Integrating Auditory and Visual Perception for Robotic Soccer Players

Hiroshi G. Okuno [†*], Yukiko Nakagawa [†], and Hiroaki Kitano [†‡]

†Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.
Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya-ku, Tokyo 150-0001, Japan
∗ Department of Information Sciences, Science University of Tokyo
‡Sony Computer Science Laboratories, Inc.
okuno@nue.org, yuki@symbio.jst.go.jp, kitano@csl.sony.co.jp

ABSTRACT

In this paper, we present a method of integrating auditory and visual perception for mobile robot for RoboCup. When humanoid robots are fielded for soccer game, they need to quickly react to the environment using all possible sensory inputs. While current robots heavily depends on visual inputs, auditory inputs actually play significant role in detecting events where visual inputs are not available, such as side and behind the face direction. Sound of other players kicking the ball, and shouting of teammates are critical cues for sophisticated teamwork play, such as offside trap. This paper presents integration of auditory and visual perception for identifying sound sources and separating sounds at high accuracy using both auditory and visual inputs.

1 INTRODUCTION

Sound plays important roles in real-world soccer games, even if a soccer field is extremely noisy. It is often the case that communication by voice is critical in key plays: At a hero interview after a game, he/she said "Since I heard my teammate's voice requesting my pass, I made a pass to an open space to which, I supposed, he would run and he made a goal." However, sound or voice has not been utilized in mobile robots. In RoboCup real robot league, small size or middle size, vision and tactile information are fully exploited, but auditory information is not used.

Sound is gathering attention as important media for multi-modal communications, but is less utilized as input media than characters or images. One reason is the lack of a general approach to recognize auditory events from a mixture of sounds. Usually, people hear a mixture of sounds, not a single sound. People with normal hearing can separate sounds from the mixture and focus on a particular voice or sound in a noisy environment. This capability is known as the *cocktail party effect* [1].

Perceptual separation of sounds, called *auditory scene analysis*, has been studied by psychoacoustic and psychophysical researchers for more than forty years. Although many observations have been analyzed and reported [2], it is only recently that researchers have begun to use computer modeling of auditory scene analysis.

This emerging research area is called *computational auditory scene analysis (CASA)* [3, 4, 5, 6], and its goal is the understanding of an arbitrary sound mixture including non-speech sounds and music. Computers need to be able to decide which parts of a mixed acoustic signal are relevant to a particular purpose – which part should be interpreted as speech, for example, and which should be interpreted as a door closing, an air conditioner humming, or another person interrupting. CASA focuses on the computer modeling and implementation for the understanding of acoustic events.

To recognize scene around us, we must be able to identify which set of perceptive input (sounds, pixels, etc) constitutes an object or an event. To understand what is in the visual scene, people (or a machine) should be able to distinguish a set of pixels which constitutes a specific object from those that are not a part of it. In Computational Auditory Scene Analysis, sound shall be separated into auditory streams each of which corresponds to specific auditory event [2, 4, 5, 6].

In this paper, we argue that integration of auditory and visual perception is significant in intelligent robotics communities, report the current status of the research and presents the research issues.

2 WHY IS SOUND NEEDED?

In this section, we argue why sound is needed in the context of RoboCup, in particular, simulator League, real robot league, and rescue.

RoboCup Simulator League Since simulator league tries to capture every aspect of human highest-level soccer, it models auditory and visual sensory information. RoboCup simulator league consists of soccer server and 22 soccer clients (players). Every communication, even between teammates, should be done via soccer server.

(1) Vision: Soccer server gives visual information to each client as a message of `see`.

(2) Audition: A player can try to send a message to other players by submitting a command `say`.

A message originated from a player is delivered to all teammates that exist within the range of voice communication by the form of `hear`. A message from the referee is delivered to all the players.

Soccer server also supports communication between coach and teammates. All messages from the coach is delivered to all teammates, since the purpose of coach is to control a training session.

RoboCup Real Robot Leagues In RoboCup small-size league and middle-size league, vision is the primary sensory information, with tactile or sonar being used as a supplementary sensory information. Sonar uses sounds, but is not considered as auditory sensory systems, because it does not hear environmental sounds.

Since a robot player moves, visual and auditory sensory systems are quite difficult to process. Since the soccer field is flat, mobility in RoboCup causes local cameras equipped with a robot moves horizontally. Therefore, adjustment of visual information is needed. This is the same for auditory sensory systems.

Additional difficulties are introduced by the mobility. When a robot moves, it is driven by motors and driving mechanism makes sounds. The auditory sensory systems should discriminate driving sounds it makes from those that other robots make.

Some simple ideas to use auditory sensory systems may be listed as follows:

(1) When a robot player hears the whistle human referee blows, it stops the game and gets ready for restart by itself.

(2) Human coach instructs a robot player or all robot players by a simple voice command. Auditory communication gives an additional communication channel between the controller and robots. This is important, because wireless communication systems sometimes failed in transmitting a command to a robot in big RoboCup matches.

RoboCup Humanoid Challenge The ultimate goal of RoboCup Initiative is proposed as follows [7]:

> By 2050, a team of fully autonomous humanoid robot soccer players shall win the soccer game, comply with the official rule of the FIFA, against the winner of the most recent World Cup.

Kitano and Asada insisted the importance of sensory systems, and presented the research issues from the following aspects:

(1) Vision — 3-D representation and real-time processing,

(2) Auditory Systems — Speech understanding and computational auditory scene analysis,

(3) Other Sensing Systems — tactile systems such as touch, and force/torque

(4) Sensor Fusion — robust multi-sonsory systems and real-time processing,

(5) Sensory-Motor Integration — concept formation in associating sensory inputs and motor commands.

At the AAAI-96, the panel entitled "Challenge Problems for Artificial Intelligence', Brooks proposed two problems concerning sounds [8]:

- Challenge 1: Speech understanding systems that are based on different principles other than hidden Markov models.
- Challenge 2: Noise understanding systems.

Although CASA shares the above interests, its ultimate goals go further; understanding general acoustic signals such as voiced speech, music and/or other sounds from real-world environments.

Speech enhancement is essential to enable automatic speech recognition to work in such environments. Conventional approaches to speech enhancement are classified as noise reduction, speaker adaptation, and other robustness techniques [9]. Speech stream separation is a novel approach to speech enhancement, and works as the front-end system for automatic speech recognition just as hearing aids for hearing impaired people.

Okuno *et al.* proposed the problem of *Understanding Three Simultaneous Speeches* as a challenge problem for Artificial Intelligence, in particular, for CASA [10]. Since psychoacoustic studies have recently showed that human cannot listen to more than two things simultaneously [11], CASA research would make computer audition more powerful than human audition.

RoboCup Rescue Challenge RoboCup Rescue is proposed as the second domain of RoboCup Initiative [12]. Disaster rescue is one of the most serious social challenges, because it involves a huge number of heterogeneous agents in dangerous environments. Kitano *et al.* argue that RoboCup Rescue is very important in large scale multi-agent domains and focus on search and rescue strategies. They do not discuss on sensory systems explicitly, but apparently new sensory systems should be designed and developed. Sense of smell or auditory systems is an important candidate to allocate victims under collaped houses by searching faint sounds they create.

In addition to new sensory systems, integration of sensory systems and motor controls is essential. When a robot enters a dameged house and tries to remove debris under which victims are supposed to be buried, it hears a strange noise made by his behavior. In this case, the robot should identify the cause of the noise and infer what will happen if it continues his jobs. If it infers that the house will be destroyed, it should stop his jobs immediately. Or if the sound is victims' voice, he may try to communicate with them.

Needless to say, research issues presented in RoboCup Humanoid Challenge also apply Rescue Challenge [7].

3 WHY IS INTEGRATION NEEDED?

In real soccer games, players shout to communicate with other players on their position, demands, alerts, opponent's moves, etc. A simple shouting is often used to notify player's position when the player is outside of the visual field of the other player. In this case, quick identification of direction of sound source is essential. In other cases,

a player may shout to alert the opponent's movement and demend for specific actions such as pass the ball, so that understanding what was said is critical. These cases take place not as an isolated action, but they could happen all at once, hence sounds are possible overlapped. While sound source allocation has substantial ambiguity, it need to be integrated with vision system. For example, suppose a player has a ball, and the the player intends to pass a ball to someone in the back who is currently not in the visual field. A teammate who is in that position to receive a ball may shout to the player who has a ball to alert the position and demand for a pass. The player with a ball recognizes the approximate location of the teammate, but not exact enough to send an accurate pass. The player then must quickly rotate face or body to capture the teammate in the visual field. Then, two players all of sudden come into the visual field. One is ready to receive a pass, and the other is not. Auditory information now has to be merged with visual information to quickly identify which one of two players is willing to accept the pass. It will cause potential mis-pass or intercept if the player sends a pass to the other player who is not ready to accept the pass. He may be in the move to run into an open space, so that the other teammate can quickly pass a ball to him.

While this is a simple example, it illustrates the case where the use of both auditory and visual information are necessary to carry out basic behaviors in soccer, and that the integration, not a separate processing, is essential to carry out the task.

In order to carry out such a task, we need to establish a method to integrate auditory and visual information. Therefore, in this paper, we exemplify a simple example of sound source separation and direction identification with auditory and vision integration. Preliminary experiments are shown on the effects of increasing modalities in separating sound streams from a mixture of sounds. The modalities to be checked is monaural sounds, binaural (a pair of stereo microphones embedded in a dummy head), and vision.

3.1 Separation by Harmonics

Nakatani *et al.* developed a harmonic stream separation system called HBSS (Harmonic-Based Stream Segregation) [5, 13], because harmonics is mathematically defined and thus easy to formulate the processing.

HBSS extracts harmonic stream fragments from a mixture of sounds by using multi-agent system. It uses three kinds of agent; the event detector, the generator, and tracers. The event detector subtracts predicted inputs from actual input by spectral subtraction [14] and gives residue to the generator. The generator generates a tracer if residue contains harmonics. Each tracer extracts a harmonic stream fragment with the fundamental frequency specified by the generator and predicts the next input by consulting the actual next input.

HBSS uses the following equations as the model of a harmonic fragment:

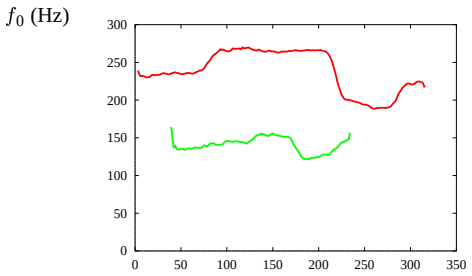
$$H(t) = \sum_{k=0} A_k(t)\sin(\theta_k(t)), \quad (1)$$

$$\dot{\theta}(t) = 2\pi f_k(t),\ (f_k(t) \simeq (k+1)\cdot f_0(t)), \quad (2)$$

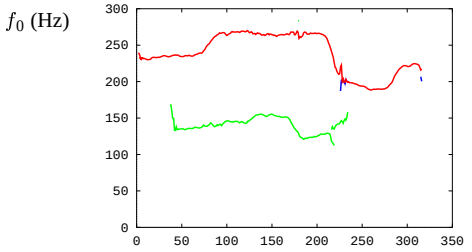
where $H(t)$ is the waveform at time t, and $A_n(t)$, $\sin(\theta_n(t))$, $\theta_n(t)$, and $f_n(t)$, respectively, are the amplitude, waveform, phase, and frequency (Hz) of the n-th harmonic component.

Extracted harmonic stream fragments are, then, grouped according to the continuity of fundamental frequencies.

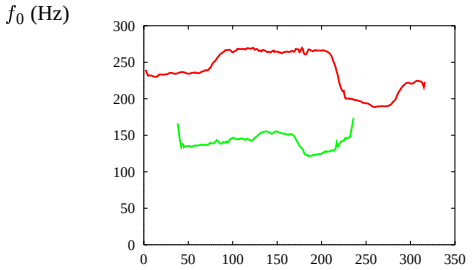
HBSS is flexible in the sense that it does not assume the number of sound sources and extracts harmonic stream fragments well. Figure 1 shows the result of experiments on harmonic stream separation from a mixture of man's and woman's utterances both saying "aiueo" (Japanese vowels,



(a) Input mixture of utterances

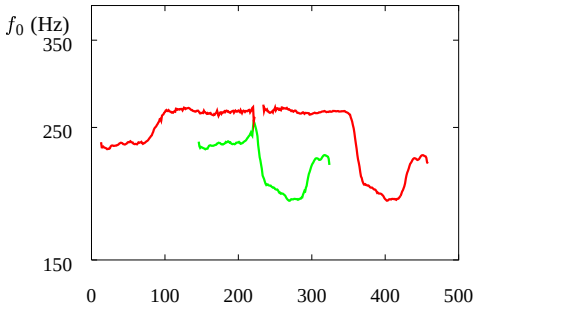


(b) Harmonic streams separated by HBSS

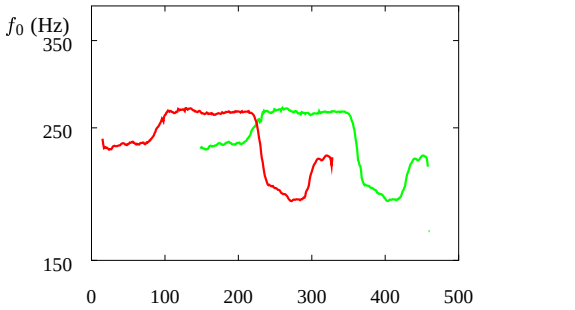


(c) Harmonic streams separated by Bi-HBSS

Figure 1: Separation of woman's and man's utterances from a mixed sound. The upper curve of fundamental frequency f_0 is woman's utterance, and the lower man's. HBSS (Monaural) fails in separation where woman's f_0 and man's f_1 are crossing.



(a) Harmonic streams separated by HBSS



(b) Harmonic streams separated by Bi-HBSS

Figure 2: Effects of Sound Source Direction: The difficulty in separation resides in the crossing of two fundamental frequencies. HBSS separates a longer stream and a short one shown in (a). Bi-HBSS separates two streams well as is shown in (b).

of course, harmonics). The input is shown in Fig. 1 (a) and fundamental frequency (f_0) of separated harmonic streams are shown in Fig. 1 (b).

However, the grouping of harmonic stream fragments may fail in some cases. For example, consider the case that two harmonic streams cross (see Fig. 2 (b)). HBSS cannot discriminate whether two harmonic streams really cross or they come closer and then go apart, since it uses only harmonics as a clue of sound source separation.

3.2 Separation by Harmonics and Direction

The use of sound source direction is proposed to overcome this problem, and Bi-HBSS (Binaural HBSS) is developed by Nakatani *et al.* [5, 15]. That is, the input is changed from monaural to binaural. Binaural input is a variation of stereo input, but a pair of microphone is embedded in a dummy head. Since the shape of a dummy head affects sounds, the interaural intensity difference is enhanced more than that for stereo microphones.

Sound source direction is determined by calculating the Interaural Time (or phase) Difference (ITD) and the Interaural Intensity Difference (IID) between the left and right channels. Usually ITD and IID are easier to calculate from binaural sounds than from stereo sounds [16].

Bi-HBSS uses a pair of HBSS to extract harmonic

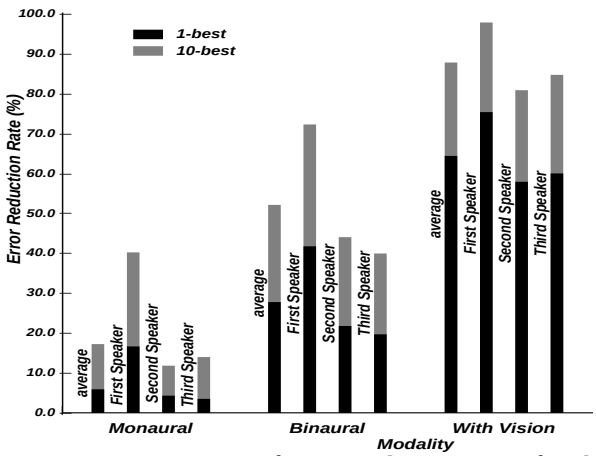


Figure 3: Improvement of Error reduction rates for the 1-best/10-best recognition of each speech by incorporating more modalities.

stream fragments for the left and right channels, respectively. The interaural coordinator adjusts information on harmonic structure extracted by the both HBSS. Then, sound source direction is determined by calculating ITD and IID between a pair of harmonic stream fragments. The sound source direction is fed back to the interaural coordinator to refine harmonic structure of harmonic stream fragment. Finally, harmonic stream fragments are grouped according to its sound source direction. Thus the problem depicted in Fig. 2 (a) is resolved (see Fig. 2 (b)).

Thus, increasing modality with sound source direction improve the performance of harmonic stream separation.

3.3 Separation by Harmonics, Direction, Visual Direction

To evaluate the performance of separation, AI challenge "*Understanding three simultaneous speeches*" [10] is attacked. Three people from different directions utter a Japanese word simultaneously. Benchmark set consists of 200 mixtures of three Japanese words uttered by different women. The second speaker utters a Japanese word 150 ms later after the first speaker, and the third speaker utters 150 ms later after the second speaker. This is because a mixture of sounds may be recognized by speech recognition system.

Since a Japanese word consists of a sequence of vowel, consonant, and vowel, speech stream is reconstructed by using harmonic streams for harmonic parts and substituting residue for non-harmonic parts [17, 18]. The idea of residue substitution is similar to the psychophysical observation known as *auditory induction* [19]. It is a phenomenon that human listeners can perceptually restore a missing sound component if it is very brief and masked by appropriate sounds.

Separated speech is tried to recognize by automatic speech recognition system, HMM-LR [20], which is based on hidden Markov model of each phonetic transition. The parameters of HMM-LR are trained by a set of 5,240 words uttered by five speakers. Of course, training data and

benchmark data are disjoint.

The performance of segregation is measured by *error reduction rates* for the 1-best and 10-best recognition [21]. Error reduction rate for the n best, $\mathcal{R}^n_{sep}$, in per cent is calculated as follows:

$$\mathcal{R}^n_{sep} = \frac{\mathcal{A}^n_{sep} - \mathcal{A}^n_{mix}}{\mathcal{A}^n_{org} - \mathcal{A}^n_{mix}} \times 100.$$

where $\mathcal{A}^n$ is a n-best accuracy of recognition, and suffix org, mix, and sep stand for the single unmixed original sounds, mixed sounds, and separated sounds, respectively. By HMM-LR, 1-best and 10-best accuracy of recognition for unmixed utterance is about 70% and 96%, respectively. 1-best and 10-best accuracy of recognition for mixed utterances is at most 5%.

Error reduction rates by speech stream separation combined with HBSS and Bi-HBSS is depicted in Fig.3. By Bi-HBSS, 55% of recognition errors are recovered, but is not satisfactory result.

One of reason why recognition errors are not recovered well is that the accuracy of sound direction is $\pm 10^\circ$ to obtain a stable separation of harmonic stream fragments.

Nakagawa *et al.* obtained the sound source direction by vision instead of the above methods. Since such a direction is accurate, sound source separation is performed by the direction filter, which uses the recalculated ITD and IID. Error reduction rates obtained by using visual direction is shown in Fig.3. By combining vision, most errors are reduced. Of course, if there are several sound sources at the same direction, this method won't work.

3.4 Visual Tracking with Auditory Direction

Visual information helps auditory sensory systems to improve the performance of sound source separation. This is also true for the opposite.

Nakagawa *et al.* reported the results of experiments to use the direction of sound source to narrow the search region of visual tracking [21]. Although visual processing consists of very simple color matching algorithm, it can track three speakers well.

In the current experiments, we are developing a tracking system whose capabilities include the followings:

(1) Someone shouts behind a robot, the robot looks back at it and starts its tracking.

(2) A robot player loses a ball. When the whistle is blown by the refree to restart the game, the robot player detects the direction of sound source, that is, the refree; it rotates toward its direction; and then it looks for a ball and resumes the tracking of the ball.

(3) A robot player tracks a ball, and then it loses the ball because the ball moves behind other robots. However, the robot hears the sound of rolling that the ball makes, it guesses where the ball will reappear, and prepares to resume the tracking.

These capabilities require auditory and visual processing as well as inference mechanism to understand what happens and to do planning for appropriate behaviors.

4 DISCUSSION

The previous section shows that increasing modalities may improve the performance of sound source separation or visual tracking. There are several observations concerning these preliminary experiments.

(1) The assumptions that each sound source separation system work well differ:

- HBSS assumes that overlapping fundamental frequencies are rare.
- Bi-HBSS assumes that overlapping fundamental frequencies from the same direction with some ambiguities are rare.
- Visual information and directional filter assumes that sound sources from the same accurate direction are rare.

Since these assumptions do not always hold, we have to combine appropriate methods that fit the current situations. To control dynamic selection of appropriate methods needs the general representation of each method for meta control. This may need a general control framework such as blackboard based systems or subsumption architectures.

(2) In soccer games, teammates are fixed and thus model-based sound source separation is promising. That is, the system acquires the characteristics of each teammate and builds mechanism of speaker identification.

(3) In soccer games, the variety of messages delivered between teamate players are relatived limited. Therefore, speech understanding systems with a limited number of vocabulary may work well even under noisy environments.

(4) Sensor fusion is needed as Kitano *et al.* already pointed out [22, 23, 24]. The actual issue is driving source of sensor fusion. One of candidates is media streams. Auditory streams and visual streams may be integrated to represent a media stream that captures various aspects of events.

(5) Real-Time Processing is mandatory. Compared with vision, auditory processing is quite slow. Digital Signal Processor (DSP) may be needed to speed up primitives in sound source separation.

(6) As the capability of CPU increases, more information is needed to understand events more precisely. This is a spiral of needs and seeds. Therefore, we need the mechanism to trade off the requirements and computing resources.

5 CONCLUSION

In this paper, we argue that auditory information is important in RoboCup and discuss some possible ways to utilize it. Since each sensory system is not enough for real-time processing, sensor fusion is essential in perception-motor control.

We thank Tomohiro Nakatani of NTT-East Multimedia Business Headquarters for his help with HBSS and Bi-HBSS. We also thank members of J-Star99, Fuminori Yamazaki and Jun Homma, for their valuable discussions.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of Acoustic Society of America*, vol. 25, pp. 975–979, 1953.

[2] A. S. Bregman, *Auditory Scene Analysis*, The MIT Press, MA., 1990.

[3] G. J. Brown and M. P. Cooke, "A computational model of auditory scene analysis," in *Proc. of Intern'l Conf. on Spoken Language Processing*, 1992, pp. 523–526.

[4] 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.

[5] T. Nakatani, H. G. Okuno, and T. Kawabata, "Auditory stream segregation in auditory scene analysis with a multi-agent system," in *Proc. of 12th National Conference on Artificial Intelligence (AAAI-94)*. 1994, pp. 100–107, AAAI.

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

[7] H. Kitano and M. Asada, "Robocup humanoid challenge: That's one small step for a robot, one giant leap for mankind," in *Proc. of International Conference on Robotics and (IROS-98)*. 1998, IEEE.

[8] B. Selman, R. A. Brooks, T. Dean, E. Horovitz, T. M. Mitchell, and Nilsson. N. J., "Challenge problems for artificial intelligence," in *Proc. of 13th National Conference on Artificial Intelligence (AAAI-96)*. 1996, pp. 1340–1345, AAAI.

[9] Y. Minami and S. Furui, "A maximum likelihood procedure for a universal adaptation method based on hmm composition," in *Proc. of 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95)*. 1995, vol. 1, pp. 129–132, IEEE.

[10] H. G. Okuno, T. Nakatani, and T. Kawabata, "Understanding three simultaneous speakers," in *Proc. of 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*. 1997, vol. 1, pp. 30–35, AAAI.

[11] M. Kashino and T. Hirahara, "One, two, many – judging the number of concurrent talkers," *Journal of Acoustic Society of America*, vol. 99, no. 4, pp. Pt.2, 2596, 1996.

[12] H. Kitano, S. Tadokoro, I. Noda, and Matsubara H., "Robocup rescue: Search and rescue in large-scale disasters as a domain for autonomous agents research," in *Proc. of International Conference on Systems, Mans and Cybernetics (SMC-99)*. 1999, IEEE.

[13] T. Nakatani, T. Kawabata, and H. G. Okuno, "A computational model of sound stream segregation with the multi-agent paradigm," in *Proc. of 1995 International Conference on Acoustics, Speech and Signal Processing (ICASSP-95)*. 1995, vol. 4, pp. 2671–2674, IEEE.

[14] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proc. of 1979 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79)*. 1979, pp. 200–203, IEEE.

[15] 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.

[16] M. Bodden, "Modeling human sound-source localization and the cocktail-party-effect," *Acta Acustica*, vol. 1, pp. 43–55, 1993.

[17] H. G. Okuno, T. Nakatani, and T. Kawabata, "Interfacing sound stream segregation to speech recognition systems — preliminary results of listening to several things at the same time," in *Proc. of 13th National Conference on Artificial Intelligence (AAAI-96)*. 1996, pp. 1082–1089, AAAI.

[18] H. G. Okuno, T. Nakatani, and T. Kawabata, "Listening to two simultaneous speeches," *Speech Communication*, vol. 27, no. 3-4, pp. 281–298, 1999.

[19] R. W. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, pp. 392–393, 1970.

[20] K. Kita, T. Kawabata, and K. Shikano, "HMM continuous speech recognition using generalized LR parsing," *Transactions of Information Processing Society of Japan*, vol. 31, no. 3, pp. 472–480, 1990.

[21] Y. Nakagawa, H. G. Okuno, and H. Kitano, "Using vision to improve sound source separation," in *Proc. of 16th National Conference on Artificial Intelligence (AAAI-99)*. 1999, AAAI, (*in print*).

[22] D. Floreano and F. Mondada, "Active perception, navigation, homing, and grasping: an autonomous perspective," in *Proc. of From Perception to Action conference*, September 1994, pp. 122–133.

[23] M. Rucci and R. Bajcsy, "Learning visuo-tactile coordination in robotic systems," in *Proc. of 1995 IEEE International Conference on Robotics and Automation*, 1995, vol. 3, pp. 2678–2683.

[24] G. J. Wolff, "Sensory fusion: integrating visual and auditory information for recognizing speech," in *Proc. of IEEE International Conference on Neural Networks*, March 1993, vol. 2, pp. 672–677.