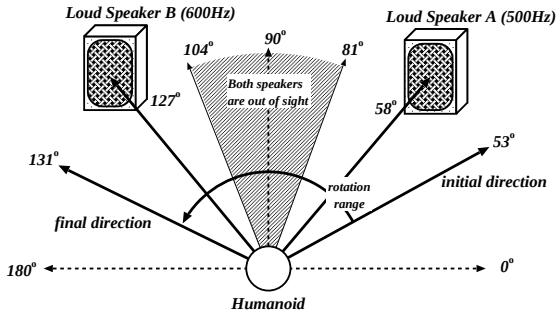


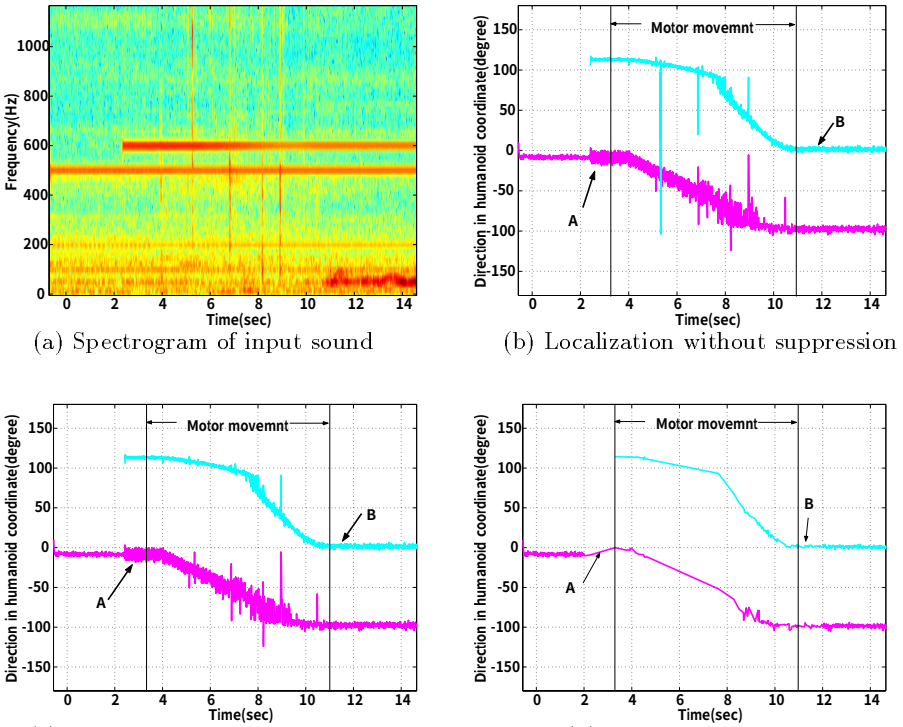
Fig. 7. Experiment: Sound source localization while *SIG* moves.

There are two sound sources: two B&W Noutilus 805 loud speakers located in a room of 12 square meters. The system is installed in a conventional residential apartment facing a road with busy traffic, and exposed to various daily life noise. Sound environment is not at controlled for experiment to ensure feasibility of the approach in daily life.

One sound source A (Loud Speaker A) plays a pure tone of 500 Hz. The other sound source B (Loud Speaker B) plays that of 600 Hz. A and B are located in front of *SIG*. *SIG* turns toward the direction of the sound source B at the velocity of 14.9 degree/sec using the direction obtained by audition under the condition that both A and B make sounds. Fig. 7 shows this situation.

Fig. 8(a) shows the captured sound spectrogram, Figs. 8(b), (c) and (d) show the localization results. The Y axis of each graph describes direction of A and B in humanoid coordinate system. Figs. 8(b), (c) and (d) show the results of sound source localization without noise cancellation, with noise cancellation by the previous method, and with our new noise cancellation, respectively.

Fig. 8(a) depicts 4 burst noises at 5.5, 7.0, 8.1 and 9.0 seconds. Fig. 8(b) also shows that sound source localization is badly impaired. Using our previous method, burst noises at 5.5 and 7.0 seconds are cancelled or weakened as shown in Fig. 8(c), but other noises still remain. Fig. 8(d) shows that our new method cancels all burst noises and suppresses vibration.



(a) Spectrogram of input sound (b) Localization without suppression

(c) localization by suppression method (d) localization by new method

Fig. 8. Localization Experiments of sound sources

However, location error of $\pm 10°$ can be observed; when the robot has rotated by $80°$, the actual sound source is located at the angle of $100°$. This *IPD* error may be caused by the mismatch between FFT window length and wave length, and by ambiguity of peak position due to discretization of FFT. And we demonstrate the noise cancellation in case of using a motor and the constant velocity of the motor. Our noise cancellation method should be extended to the variable velocities of motors.

6 Discussion

We propose *auditory epipolar geometry*, which provides a sound source localization method without using *Head Related Transfer Function (HRTF)*. In real world, localization without using *HRTF* is needed because *HRTF* depends on environments. And the method can be easily used for integrating auditory and visual information as shown in section 2.4. Though the experiments demonstrate effective noise cancellation method using noise database in real-world environments, the system still has about $\pm 10°$ error in sound source localization. It may be difficult to solve the problem only using auditory information because it is said that even human auditory capability has $\pm 8°$ error in sound source localization [5].

Therefore, other sensory information such as vision and tactile information is required to compensate the error. The integration of vision and audition was done by Nakagawa *et al* [15]. However, their system fails in sound source localization and separation in real environments because it works only in simulated environments and using *HRTF*. We have already demonstrated the feasibility of the integration of audition, vision, and motor information in real-world environments [12]. This performance will be improved by incorporating the proposed method. Real time processing of active audition is critical in real world applications. To calculate *IPD* by *auditory epipolar geometry* is not difficult to speed up, but an *IPD* error described in Sec. 5 needs more sophisticated theoretical treatment. Other future work includes incorporating various acoustic features such as harmonics, onset, offset, common amplitude modulation, common frequency modulation, formants, timbre, and so on.

7 Conclusion

We discuss the importance of this research with respect to active audition since it has not been studied so far. We also discuss that the cover is important for active audition. By analyzing the acoustics of the cover, we demonstrate the effectiveness of noise cancellation method which improves sound source localization even while the humanoid is moving. In addition, we show that *auditory epipolar geometry* method without using *HRTF* plays an important role of sound source localization in real environments. Because this method can be easily expanded to combine visual information, it can be a useful method not only for active audition but also for active perception which integrates various sensory information. Active perception is important for the integration of perceptual information as well as to understand fundamental principles of intelligence.

Acknowledgments

We thank *NITTOBO Acoustic Engineering Co., Ltd.* for the acoustic measurements and offering the anechoic room. We thank our colleagues of Symbiotic Intelligence Group, Kitano Symbiotic Systems Project; Dr. Theo Sabish, Dr. Tino Lourence, Yukiko Nakagawa and Dr. Iris Fermin for their discussion.

References

1. Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1987.
2. S. F. Boll. A spectral subtraction algorithm for suppression of acoustic noise in speech. In *Proceedings of 1979 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79)*, pages 200–203. IEEE, 1979.
3. R. Brooks, C. Breazeal, M. Marjanovie, B. Scassellati, and M. Williamson. The cog project: Building a humanoid robot. Technical report, MIT, 1999.
4. G. J. Brown. *Computational auditory scene analysis: A representational approach.* PhD thesis, Dept. of Computer Science, University of Sheffield, 1992.

5. J. Cavaco, S. ad Hallam. A biologically plausible acoustic azimuth estimation system. In *Proceedings of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99)*, pages 78–87. IJCAI, Aug. 1999.
6. M. P. Cooke, G. J. Brown, M. Crawford, and P. Green. Computational auditory scene analysis: Listening to several things at once. *Endeavour*, 17(4):186–190, 1993.
7. O. D. Faugeras. *Three Dimensional Computer Vision: A Geometric Viewpoint*. The MIT Press, MA., 1993.
8. M. Kawato. Bi-directional theory approach to consciousness. In *Cognition, Computation, and Consciousness*. Oxford University Press, 1996.
9. N. Kita, S. Rougeaux, Y. Kuniyoshi, and S. Sakane. Real-time binocular tracking based on virtual horopter. *Journal of Robotics Society Japan*, 13(5):101–108, 1995.
10. H. Kitano, H. G. Okuno, K. Nakadai, I. Fermin, T. Sabish, Y. Nakagawa, and T. Matsui. Designing a humanoid head for robocup challenge. In *Proceedings of 4th Internatianal Conference on Autonomous Agents (Agents 2000)*. ACM, 2000.
11. 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*, pages 1723–1726. ESCA, 1999.
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)*. AAAI, 2000. (*to appear*).
13. K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano. Active audition system and humanoid exterior design. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS 2000)*. IEEE, 2000. (*accepted*).
14. K. Nakadai, H. G. Okuno, and H. Kitano. A method of peak extraction and its evaluation for humanoid. In *SIG-Challenge-99-7*, pages 53–60. JSAI, 1999.
15. Y. Nakagawa, H. G. Okuno, and H. Kitano. Using vision to improve sound source separation. In *Proceedings of 16th National Conference on Artificial Intelligence (AAAI-99)*, pages 768–775. AAAI, 1999.
16. T. Nakatani, H. G. Okuno, and T. Kawabata. Auditory stream segregation in auditory scene analysis with a multi-agent system. In *Proceedings of 12th National Conference on Artificial Intelligence (AAAI-94)*, pages 100–107. AAAI, 1994.
17. 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)*, volume 1, pages 165–172. AAAI, 1995.
18. D. Rosenthal and H. G. Okuno, editors. *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, NJ., 1998.
19. 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*, 20:24–32, 1995.