# Using Vision to Improve Sound Source Separation

**Yukiko Nakagawa [†], Hiroshi G. Okuno [†], 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
Tel: +81-3-5468-1661, Fax: +81-3-5468-1664
‡Sony Computer Science Laboratories, Inc.
yuki@symbio.jst.go.jp, okuno@nue.org, kitano@csl.sony.co.jp

**Abstract**

We present a method of improving sound source separation using vision. The sound source separation is an essential function to accomplish auditory scene understanding by separating stream of sounds generated from multiple sound sources. By separating a stream of sounds, recognition process, such as speech recognition, can simply work on a single stream, not mixed sound of several speakers. The performance is known to be improved by using stereo/binaural microphone and microphone array which provides spatial information for separation. However, these methods still have more than 20 degree of positional ambiguities. In this paper, we further added visual information to provide more specific and accurate position information. As a result, separation capability was drastically improved. In addition, we found that the use of approximate direction information drastically improve object tracking accuracy of a simple vision system, which in turn improves performance of the auditory system. We claim that the integration of vision and auditory inputs improves performance of tasks in each perception, such as sound source separation and object tracking, by bootstrapping.

## Introduction

When we 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, we (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 auditory scene analysis, sound shall be separated into auditory streams each of which corresponds to specific auditory event (Bregman 1990; Cooke *et al.* 1993; Rosenthal & Okuno 1998).

Separation of streams from perceptive input is nontrivial task due to ambiguities of interpretation on which elements of perceptive input belong to which stream. This is particularly the case for auditory stream separation. Assume that there are two independent sound sources (this can be machines or human speakers) which create their own auditory stream, illustrated as harmonic structures shown in Fig. 1 (a). When these sound sources create sound at the same time, two auditory streams come together to a listener, superimposed harmonic structure may look like Fig. 1 (b). In this case there are two possible ways to separate auditory streams, only one of them is correct (Fig. 1 (c)).

While many research has been carried out to accurately separate such auditory streams using heuristics, there are essential ambiguities which cannot be removed by such a method. The use of multiple microphones, such as stereo microphone, binaural microphone, and microphone array is known to improve separation accuracy (Bodden 1993; Wang *et al.* 1997). However, so far there is no research to use visual information to facilitates auditory scene analysis.

At the same time, there are many research on integration of visual, auditory, and other perceptive information. Most of these studies basically use additional perceptive input in order to provide clue to shift attention of other perceptive input. For example, research of sound-driven gaze are addressing how sound source can be used to control gaze to the object which generates sound (Ando 1995; Brooks *et al.* 1998; Wolff 1993). Similarly, integration of vision and audition to find an objects using active perception has been proposed for autonomous robot (Wang *et al.* 1997; Floreano & Mondada 1994). By the same token, touch-driven gaze is the fusion of visuo-tactile sensing in order to control gaze using tactile information (Rucci & Bajcsy 1995).

However, in these research the processing of each perceptive input is handled separately except for gaze control. Therefore, there is no effect of increased modality for each perceptive input processing.

In this paper, we argue that the use of visual information drastically improves auditory stream separation accuracy. The underlying hypothesis is that ambiguities in stream separation arise from two reasons:

- there are missing dimensions in the state-space which represents perceptive inputs, and
- some constraints are missing which can be used to eliminate spurious trajectories in the state-space.

We will demonstrate viability of the hypothesis using auditory stream separation of three-simultaneous speeches carried out by (1) a monaural microphone system, (2) a bin-

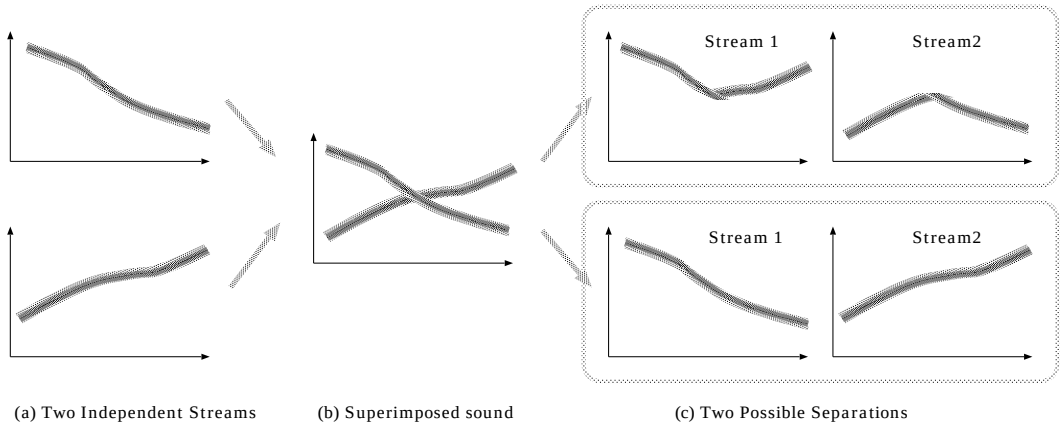



Figure 1: Example of Overlapped Auditory Streams Separation

aural microphone system, and (3) a binaural microphone[1] system with vision.

Separation of sound source is significant challenge for auditory system for the real world. In real world environment, multiple objects create various sounds, such as human voice, door noise, automobile sounds, music, and so forth. Human being with normal hearing capability can separate these sounds even if these sounds are generated at the same time, and understand what is going on. In this paper, we focus on separation of multiple and simultaneous human speeches, where up to three persons speak simultaneously. At first glance, it may look a bit odd to assume three persons speak simultaneously. However, it turns out that this situation has many potential applications. In many voice-controlled devices, such as a voice-commanded car-navigation system, the system needs to identify and separate auditory stream of the specific speaker from environmental noise and speeches of other people. Most of commercial level speech recognition system built-in into portable devices needs identify and separate owner's voice from background noise and voices of other person happened to be talking to someone.

In addition, due to the complexity of the task, if we can succeed in separation of multiple simultaneous speeches, it would be much easier to apply the method to separate various sounds that has drastically different from human voice. "Understanding Three Simultaneous Speeches" (Okuno, Nakatani, & Kawabata 1997). is also one of the AI challenge problem chosen at IJCAI.

## Needs for Visual Information

There are many candidates for clues for sound source separation; Some acoustic attributes include harmonics (fundamental frequency and its overtones), onset (starting point of sound), offset (ending point of sound), AM (Amplitude Modulation), FM (Frequency Modulation), timbre, formants, and sound source localization (horizontal and vertical directions, distance). Case-based separation with sound database may be possible.

The most important attribute is harmonics, because it is mathematically defined and thus easy to formulate the processing. Nakatani *et al.* developed Harmonic Based Stream Segregation System (HBSS) to separate harmonic streams from a mixture of sounds (Nakatani, Okuno, & Kawabata 1994). 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 (Boll 1979) 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. Then, extracted harmonic stream fragments are 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. However, the grouping of harmonic stream fragments may fail in some cases. For example, consider the case that two harmonic streams cross (see Fig. 1 (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.

The use of sound source direction is proposed to overcome this problem and Bi-HBSS (Binaural HBSS) is developed by Nakatani *et al.* (Nakatani, Okuno, & Kawabata 1994; Nakatani & Okuno 1999). In other words, 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

[1] A binaural microphone is a pair of microphones embedded in a dummy head.

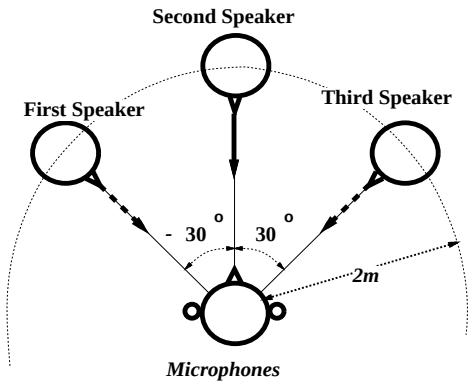



Figure 2: Position of Three Speakers for Benchmark

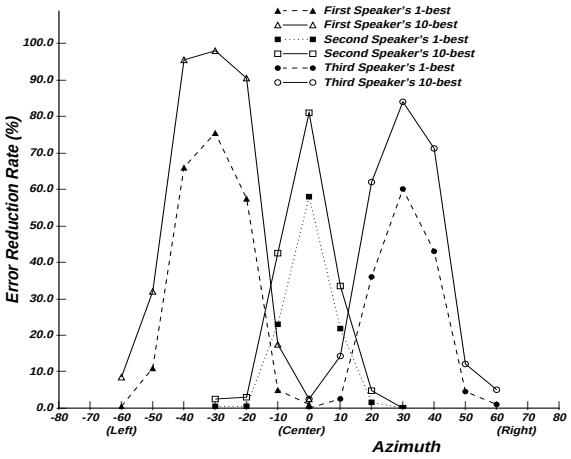



Figure 3: Error Reduction rates for the 1-best and 10-best recognition by assuming the sound source direction

binaural sounds than from stereo sounds (Bodden 1993).

Bi-HBSS uses a pair of HBSS to extract harmonic 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. 1 (b) is resolved. Speech stream is reconstructed by using harmonic streams for harmonic parts and substituting residue for non-harmonic parts (Okuno, Nakatani, & Kawabata 1996).

## Preliminary Experiment

Since the direction determined above in Bi-HBSS may contain an error of $\pm 10^\circ$, which is considered very large, its influence on the error reduction rates of recognition is investigated. For this purpose, we construct a direction-pass filter which passes only signals originating from the specified direction and cuts other signals. We measured the IID and ITD in the same anechoic room for every $5^\circ$ azimuth in the horizontal plane. A rough procedure of direction-pass filter is as follows:

1. Input signal is given to a set of filter banks for the left and right channels and analyzed by discrete Fourier transformation,
2. IID and ITD for each frequency band are calculated and its direction is determined by comparing IID and ITD. This is because ITD is more reliable in lower frequency regions, while IID is more reliable in higher frequency regions.
3. Then, each auditory stream is synthesized by applying inverse Fourier transformation to the frequency components originating from the direction.

**Benchmark Sounds** The task is to separate simultaneous three sound sources using binaural microphone and vision. (See Fig. 2) The benchmark sound set used for the evaluation of sound source separation and recognition consists of 200 mixture of three utterances of Japanese words. The mixture of sounds are created analytically in the same manner as (Okuno, Nakatani, & Kawabata 1996). Of course, a small set of benchmarks were actually recorded in an anechoic room, and we confirmed that the synthesized and actually recorded data don't cause a significant difference in speech recognition performance.

1. All speakers are located at about 2 meters from the pair of microphones installed on a dummy head as is shown in Fig. 2.
2. The first speaker is a woman located at $30^\circ$ to the left from the center.
3. The second speaker is a man located in the center.
4. The third speaker is a woman located at $30^\circ$ to the right from the center.
5. The order of utterance is from left to right with about 150ms delay.

   This delay is inserted so that the mixture of sounds was to be recognized without separation.
6. The data is sampled by 12KHz and the gain of mixture of sounds is reduced if the data overflows in 16 bit. Most mixtures are reduced by 2 to 3 dB.

**Evaluation Criteria** The recognition performance is measured by *the error reduction rate for the 1-best and 10-best recognition*. First, *the error rate caused by interfering sounds* is defined as follows. Let the $n$-best recognition rate be the cumulative accuracy of recognition up to the $n$-th candidate, denoted by $\mathcal{CA}^{(n)}$. The suffix, *org*, *sep*,

or *mix* is added to the recognition performance of the single unmixed original sounds, mixed sounds, and separated sounds, respectively. The error rate caused by interfering sounds, $\mathcal{E}^{(n)}$, is calculated as $\mathcal{E}^{(n)} = \mathcal{CA}^{(n)}_{org} - \mathcal{CA}^{(n)}_{mix}$.

Finally, the error reduction rate for the $n$-best recognition, $\mathcal{R}^{(n)}_{sep}$, in per cent is calculated as follows:

$$\mathcal{R}^{(n)}_{sep} = \frac{\mathcal{CA}^{(n)}_{sep} - \mathcal{CA}^{(n)}_{mix}}{\mathcal{CA}^{(n)}_{org} - \mathcal{CA}^{(n)}_{mix}} \times 100 = \frac{\mathcal{CA}^{(n)}_{seg} - \mathcal{CA}^{(n)}_{mix}}{\mathcal{E}^{(n)}} \times 100.$$

**Preliminary Results** 200 mixtures of three sounds are separated by using a filter bank with the IID and ITD data. We separate sounds in every $10^\circ$ azimuth (direction) from $60^\circ$ to the left to $60^\circ$ to the right from the center. Then each separated speech stream is recognized by a Hidden Markov Model based automatic speech recognition system (Kita, Kawabata, & Shikano 1990).

The error reduction rates for the 1-best and 10-best recognition of separated sound for every $10^\circ$ azimuth are shown in Fig. 3. The correct azimuth for this benchmark is $30^\circ$ to the left (specified by $-30^\circ$ in Fig. 3), $0^\circ$, and $30^\circ$ to the right. For these correct azimuths (directions), recognition errors are reduced significantly. The sensitivity of error reduction rates to the accuracy of the sound source depends on how other speakers are close to. That's why the curve of error reduction rates for the center speaker is the steepest in Fig. 3.

This experiment proves that if the correct direction of the speaker is available, separated speech is of a high quality at least from the viewpoint of automatic speech recognition. In addition, the error reduction rates is quite sensible to the accuracy of the sound source direction if speech is interfered by closer speakers.

While binaural microphone provides direction information at certain accuracy, it is not enough to separate sound source in realistic situations. There are inherent difficulties in obtaining high precision direction information by solely depending on auditory information.

The fundamental question addressed in this paper is that how the use of visual information can improve the sound source separation by providing more accurate direction information.

## Integration of Visual and Auditory Stream

In order to investigate how the use of visual input can improve auditory perception, we developed a system consists of binaural microphone and CCD camera, as input devices, and sound source separation system (simply, *auditory system*) and color-based real time image processing system (simply, *vision system*), that interacts to improve accuracy of processing in both modalities. The concept of integrated system is depicted in Fig. 4.

If auditory scene analysis module detects a new sound source, it may trigger vision module to focus on it. If vision module identifies the position of the sound source, it returns the information to auditory scene analysis module and conflict resolution module checks whether the both

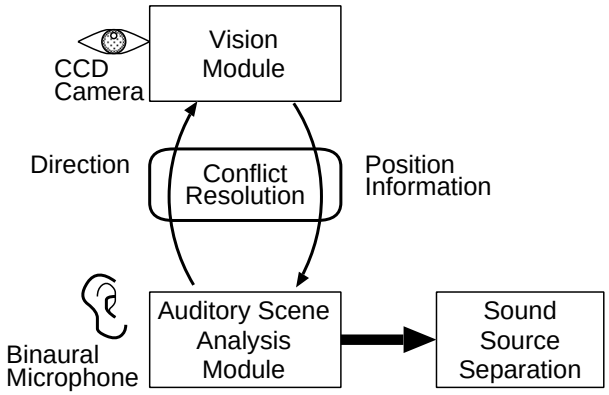



Figure 4: Concept of Integrated Vision and Auditory Systems

information specifies the same sound source. In case of the same sound source, the position information subsumes the direction information as long as the sound source exists.

While there are several ways for vision and auditory perceptions to interact, we focus on how information on position of possible sound sources derived from both vision and auditory perception interact to improve auditory stream separation. In essence, a visual input provides information on directions of possible sound sources, which can be used to better separate auditory stream. At the same time, as we will discuss in depth later, information of approximate direction of sound sources significantly improve accuracy of vision system in tracking possible sound sources by constraining possible location of target objects.

### Auditory Streams

The task of audition is to understand *auditory events*, or the sound sources. An auditory event is represented by *auditory streams*, each of which is a group of acoustic components that have consistent attributes. Since acoustic events are represented hierarchically (e.g. orchestra), auditory streams have also a hierarchical structure.

Auditory system should separate auditory streams by using the sound source direction and do the separation incrementally and in real-time, but such a system has not been developed so far. Therefore, as a prototype of auditory system, we use Bi-HBSS, because it separates harmonic structures incrementally by using harmonic structure and the sound source direction.

### Visual Streams

The task of vision is to identify possible sound sources. Among various methods to track moving objects, we used a simple color-based tracking. This is because we are also interested in investigating how accuracy of visual tracking can be improved using information from the auditory system, particularly sound source position.

Images are taken by a CCD camera (378K pixel 2/3" CCD) with a wide conversion lens, video capture board in a personal computer (Pentium II 450MHz, 384MB RAM), a

Frame 7 Frame 51

Frame 78 Frame 158

Figure 5: Some Visual Images for Tracking Experiments

the rate of six frames per second for forty seconds. Captured images are $640 \times 480$ pixels with 16 bit color. R, G and B in a pixel is represented by 5 bit, respectively. The pixel color (RGB) is translated into HSV color model to attain higher robustness against small changes in lighting condition.

In this experiment, we assume that a human face, especially mouth is a possible sound source and that the mouth is around the gravity center of face. Therefore, the vision system computes clusters of skin colors, and their center of gravity to identify the mouth.

Since there are multiple clusters of skin color, such as face, hands, and legs, clusters that are not considered as face shall be eliminated using various constraints. Such constraints includes positional information from auditory system, heights, velocity of cluster motion, etc.

## Experiments

### Test Data

**Auditory Sounds and Criteria of Evaluation** Since the preliminary experiment is already reported in this paper, the same benchmark sounds are used and the same evaluation criteria for performance is adopted.

**Visual Images** The auditory situation described above was realized in a visual image that has three people sitting around the table and discussing some business issues. Image is taken by a CCD camera positioned two meters from the speakers. Excerpts of frames from the image are shown in Fig. 5. Apart from face of each person, there are few objects that causes false tracking. One is a yellow box just left side of the person in the center, and the other is a knee (under the table) of the person in the left. In addition, hands can be mis-recognized as it has similar color with face.

### Experiment 1: Effect of Modalities

In Experiment 1, we investigate the effect of three modalities. They are listed in the order of increasing modalities:

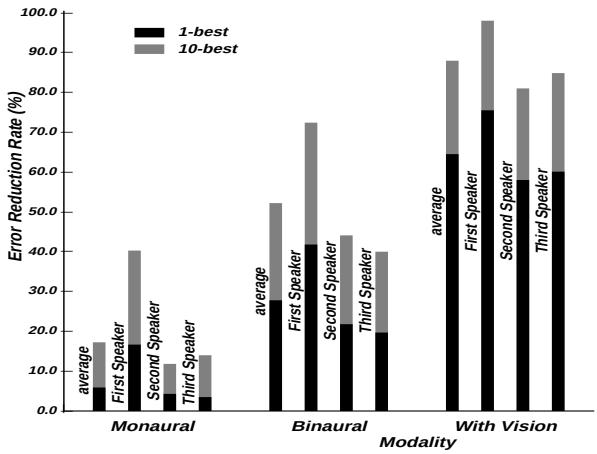



Figure 6: Experiment 1: Improvement of Error reduction rates for the 1-best/10-best recognition of each speech by incorporating more modalities

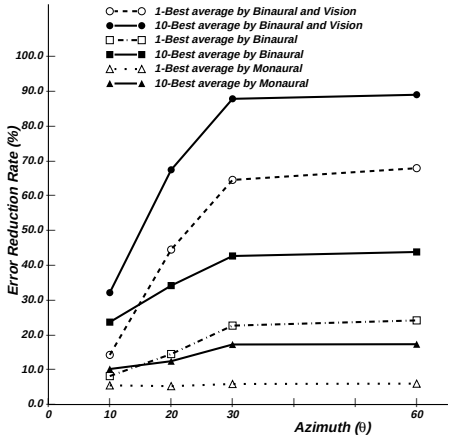



Figure 7: Experiment 2: How average of error reduction rates for the 1-best/10-best recognition of each speech by incorporating modalities vary when the position of each speaker varies.

1. Speech stream separation by monaural inputs,
2. Speech stream separation by binaural inputs, and
3. Speech stream separation by binaural inputs with visual information.

We use HBSS, Bi-HBSS and simulator for integrated systems depicted in Fig. 4 for the three experiments, respectively.

Error reduction rates for the 1-best and 10-best recognition of each speech is shown in Fig. 6. As more modalities are incorporated in auditory system, error reduction rates are improved drastically.

### Experiment 2: Robustness of Modality against Closer Speakers

In Experiment 2, we investigate the robustness of the three speech stream separation algorithms by changing the directions of each speakers. The azimuth between the first and second speakers and that between the second and third speakers are the same, say "$\theta$". We measured the average error reduction rates for the 1-best and 10-best recognition for $10^\circ$, $20^\circ$, $30^\circ$, and $60^\circ$.

The result of error reduction rates by the three algorithms is shown in Fig. 7. Error reduction rates saturate around the azimuth of more than $30^\circ$. For the azimuth of $10^\circ$ and $20^\circ$, error reduction rates for the second (center) speaker are quite poor compared with the other speakers (this data is not shown in Fig. 7).

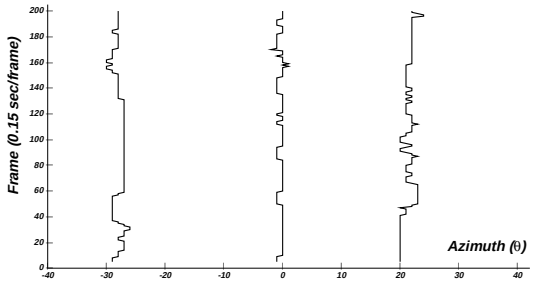
## Experiment 3: Accuracy of Vision System with Auditory Feedback

Experiments 1 and 2 assume that vision system provides precise direction information, and thus the auditory system can disambiguate harmonic structures without checking its validity. However, question can be raised on the accuracy of vision system. If the vision system provides wrong direction information to the auditory system, the performance of sound source separation may be drastically deteriorated, because it must operate under wrong assumptions. Therefore, Experiment 3 focuses on how the accuracy of vision system is improved as more constraints are incorporated.

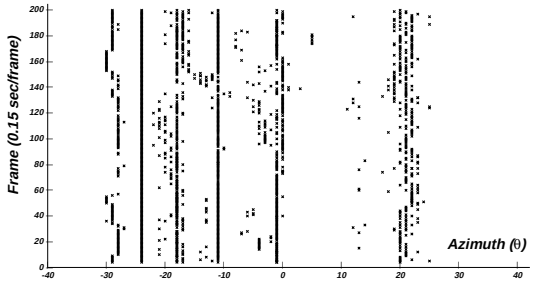
We measured tracking accuracy of a simple color-based tracking system with (1) no constraints (purely rely on cluster of color), (2) presumed knowledge on human heights, (3) approximate direction information ($-40^\circ \sim -20^\circ$, $-10^\circ \sim -10^\circ$, and $20^\circ \sim 40^\circ$) from the auditory system, and (4) using both height and direction information. Fig. 8 shows actual tracking log for each case.

In this experimental data, speakers are sitting around the table where they can be seen at $-30^\circ$, $0^\circ$, and $20^\circ$ in the visual field of the camera.

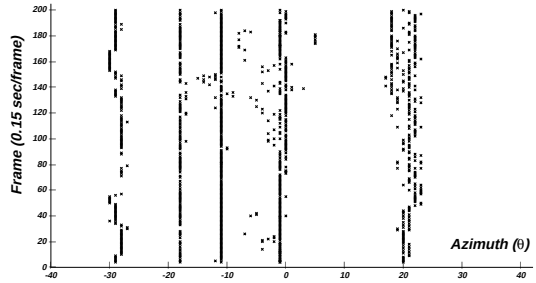
The result of tracking accuracy is shown in Fig. 8. As a reference for comparison, accurate face position is annotated manually (Fig. 8 (R)). When only color is used for tracking, there are numbers of spurious clusters that are mistakenly recognized as face (Fig. 8 (a)). Using knowledge on human height, some clusters can be ruled out when it is located at position lower than table or higher than2m. Nevertheless, many spurious clusters remains. For example, clusters at azimuth $-12^\circ$ and $-18^\circ$ are a yellow box at left of the person in the center. Imposing direction information from the auditory system drastically reduced spurious tracking (Fig. 8 (c)). However, there are a few remaining mis-recognition. A cluster at $-25^\circ$ is actually a knee of the person at the left. Use of direction information cannot rule out possible cluster even if it violate height constraints, because it cannot provide position information on elevation in the current implementation. Combining direction information and height constraints drastically improve accuracy of the tracking (Fig. 8 (d)).



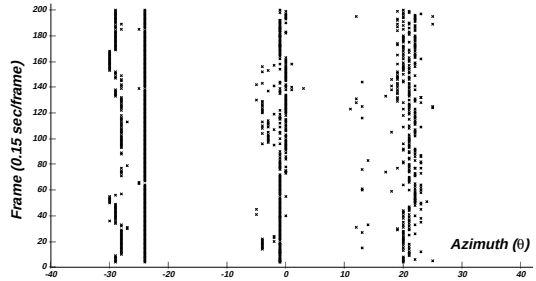

(R) Accurate face position annotated manually.



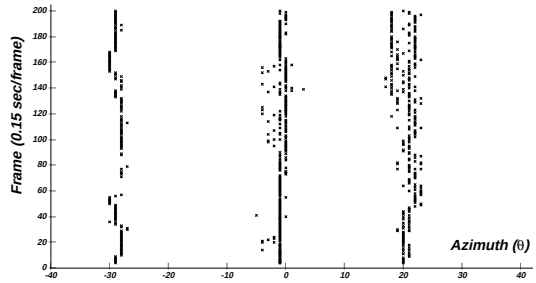

(a) By Color Only




(b) By Color and Height

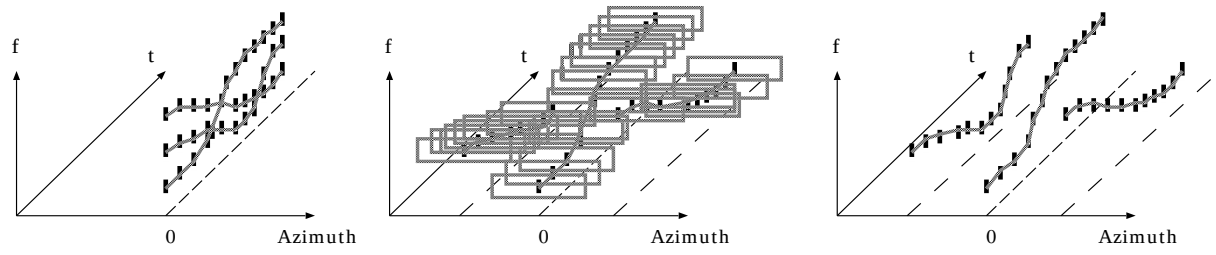



(c) By Color and Audio




(d) By Color, Height, and Audio

Figure 8: Tracking Accuracy of the Vision System under various Constraints.




(a) Monaural (b) Binaural (c) Binaural with vision

Figure 9: Spatial Feature of Auditory Streams

### Observations on Experiments

Some observations on the experiments are summarized below:

1. The error reduction rates for the 1-best and 10-best is greatly improved by fixing the direction of sound sources to the correct one. Since Bi-HBSS separates auditory streams by calculating the most plausible candidate, the direction of sound source is not stable. This is partially because some acoustic components may disappear by mixing sounds.
2. If the precise direction of visual information is available, the error reduction rates are drastically improved. Allowable margin of errors in the direction of speaker is narrower for the second (center) speaker than for the others, because he is located between them.
3. The direction of sound source can be obtained with $\pm 10^\circ$ errors by Bi-HBSS, while our simple experiments with cameras show that error margin is about $\pm 2 \sim 3^\circ$ even using rather simple vision system when combined with direction information from auditory system and height constraints.

   Therefore, information fusion of visual and auditory information is promising.
4. By fixing the direction supplied by vision module, precalculated IID and ITD data are required. However, this prerequisite may not be fulfilled in actual environments. Online adjustment of IID and ITD data is required to be apply to more realistic environment.
5. Another problem with Experiment 3 is that the number of auditory streams and that of visual streams differ. For example, some sound sources may be occluded by other objects. Or some possible sound source (speaker) does not speak actually but listens to other people's talk. In this paper, the latter case is excluded, but the former case remains as future work.

## Discussions

The central issue addressed in this paper is that how different perceptive inputs affect recognition process of a specific perceptive input. Specifically, we focused on the issue of auditory scene analysis in the context of separating streams of multiple simultaneous speeches, and how visual inputs affects the performance of auditory perception.

As briefly discussed already, the difficulties in the auditory stream separation lies in the fact that trajectories of independent streams overlap in the state space, so that clear discrimination cannot be maintained throughout the stream. Perception based on monaural auditory input has very limited dimension as it can only use amplitude and frequency distribution. There is no spatial axis. As illustrated in Fig. 9 (a), auditory streams overlap on the same spatial plane. Using binaural inputs expands dimension as it can now use amplitude and phase difference of sound sources, which adds spatial axis to the state space.

However, spatial resolution based on sound is limited due to velocity of sounds and limitation in determining amplitude and phase differences between two microphones. This is particularly difficult in reverberant environment, where multiple paths exist between sound sources and microphone due to reflection of room walls. Thus, as illustrated in the Fig. 9 (b), there are significant overlap in the auditory streams. (Ambiguities are shown as shaded boxes.)

Introduction of visual inputs, when appropriately used, adds significantly large dimensions, such as precise position, color, object shape, motion, etc. Among these features, information on positions of objects contribute substantially to the auditory perception. With visual information, the location of sound sources can be precisely determine with an accuracy of few degrees for a point source at 2-meter distance. With this information, overlap of trajectories are significantly reduced (Fig. 9 (c)). Experimental results clearly demonstrates this is actually the case for sound source separation.

By the same token, the performance of the vision system can be improved with the information from the auditory system. As the third experiments demonstrates, even a simple color-based visual tracking system can be highly accurate if approximate position information on possible sound source were provided from the auditory system, together with other constraints such as height constraints for human face positions.

These results suggests that interaction between different

perception can bootstrap performance of each perception system. This implies that even if performance of each perception module is not highly accurate, an integrated system can exhibit much higher performance than simple combination of subsystems. It would be a major open issue for future research to identify what are conditions and principles which enables such bootstrapping.

## Conclusion

The major contribution of this work is that the effect of visual information in improving auditory stream separation was made clear. While many research has been performed on integration of visual and auditory inputs, this is the first study to clearly demonstrate that information from a sensory input (e.g. vision) affects processing quality of other sensory inputs (e.g. audition). In addition, we found that accuracy of the vision system can be improved by using information derived from the auditory system. This is a clear evidence that integration of multiple modality, when designed carefully, can improve processing of other modalities, thus bootstrap the coherence and performance of the entire system.

Although this research focused on vision and audition, the same principle applies to other pairs of sensory inputs, such as tactile sensing and vision. The important research topic now is to explore possible interaction of multiple sensory inputs which affects quality (accuracy, computational costs, etc) of the process, and to identify fundamental principles for intelligence.

## Acknowledgments

We thank Tomohiro Nakatani of NTT Multimedia Business Headquarter for his help with HBSS and Bi-HBSS. We also thank members of Kitano Symbiotic Systems Project, Dr. Takeshi Kawabata of NTT Cyber Space Laboratories, and Dr. Hiroshi Murase of NTT Communication Science Laboratories for their valuable discussions.

## References

Ando, S. 1995. An autonomous three–dimensional vision sensor with ears. *IEICE Transactions on Information and Systems* E78–D(12):1621–1629.

Bodden, M. 1993. Modeling human sound-source localization and the cocktail-party-effect. *Acta Acustica* 1:43–55.

Boll, S. F. 1979. 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)*, 200–203. IEEE.

Bregman, A. S. 1990. *Auditory Scene Analysis*. MA.: The MIT Press.

Brooks, R. A.; Breazeal, C.; Irie, R.; Kemp, C. C.; Marjanovic, M.; Scassellati, B.; and Williamson, M. M. 1998. Alternative essences of intelligence. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)*, 961–968. AAAI.

Cooke, M. P.; Brown, G. J.; Crawford, M.; and Green, P. 1993. listening to several things at once. *Endeavour* 17(4):186–190.

Floreano, D., and Mondada, F. 1994. Active perception, navigation, homing, and grasping: an autonomous perspective. In *Proceedings of From Perception to Action conference*, 122–133.

Kita, K.; Kawabata, T.; and Shikano, K. 1990. HMM continuous speech recognition using generalized LR parsing. *Transactions of Information Processing Society of Japan* 31(3):472–480.

Nakatani, T., and Okuno, H. G. 1999. Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication* 27(3-4). (*in print*).

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

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

Okuno, H. G.; Nakatani, T.; and Kawabata, T. 1997. Understanding three simultaneous speakers. In *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, volume 1, 30–35. AAAI.

Rosenthal, D., and Okuno, H. G., eds. 1998. *Computational Auditory Scene Analysis*. NJ.: Lawrence Erlbaum Associates.

Rucci, M., and Bajcsy, R. 1995. Learning visuo-tactile coordination in robotic systems. In *Proceedings of 1995 IEEE International Conference on Robotics and Automation*, volume 3, 2678–2683.

Wang, F.; Takeuchi, Y.; Ohnishi, N.; and Sugie, N. 1997. A mobile robot with active localization and discrimination of a sound source. *Journal of Robotic Society of Japan* 15(2):61–67.

Wolff, G. J. 1993. Sensory fusion: integrating visual and auditory information for recognizing speech. In *Proceedings of IEEE International Conference on Neural Networks*, volume 2, 672–677.