# Incorporating Visual Information into Sound Source Separation

**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

Tel: +81-3-5468-1661, Fax: +81-3-5468-1664

* Department of Information Science, Science University of Tokyo

‡SONY Computer Science Laboratories, Inc.

okuno@nue.org, yuki@symbio.jst.go.jp, 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 a 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 binaural microphone and microphone array which provide spatial information for separation. However, these methods still have around 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. We argue, from the experiments, in this paper, that integration of vision and auditory sensory inputs improves cognitive tasks such as auditory stream separation.

## 1 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 and 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; Morita, 1990; Sugie *et al.*, 1988; Schmids, 1986; Stadler and Rabinowitz, 1993; Wang *et al.*, 1997]. However, so far there is no research to use visual information to facilitate computational 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 and 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 and 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

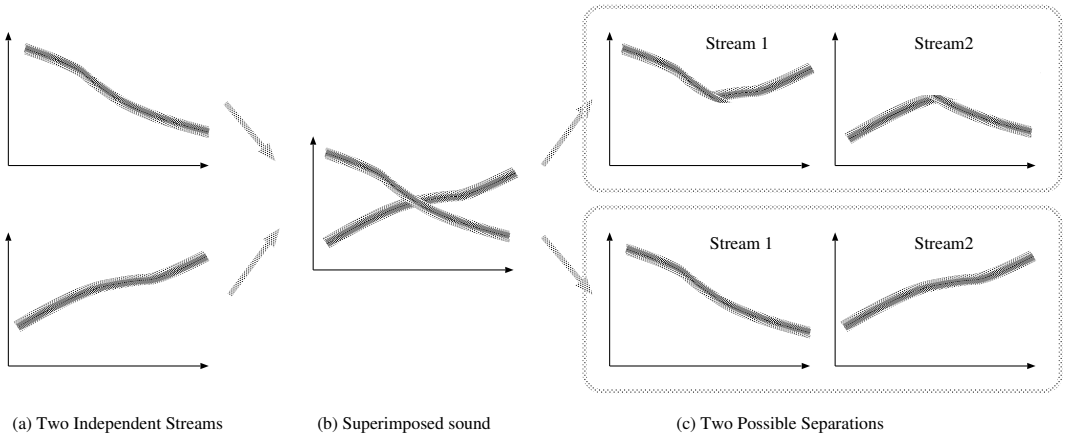



Figure 1: Example of Overlapped Auditory Streams Separation

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 binaural microphone system, and (3) a binaural microphone 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 *et al.*, 1997]. is also one of the AI challenge problem chosen at IJCAI.

## 2 Why Visual Information is Needed?

### 2.1 Clues for Sound Source Separation

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.

**Harmonic Structure**

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 *et al.*, 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.

**Sound Source Direction**

The use of sound source direction is proposed to overcome this problem and Bi-HBSS (Binaural HBSS) is developed by Nakatani *et al.* [Nakatani *et al.*, 1994; Nakatani and 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 sound signals which is called "*Head-Related Transfer Function (HRTF)*", 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 or IPD) 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 [Bodden, 1993].

Bi-HBSS uses a pair of HBSS to extract harmonic stream fragments for the left and right channels, respectively. A coordinating agent called *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 obtained 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. Once harmonic stream fragments are separated, a speech stream is reconstructed by using harmonic stream fragments for harmonic parts and substituting residue for non-harmonic parts [Okuno *et al.*, 1996; 1999].

**Determining Sound Source Direction**

To determine the direction of harmonic structure, the directions of each harmonic component, i.e. fundamental frequency and overtones, is calculated. First, the IPD is calculated for the left and right channel of the same harmonic component. Then it is converted to the ITD. The direction of a harmonic component is determined by comparing its IID and ITD with the IID and ITD data for the binaural microphone specified by HRTF.

The IID $\triangle p_k$ and the IPD $\triangle \phi_k$ of a harmonic component $f_k$ are calculated from its attributes as follows:

$$\triangle p_k = 10 \cdot \log_{10}(\mid \mathrm{FA}^r_\tau(f_k) \mid^2 / \mid \mathrm{FA}^l_\tau(f_k) \mid^2), \quad (1)$$

$$\triangle \phi_k = \arg(\mathrm{FA}^r_\tau(f_k)) - \arg(\mathrm{FA}^l_\tau(f_k)), \quad (2)$$

where $l$ and $r$ indicate the left and right channels, respectively, and $\arg(\cdot)$ is the function calculating the phase of the component. To convert IPD $\phi_k$ to ITD $\triangle t_{k,n}$, constraints of Eqs. (3) and (4) are applied.

$$2\pi f_k \triangle t_{k,n} = \triangle \phi_k + 2\pi n, \quad (n = 0, \pm 1, \ldots), \quad (3)$$

$$\triangle t_{\min} < \triangle t_{k,n} < \triangle t_{\max}, \quad (4)$$

where $\triangle t_{\min}$, and $\triangle t_{\max}$ are the minimum and the maximum ITDs calculated from the property of the HRTF. Here, multiple candidates may be obtained because the phase may rotate more than $2\pi$ in high frequency region. Therefore, more candidates is generated for the higher frequency component.

Please note that ITD is measured by a unit of millisecond. Here, we use the following values for constants.:

$$(\triangle t_{\max} = 0.83 \text{ ms}, \triangle t_{\min} = -0.83 \text{ ms})$$

Next, a direction candidate $d_{k,n}$ of $f_k$ is calculated from the ITD $\triangle t_{k,n}$, and then verified by the IID $\triangle p_k$. The time difference property $D_{hrtf}(\cdot)$ of the HRTF is used to obtain the $d_{k,n}(\tau)$.

$$d_{k,n}(\tau) = D_{hrtf}(\triangle t_{k,n}), \quad (5)$$

Then, the obtained direction $d_{k,n}(\tau)$ is verified whether the following conditions are satisfied:

$$P_{\min}(f_k, d_{k,n}(\tau)) < \triangle p_k < P_{\max}(f_k, d_{k,n}(\tau)), \quad (6)$$

where $P_{\min}(f_k, d_{k,n}(\tau))$ and $P_{\max}(f_k, d_{k,n}(\tau))$ are minimum and maximum values of the IID for the frequency $f_k$ and the direction $d_{k,n}(\tau)$, which are specified by HRTF.

Now, the direction of a harmonic structure is determined; by using the *direction histogram*. The *direction histogram* $h(m)$ divides the direction of sound sources from right to left into $m$ segments ($m = 21$), each of which has the same width of the ITD ($= 0.083$ ms, $\simeq 10$ degrees). The sum of the intensity is calculated in each discrete direction as follow:

$$h(m) = \sum_{f_k \in L(m)} \sum_{ch} \mid \mathrm{FA}^{ch}_\tau(f_k) \mid, \quad (7)$$

where $L(m)$ is a set of harmonic components whose directions are located in the discrete direction $m$. Finally, the direction is determined as $m$ that gives the maximum value $h(m)$.

The direction of a harmonic structure is defined as stable when the direction remains within a limited region (say, within 0.41 ms) for a specific time (say, for more than 75 ms). Once a stable direction is obtained for a fragment, the latest stable direction substitutes for the direction of the fragment at a frame. This additional procedure provides a long-term analysis to obtain a reliable direction information while the direction obtained just by a short-term analysis is not always reliable because it might be affected by sound interference.

## 2.2 Preliminary Experiment

Since the direction determined above in Bi-HBSS may contain an error of $\pm 10°$, 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° azimuth in the horizontal plane. A rough procedure of direction-pass filter is as follows:

The procedure for extracting sound components originating from a given sound direction is as follows [Bodden, 1993; Lyon, 1984]:

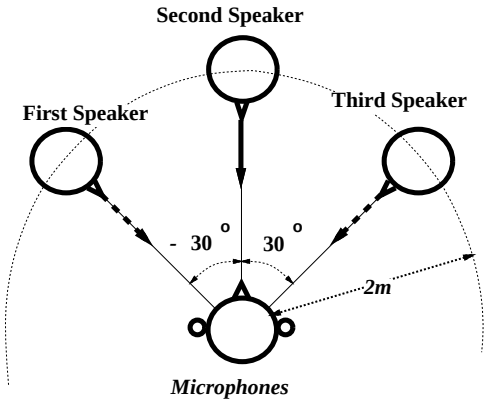



Figure 2: Position of Three Speakers for Benchmark

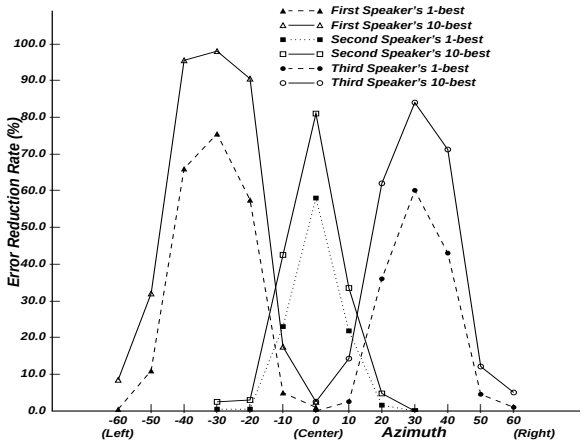



Figure 3: Error Reduction rates for the 1-best and 10-best recognition by assuming the sound source direction

1. the input signal is put into a set of filter banks for the left and right channels,
2. the IID and IPD for each frequency band are calculated and its direction is determined by thresholding the IID and IPD,
3. then, the outputs of the filters coming from the target direction are combined to obtain the target signal.

The direction threshold of the IID and IPD is determined by taking account of the deviation of the direction estimation error; that is, the IPD is dominant in low-frequency regions, while the IID is dominant in the high frequency regions.

In the experiment, we used discrete fourier transformation for the frequency analysis in both channels [Sugie *et al.*, 1988], calculated IPD, $\triangle\phi$, and IID, $\triangle p$, at each frequency point $f$, then extracted the direction by the same way as obtaining that of harmonic component described in the section 2.1. Then, sounds were synthesized by applying inverse fourier transformation to the frequency components coming from the direction neighboring to the target direction ($\pm$0.25 ms).

**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 *et al.*, 1996; 1999]. 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° 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° 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° azimuth (direction) from 60° to the left to 60° to the right from the center. Then each separated speech stream is recognized by a Hidden Markov Model based automatic speech recognition system [Kita *et al.*, 1990].

The error reduction rates for the 1-best and 10-best recognition of separated sound for every 10° azimuth are shown in Fig. 3. The correct azimuth for this benchmark is 30° to the left (specified by −30° in Fig. 3), 0°, and 30° 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.

# 3 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 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.

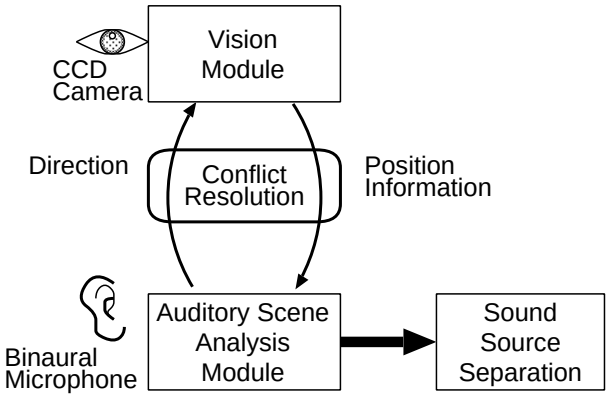



Figure 4: Concept of Integrated Vision and Auditory Systems

## 3.1 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.

## 3.2 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.

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.

# 4 Experiments

## 4.1 Test Data

Since the preliminary experiment is already reported in this paper, the same benchmark sounds are used and the same evaluation criteria for performance, that is, error reduction rates, are adopted.

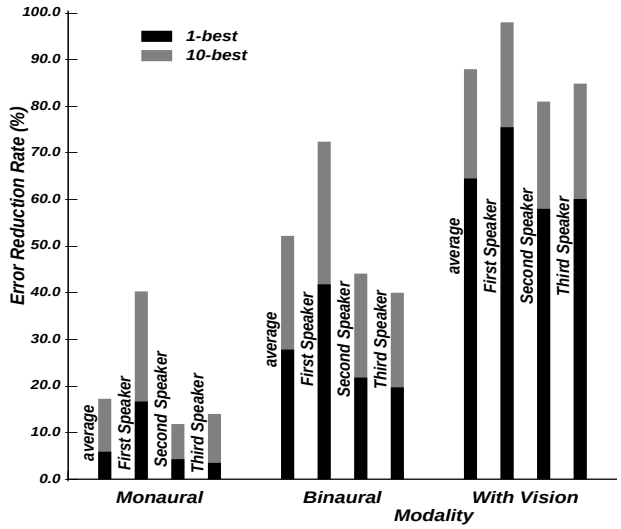



Figure 5: Experiment 1: Improvement of Error reduction rates for the 1-best/10-best recognition of each speech by incorporating more modalities

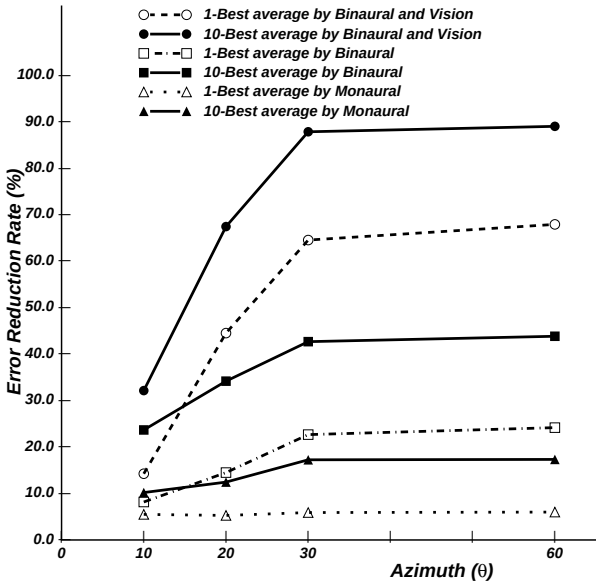



Figure 6: Experiment 2: How average of error reduction rates for the 1-best/10-best recognition of each speech by incorporating modalities vary when the position of each speaker varies.

### 4.2 Experiment 1: Effects of Modalities

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

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. 5. As more modalities are incorporated in auditory system, error reduction rates are improved drastically.

### 4.3 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°, 20°, 30°, and 60°.

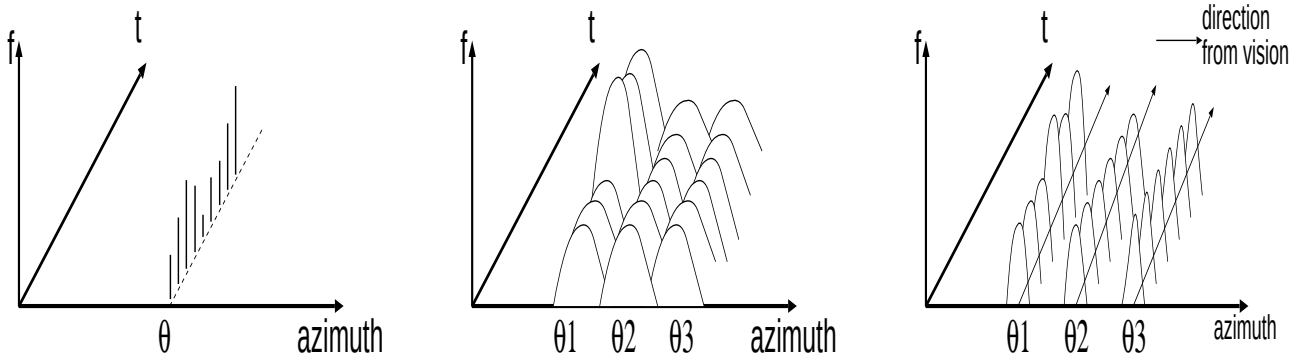
The result of error reduction rates by the three algorithms is shown in Fig. 6. Error reduction rates saturate around the azimuth of more than 30°. For the azimuth of 10° and 20°, error reduction rates for the second (center) speaker are quite poor compared with the other speakers (this data for each speaker is not shown in Fig. 6).

### 4.4 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 $\pm 10°$ errors by Bi-HBSS, while our simple experiments with cameras show that error margin is about $\pm 2 \sim 3°$ 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




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

Figure 7: Spatial Feature of Auditory Streams

data is required to be apply to more realistic environment.

5. Experiments on listening to two simultaneous speakers showed that error reduction rates by Bi-HBSS are superior to by direction-pass filter [Nakatani and Okuno, 1999]. In this case, the azimuth between two speakers is 60°.

   This is partially because direction-pass filter does not allocate signals exclusively to some sound source unlike Bi-HBSS, and partially because error reduction rates are saturated for more than 30° for direction-pass filter.

   Therefore, we need to integrate Bi-HBSS and direction-pass filter to attain more error reductions.

### 4.5 Improvement of Visual Tracking by Using Audition

Visual tracking is reciprocally improved by using audition, although this paper focuses on the improvement of sound source separation by using vision [Nakagawa *et al.*, 1999].

The benchmark is to track each face of three persons sitting at the table. A face is considered as the sound source. Tracking accuracy of a simple color-based tracking system is measured under the following four conditions:

1. No constraints (purely rely on cluster of color) are assumed.
2. Human heights are used to exclude non-face parts.
3. rough direction of sound source, i.e. a face, from the microphone. A range of 20o is given to each face.
4. Both height constraints and rough direction of sound sources are used.

Tracking errors are more reduced as incorporating more constraints, vison or/and audition. In particular, combining rough direction information and height constraints drastically improve accuracy of the tracking.

## 5 Discussions

### 5.1 Theoretical Implication

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. 7 (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. 7 (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. 7 (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.

### 5.2 Do human use vision for auditory stream separation?

There is an interesting question which came to our mind. A series of experiments indicate that the use of multimodal sensory inputs improves three sound source separation. Similar phenomena is also observed in human subject. Thus, it can be inferred that there may be an interaction between the cortical area which determine sound source location and auditory stream separation. In addition, location information derived visual cortex may be delivered to cortical area involved in sound source localization or stream separation. At this moment, we have no clue whether such interaction actually takes place in our brain. But, we are not surprised even if such interaction actually exists.

From engineering perspective, we may be able to further improve auditory stream separation by aligning additional microphone a vertical position (above two microphones). This enables the system to resolve vertical spatial relationship, which is not possible with two horizontally aligned microphones.

Similarly, with a binaural system, head position can be moved during auditory stream separation, which gives clue for positional information for vertical axis.

In fact, human continuously moved head during observation as such action make compensate vertical positional information.

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

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

[Boll, 1979] 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.

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

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

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

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

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

[Lyon, 1984] R. F. Lyon. Computational Models of Neural Auditory Processing. In *Proceedings of 1984*

*International Conference on Acoustics, Speech, and Signal Processing (ICASSP-84)*, IEEE, 1984.

[Morita, 1990] T. Morita. Separation of mixed voices by acoustic parameter optimization. *Trans. of Inst. of Electronics, Inf. and Comm. Engineers*, vol.J73A, No.10. 1990.

[Nakagawa *et al.*, 1999] Y. Nakagawa, H. G. Okuno, and H. Kitano. Using vision to improve sound source separation. In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI-99)*, page (*in print*). AAAI, 1999.

[Nakatani and Okuno, 1999] T. Nakatani and H. G. Okuno. Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication*, 27(3-4):209–222, 1999.

[Nakatani *et al.*, 1994] 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.

[Nakatani *et al.*, 1995] 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.

[Okuno *et al.*, 1996] 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 *Proceedings of 13th National Conference on Artificial Intelligence (AAAI-96)*, pages 1082–1089. AAAI, 1996.

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

[Okuno *et al.*, 1999] H. G. Okuno, T. Nakatani, and T. Kawabata. Listening to two simultaneous speeches. *Speech Communication*, 27(3-4):281–298, 1999.

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

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

[Schmids, 1986] R. O. Schmids. Multiple Emitter Location and Signal Parameter Estimation. *IEEE Trans. on AP*, Vol.AP-34, No.3. 1986.

[Stadler and Rabinowitz, 1993] R. W. Stadler, and W. M. Rabinowitz. On the potential of fixed arrays for hearing aids. *Journal of Acoustic Society of America*, **94**(3) Pt.1:1332–1342, 1993.

[Sugie *et al.*, 1988] N. Sugie, J. Huang, and N. Ohnishi. Localizing sound source by incorporating biological auditory mechanism. In *Proceedings of IEEE International Conference on Neural Networks*, 1988.

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

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