Combining Independent Component Analysis and Sound Stream Segregation

Hiroshi G. Okuno[†*], **Shiro Ikeda**[‡], **and Tomohiro Nakatani** [§]

† 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 Science, Science University of Tokyo
‡ "Information and Human Activity", PRESTO, Japan Science and Technology Corp.
§ Provisional Headquaters, Nippon Telegraph and Telephon East Corp.
okuno@nue.org, shiro@brain.riken.go.jp, nak@mbd.mbc.ntt.co.jp

Abstract

This paper reports the issues and results of AI Challenge: "Understanding Three Simultaneous Speeches". First, the issues of the Challenge are revisited. We emphasis the importance of information fusion of various attributes of speeches (sounds) in separating speeches from a mixture of sounds. This emphasis is supported by comparing two methods of speech separation; *computational auditory scene analysis approach* that employs the attributes of sound sources and sound transmitting channel, and *blind source separation approach* that dispenses with these attributes. Although these two approaches are usually considered as opposite with regards to whether sound attributes is used or not, we conclude that they differ in the ways of using sound attributes. Next, a new algorithm for information fusion is proposed. Sound attributes extracted by tracking harmonic structures and sound source directions as well as by independent component analysis are fused according to sound ontology. Finally, the error reduction rate of the 1-best/10-best word recognition of each speaker performed on 200 mixtures of two women's and one man's utterances of an isolated word is reported.

1 Introduction

"Listening to several things simultaneously", or "Price Shotoku" computer after Japanese legend, is a next goal to automatic speech recognition systems in AI-related audition research [Cooke *et al.*, 1993; Nakatani *et al.*, 1994; Okuno *et al.*, 1995; Rosenthal and Okuno, 1998], since automatic speech recognition systems have been widely available recently on personal computers and some commercial personal computers come equipped with automatic speech recognition system. At a crowded party, people can attend one conversation and switch to another. This capability is well known as "cocktail party effect". However, psychoacoustic observation proves that people with normal hearing capability can listen to at most two things simultaneously [Kashino and Hirahara, 1996]. Therefore, "listening to several things simultaneously" is a challenging problem in audition research, and is expected to augment human's auditory capability or create new application areas in sound processing, in particular, sound recognition systems.

The AI Challenge "Listening to Three Simultaneous Speeches" (hereafter, *the Challenge*) proposed at IJCAI-97 by Okuno *et al.* [Okuno *et al*, 1997] is important in modeling computer audition, because most researches in sound processing assume some specific sounds as input. This assumption may not hold in the real-world, because we usually hear a mixture of sounds, not a single sound. The Challenge is restated as follows:

- **Speakers**: The distance between a microphone and a speaker is at least 1.4m. Three speakers utter a word simultaneously. They do not have to start at once.
- **Microphones**: At most two microphones are used.
- **Performance Measurement**: Error reduction rates for the 1-best and 10-best of recognition.
- **System Design**: The system should be extensible and adaptive for future challenging problems such as moving talkers or additional speakers.

There may be two main approaches to attack the Challenge, that is, understanding three simultaneous speeches. One is the *cascaded approach* which first separates each speech from a mixture of sounds and then recognize each separated speech by automatic speech recognition system. The other is the *integrated approach*

which exploits speech separation and speech recognition in concurrent and integrated ways [Cooke *et al.*, 1993; Ellis, 1996; Lesser *et al.*, 1993].

In this paper, we focus on the cascaded approach toward the Challenge. The cascaded approach consists of two main processes; one is to separate a speech from a mixture of sounds and the other is to recognize each separated speech by automatic speech recognition system.

Okuno *et al.* [Okuno *et al*, 1997] assume the cascaded approach and investigate the issues of the Challenge from the following aspects:

I-1. *robust automatic speech recognition* [Luo and Denbigh, 1994]

I-2. *signal processing* [Murata and Ikeda, 1998]

I-3. *general sound understanding, or computational auditory scene analysis (CASA)*

I-4. *psychoacoustics* [Kashino and Hirahara, 1996]

Since the Challenge involves understanding arbitrary sound which is one of the main topics of computational auditory scene analysis (CASA) [Cooke *et al.*, 1993; Nakatani *et al.*, 1994; Rosenthal and Okuno, 1998], it is also considered as a challenge for CASA.

Nakatani and Okuno point out another issue from the viewpoint of sound representation [Nakatani and Okuno, 1998].

I-5. *common sound representation, or, sound ontology.*

They proposed to use common representation of sounds, or *sound ontology*, in the following three purposes:

(1) to integrate sound separation systems such as speech and music,

(2) to interface sound separation systems with applications such as automatic speech recognition systems, and

(3) to integrate bottom-up and top-down processings in sound separation.

In this paper, we point out another issues in sound representation and utilization of sound source attributes, and present an integrated systems that fuses sound attributes extracted by different sound separation systems. The proposed system is then applied to the Challenge and the results are reported.

The rest of this paper is organized as follows: Section 2 revisits the issues of the Challenge. Section 3 presents the integrated systems to attack the Challenge, and Section 4 shows preliminary results. Discussion and Concluding remarks are given in Section 5 and 6.

2 Research Issues Revisited

Consider the additional issue for speech separation:

I-6. *role of sound attributes in sound source separation*

There have been a lot of dispute on whether sound attributes play an important role in separating speeches from a mixture of sounds [Rosenthal and Okuno, 1998].

Usually, CASA uses various sound attributes such as common onset, offset, AM (amplitude modulation), FM (frequency modulation), formants, and sound source direction as clues to separate speeches from a mixture of sounds. For example, Nakatani *et al.* developed speech separation system, Bi-HBSS, based on CASA [Nakatani *et al.*, 1995]. Bi-HBSS extracts speech from a mixture of sounds as follows (see part of Fig 3):

- **Harmonic Fragment Extraction**:

  First, it extracts all harmonic fragments by tracing harmonic structures of the input signal and the direction of sound sources. It also calculates the residue by subtracting all extracted harmonic fragments in a wave form.

- **Harmonic Grouping**:

  Next, harmonic fragments are grouped according to the proximity of fundamental frequency and the direction of sound sources.

  The first two steps reconstruct harmonic structure of original sound.

- **Residue Substitution**:

  Finally, speech is reconstructed by substituting the residue for non-harmonic parts.

Extensive experience with Bi-HBSS on various kinds of benchmarks has given the following observations:

(1) The criteria of the proximity for harmonic grouping depend on the allocation of speakers and the characteristics of successive processing.

  Usually the proximity in the direction of speakers plays an important role. more than 45°. However, the sensitivity or resolution of direction is about 20° [Nakatani *et al.*, 1995], because the direction of speakers is calculated by the interaural difference in time and intensity. When more speakers utter at different places, the difference of direction becomes less than 20° and thus directional information becomes useless.

(2) Which part of the residue is used for residue substitution depends on the characteristics of successive applications [Nakatani and Okuno, 1998].

Therefore, additional framework is needed to apply Bi-HBSS to the Challenge.

Blind source separation, or, *independent component analysis* is an opposite approach to CASA, and solves the problem by signal processing techniques. It is often said that blind source separation does not use such auditory clues. Murata and Ikeda invented a new algorithm "online algorithm" for blind source separation and applied it to separate each speeches from a mixture of two speeches with successful results [Murata and Ikeda, 1998].

2.1 Blind Source Separation

Blind source separation is sketched roughly. Let source signals consisting of n components (sound

sources) be denoted by the vector (1), and observed signals by n sensors (microphones) be denoted by the vector (2) specified as below:

$$\boldsymbol{s}(t) = (s_1(t), \cdots, s_n(t))^T, \quad t = 0, 1, 2, \ldots \tag{1}$$

$$\boldsymbol{x}(t) = (x_1(t), \cdots, x_n(t))^T, \quad t = 0, 1, 2, \ldots \tag{2}$$

Each component of $\boldsymbol{s}(t)$ is assumed to be independent of each other, that is, the joint density function of the signals is factorized by their marginal density function

$$p(s_1(t), \cdots, s_n(t)) = p(s_1(t)) \times \cdots \times p(s_n(t)).$$

In addition, observations are assumed to be linear mixtures of source signals:

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t)$$

Note that A is an unknown linear operator.

Let $a_{aj}(\tau)$ be a unit impulse response from source j to sensor i with time delay τ. The observation at sensor i can be represented as

$$\boldsymbol{x}(t) = \left(\sum_k a_{ik} * s_k(t) \right),$$

$$\text{where, } a_{ik} * s_k(t) = \sum_{r=0}^{\tau_{max}} a_{ik}(\tau) * s_k(t - \tau)$$

Thus, A can be represented in matrix form as

$$A(t) = \begin{pmatrix} a_{11}(t) & \cdots & a_{1n}(t) \\ \vdots & \ddots & \vdots \\ a_{n1}(t) & \cdots & a_{nn}(t) \end{pmatrix}.$$

The goal of blind source separation is to find a linear operator $B(t)$, such that the components of reconstructed signals

$$\boldsymbol{y}(t) = B * \boldsymbol{x}(t)$$

are mutually independent, *without* knowing the operator $A(t)$ and the probability distribution of source signal $\boldsymbol{s}(t)$.

Ideally we expect $B(t)$ to be the inverse operator of $A(t)$, but there remains indefiniteness of scaling factors and permutation due to lack of information on the amplitude and the order of the source signals.

"On-line ICA (Independent Component Analysis)" algorithms [Murata and Ikeda, 1998] separates source signals from a mixture of signals in the following steps:

(1) First, mixed signals are converted to the spectrogram, or to time-frequency domain. That is, the windowed-Fourier transformation is applied to observed signals by shifting Hamming window of 128 points.

(2) Then, blind source separation algorithm is applied to each frequency channel independently. That is, on-line ICA (Independent Component Analysis) is applied to the frequency components of the non-symmetric 65 points.

(3) Next, the correspondence of separated components in each frequency is determined based on temporal structure of signals.

Since the output of ICA carries ambiguities in permutation of the frequent components and in the amplitudes, the permutation of components is determined on the basis of correlation between their envelops.

(4) Finally, separated spectrogram of the source signals is constructed.

Now, we apply on-line ICA to the same benchmark sets as Okuno *et al.* [Okuno *et al.*, 1996] to get better knowledge on two opposite approaches.

2.2 Preliminary Experiments

To evaluate the performance of separation, we adopt the same three benchmark sets that are used by Okuno *et al.* [Okuno *et al.*, 1996]. The first benchmark set, called **Double**, consists of 500 two-sound mixtures of women's utterance of Japanese words. The first speaker utters at 30° to the left from the center and the second speaker utters at 30° to the right from the center. The second benchmark set, called **Triple**, consists of three sounds, that is, two sounds used in the first benchmark set and additional intermittent harmonic sounds that comes from the center. The power of additional sound is about half of the average power of two women's utterances. The third benchmark set, called **Triple'**, differs from the second one in that the power of additional sound is almost the same as the average power of two women's utterance. Each mixed sound is recorded by a pair of binaural microphones in two channels. Sampling rate is 12KHz and data size is 16 bit. Most mixed sounds are created by using Head-Related Transfer Functions.

We also use the same automatic speech recognition system, called HMM-LR, a hidden Markov Model based system [Kita *et al.*, 1990; Okuno *et al.*, 1996]. HMM-LR is also used for the Challenge.

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 $\mathcal{E}^{(n)} = \mathcal{CA}^{(n)}_{org} - \mathcal{CA}^{(n)}_{mix}$.

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

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

Figure 1 and Figure 2 show the error reduction rates for the 1-best and 10-best recognition by Bi-HBSS (created from the data in [Okuno *et al.*, 1996]), and blind source separation, respectively.

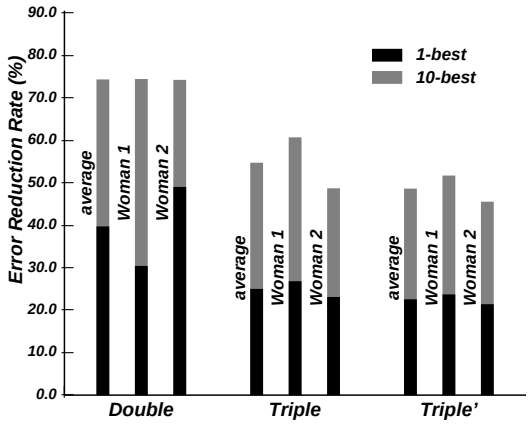

Figure 1: Error reduction rates for the 1-best/10-best recognition of each speech by Bi-HBSS

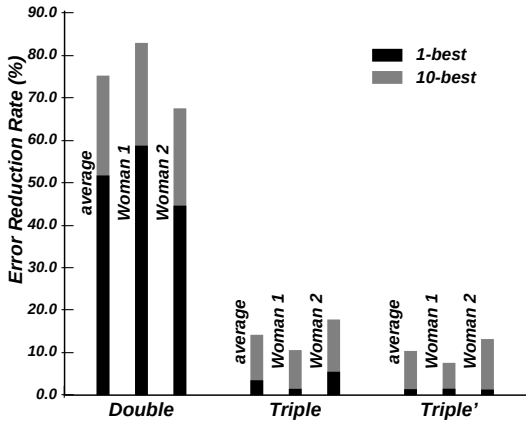

Figure 2: Error reduction rates for the 1-best/10-best recognition of each speech by Blind Source Separation

2.3 Observations

(1) Blind source separation gives better error reduction rates for the benchmark set **Double**, but quite worse ones for **Triple** and **Triple'**. These poor results on the benchmark of three sounds are quite reasonable, because blind source separation assumes that the number of sound sources is equal to or less than the number of microphones.

Another restriction of blind source separation, that is, sound sources are independent of each other, does not influence on these benchmark sets. However, when a mixture of musical signals that contains harmony are manipulated by blind source separation, we must be careful about this restriction.

(2) Blind source separation needs attributes of sounds to dissolve ambiguities of the order of independent components and their amplitudes to reconstruct each original sound.

In other words, simple signal processing, or mathematical treatment is not enough to separate sounds from a mixture of sounds.

(3) The essential issue of **I-6** can be paraphrased that *what kinds of attributes should be used to reconstruct original sounds?*

In blind source separation, common amplitude modulation (AM) is used to reconstruct original sounds. This reconstruction process is considered similar to harmonic grouping of Bi-HBSS. Bi-HBSS uses fundamental frequency (pitch) and the direction of sound source as the criteria of grouping.

(4) The recognition rate of separated speeches by blind source separation is not so affected by the distance between speakers unlike the case of Bi-HBSS.

(5) To fuse sound attributes, common sound representation is required as Nakatani and Okuno pointed out [Nakatani and Okuno, 1998].

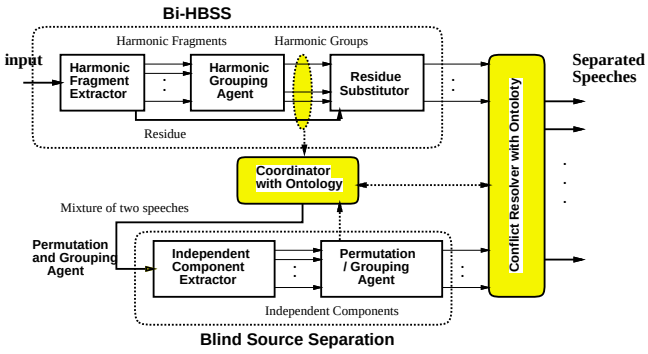

Figure 3: Integrated Systems with Bi-HBSS and Blind Source Separation

3 System Integration

To attack the Challenge, Bi-HBSS is not only upgraded but also integrated with sound source separation.

First, the longer-term analysis of harmonic structure and direction of sound source is incorporated in the tracking mechanism of Bi-HBSS. The direction of a fragment is defined as stable when the direction is continuously (for more than 75 ms) within a limited region (within 0.41 ms). Once a stable direction is obtained for a fragment, the most recently stable direction becomes the direction of the fragment at the time-frame. This long-term analysis leads to a reliable direction information, because sound interference often makes a short-term analysis erroneous. In addition, frequency modulation (FM) of each component of a harmonic structure is analyzed so that the harmonic structure becomes more accurate. These modifications are motivated by the fact that blind source separation uses common AM to reconstruct speech signals. The details are described in [Nakatani and Okuno, 1999].

Second, Bi-HBSS and blind source separation are in-

tegrated to exploit each merits and overcome each weak points. The idea is very simple. Since blind source separation can separate better than Bi-HBSS for a mixture of two sounds, Bi-HBSS generates a mixture of sounds. The whole system is depicted in Fig 3.

The flow of processing is roughly sketched as below:

(1) When Bi-HBSS gets input signals, its Harmonic Fragment Extractor extracts harmonic fragments, which Harmonic Grouping Agent groups to harmonic groups.

(2) A newly designed agent, Coordinator, always watches the processing of Harmonic Grouping Agent and bookkeeping information on harmonic groups. When Harmonic Grouping Agent finishes all processing, Coordinator generates a mixture of two speeches, and gives it to blind source separation. This mixture usually consists of the latestly separated two speeches. As described in Section 2, Independent Component Extractor extracts independent components and Permutation/Grouping Agent calculates a correct combination of independent components and reconstructs speeches. In this stage, information on independent components is fed back to Coordinator, which bookkeeps the information. Since the information supplied to Coordinator may have different formats, Coordinator converts it to a standard format by using ontology [Nakatani and Okuno, 1998].

(3) Finally, speeches separated by Bi-HBSS and blind source separation are given to Conflict Resolver, which checks whether speech separated by blind source separation has a corresponding speech separated by Bi-HBSS. If found, Bi-HBSS's output is adopted. Otherwise, Conflict Resolver calls Harmonic Grouping Agent to do regrouping according to blind source separation.

Since a mixture of two speeches Coordinator gives to blind source separation may contains errors, the system pays more respects to Bi-HBSS.

Note that their integration is possible in spite of their opposite approaches, because there are many common functionalities between these two systems.

4 Experiments

The benchmark set used for the evaluation 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]. Of course, a small set of benchmarks were actually recorded in an anechoic room. We confirmed that the synthesized and actually recorded data do not cause a significant difference in speech recognition evaluation.

(1) All speakers are located at about 2m from a pair of microphones installed on a dummy head.

(2) The first speaker is a woman located at 30° to the left from the center.

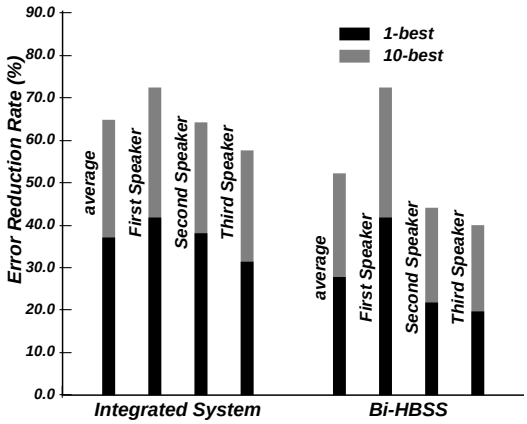

Figure 4: Error reduction rates for the 1-best/10-best recognition of each speech by Integrated System and Bi-HBSS

(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. For most mixtures, the power is reduced by 2 to 3 dB.

The error reduction rates for the 1-best and 10-best recognition of each speech by integrated system and Bi-HBSS are shown in Fig 4. By integrating blind source separation, the error reduction rates for speaker 2 and 3 are improved. Since these two speakers are within 30° from the microphones and the sensitivity of the direction in Bi-HBSS is about 20°, the direction of speakers is sometimes incorrect and thus the error of recognition is not recovered so much. In average, 37.1% and 64.8% of errors in recognition is reduced. The error reduction rates by Bi-HBSS is better than that in Fig 1, since new tracing mechanism is incorporated in Bi-HBSS.

5 Discussion and Future Work

(1) Integration of existing systems with monir revision can reduce 64.8% of errors in recognition in average. We think that the current Bi-HBSS designed with multi-agent systems [Nakatani *et al.*, 1995] and the architecture for integration with ontology, or common representation of sounds, is proved effective in audition research.

(2) One of the most important future work is to apply bottom-up and top-down processing to the Challenge. We are currently designing bottom-up and

top-down control for the Challenge based on Bi-HBSS.

(3) If the direction of speaker is precisely obtained, the quality of separated speeches can be improved by a filter function. Our preliminary experiments on the same benchmark set show that the error reduction rates for the 1-best and 10-best recognition of the first speaker are 58.8% and 87.6%, respectively.

Directional information can be extracted by binaural input [Blauert, 1983; Bodden, 1993] or by microphone arrays [Inoue *et al.*, 1997; Stadler and Rabinowitz, 1993].

(4) Filter can extract one speech and the remaining signals are severely distorted. Therefore, this approach is more adequate to "cocktail party" computer than "Prince Shotoku" computer that can listen to several things simultaneously.

(5) Directional information can be obtained by vision. Fusing visual and auditory information is an interesting research theme. By detecting a new auditory event by vision or audition, a camera moves toward the sound source or auditory system extracts sound that comes from the specific direction. Multi-modal cocktail party computer is an important and exciting future work.

(6) Other future work includes application of the Challenge, and design of more universal media ontology to make integration of visual aud auditory processings easier and more feasible.

(7) If the Challenge is revised as follows, another interesting research issues emerge.

> *How much the error rates for recognition of separated speeches can be reduced by using vision?*

Okuno *et al.* reported that incorporating visual information on the sound source direction improves the error reduction rates [Okuno *et al*, 1999]. Nakagawa *et al.* also reported that visual tracking is also improved by using auditory information on the sound source direction [Nakagawa *et al.*, 1999].

6 Conclusion

In this paper, we present the role of sound attributes as an additional issue concerning the Challenge "Understanding Three Simultaneous Speeches". There often says that computational auditory scene analysis exploits sound attributes in separating sounds from a mixture of sounds, while blind source separation does not. We investigate blind source separation and show that blind source separation also uses sound attributes in separating speeches and thus it is possible to integrate both systems for speech separation. Then we integrate Bi-HBSS and blind source separation systems to attack the Challenge. The average error reduction rates for the 1-best and 10-best recognition in the Challenge is 37.1% and 64.8%, respectively. We believe that our results will encourage AI community to engage in audition research more actively.

Acknowledgments

We thank Dr. Hiroaki Kitano, the project director of Kitano Symbiotic Systems Project, Dr. Takeshi Kawabata of NTT Cyber Space Laboratories, Drs. Hiroshi Murase and Kunio Kashino of NTT Communication Science Labs, for their valuable discussions.

References

[Blauert, 1983] J. Blauert. *Spatial Hearing: the Psychophysics of Human Sound Localization.* MIT Press, 1983.

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

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

[Ellis, 1996] D. P. W Ellis. Prediction-driven computational auditory scene analysis, PhD thesis, MIT Media Lab, 1996.

[Inoue *et al.*, 1997] Masaaki Inoue, T. Yamada, Satoshi Nakamura, and Kiyohiro Shikano. Microphone Array Design Measures for Hands-Free Speech Recognition *Proceedings of EUROSPEECH 97*, pages 331–334, 1997.

[Kashino and Hirahara, 1996] M. Kashino, and T. Hirahara. One, two, many – Judging the number of concurrent talkers. *J. of Acoustical Society of America*, 99 (4) Pt.2, 2596.

[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.

[Lesser *et al.*, 1993] V. Lesser, S.H. Nawab, I. Gallastegi,and F. Klassner. IPUS: An Architecture for Integrated Signal Processing and Signal Interpretation in Complex Environments. *Proceedings of Eleventh National Conference on Artificial Intelligence*, 249–255, AAAI, 1993.

[Luo and Denbigh, 1994] H.Y. Luo, and P.N. Denbigh. A speech separation system that is robust to reverberation, In *Proceedings of International Conference on Speech, Image Processing and Neural Networks*, vol.1:339-42, IEEE, 1994.

[Minami and Furui, 1995] Y. Minami, and S. Furui. A Maximum Likelihood Procedure for A Universal Adaptation Method based on HMM Composition. *Proceedings of 1995 International Conference on Acoustics, Speech and Signal Processing*, vol.1:129–132, IEEE, 1995.

[Murata and Ikeda, 1998] N. Murata and S. Ikeda. An Online Algorithm for Blind Source Separation on Speech Signals, *Proceedings of 1998 International Symposium on Nonlinear Theory and its Applications*, pages 923–927, 1998.

[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)*, (*in print*). AAAI, 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. *Proceedings of 12th National Conference on Artificial Intelligence (AAAI-94)*, pages 100–107, AAAI.

[Nakatani *et al.*, 1995] T. Nakatani, H. G. Okuno, and T. Kawabata. Residue-driven architecture for Computational Auditory Scene Analysis. *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, vol.1:165–172, IJCAI. 1995.

[Nakatani and Okuno, 1998] T. Nakatani and H. G. Okuno. Sound Ontology for Computational Auditory Scene Analysis. In *Proceedings of 15th National Conference on Artificial Intelligence*, pages 1004–1010, AAAI, July 1998.

[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.

[Okuno *et al.*, 1995] H. G. Okuno, T. Nakatani, and T. Kawabata. Cocktail-Party Effect with Computational Auditory Scene Analysis — Preliminary Report —. *Symbiosis of Human and Artifact* vol.2:503–508, Elsevier, 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)*, Vol.1:30–35, IJCAI, 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.

[Okuno *et al*, 1999] H. G. Okuno, Y. Nakagawa, and H. Hitano. Incorporating Visual Information into Sound Source Separation. In *Working Notes of IJCAI Workshop on Computational Auditory Scene Analysis (CASA'99)*, Aug. 1999.

[Ramalingam, 1994] C.S. Ramalingam and R. Kumaresan. Voiced-speech analysis based on the residual interfering signal canceler (RISC) algorithm. In *Proceedings of 1994 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-94)*, pp.473–476, IEEE, 1994.

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

[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.