Understanding Three Simultaneous Speeches

Hiroshi G. Okuno, Tomohiro Nakatani, and Takeshi Kawabata
NTT Basic Research Laboratories
3-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-01, Japan
okuno@nue.org, nakatani@horn.brl.ntt.co.jp, kaw@idea.brl.ntt.co.jp

Abstract

Understanding three simultaneous speeches is proposed as a challenge problem to foster artificial intelligence, speech and sound understanding or recognition, and computational auditory scene analysis research. Automatic speech recognition under noisy environments is attacked by speech enhancement techniques such as noise reduction and speaker adaptation. However, the signal-to-noise ratio of speech in two simultaneous speeches is too poor to apply these techniques. Therefore, novel techniques need to be developed. One candidate is to use speech stream segregation as a front-end of automatic speech recognition systems. Preliminary experiments on understanding two simultaneous speeches show that the proposed challenge problem will be feasible with speech stream segregation. The detailed plan of the research on and benchmark sounds for the proposed challenge problem is also presented.

1 Introduction

Recently emerges a new research on understanding arbitrary sound mixtures including non-speech sounds and music. Their understanding represents a challenging and little-studied area of artificial intelligence, automatic speech recognition/understanding, and signal processing. This interdisciplinary research area is called *computational auditory scene analysis* (hereafter, CASA).

At a crowded party, one can attend one conversation and then switch to another one. This phenomenon is known as the *cocktail party effect* [Cherry, 1953]. As seen in the cocktail-party effect, humans have the ability to selectively attend to sound from a particular source, even when it is mixed with other sounds. Current automatic speech recognition systems can understand *clean* speech well in relatively noiseless laboratory environments, but break down in more realistic, noisier environments.

Computers also need to be able to decide which parts of a mixed acoustic signal are relevant to a particular purpose – which part should be interpreted as speech, for example, and which should be interpreted as a door closing, an air conditioner humming, or another person interrupting. CASA focuses on the computer modeling and implementation for the understanding of acoustic events

The research topics concerning CASA include modeling, signal processing, sound representational, control and system architecture, and applications as well as sensor integration. Some of these topics were discussed at the IJCAI-95 workshop on Computational Auditory Scene Analysis [Rosenthal and Okuno, 1997].

At the AAAI-96, the panel entitled "Challenge Problems for Artificial Intelligence", Brooks proposed two problems concerning sounds [Selman *et al.*, 1996]:

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

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

We propose the problem of *Understanding Three Simultaneous Speeches*[1] (hereafter, *the challenge*) as a challenge problem for artificial intelligence, in particular, for CASA. A computer capable of listening to several things simultaneously is called *Prince Shotoku Computer* after the Japanese legendary that Prince Shotoku (A.D. 574–622) could listen to ten people's petitions simultaneously [Okuno *et al.*, 1995]. Since psychoacoustic studies have recently showed that humans cannot listen to more than two things simultaneously [Kashino and Hirahara, 1996], CASA research would make computer

[1] The selection of the word "*simultaneous*" or "*concurrent*" is controversial. The former carries more physical senses, while the latter carries more mental senses; e.g., "separation of simultaneous talkers", "simultaneous voices separation", and "separation of concurrent sentences" make sense. We adopt "simultaneous" because the proposed challenge problem won't pursuit understanding what each speaker talks about. Understanding what a speaker says without speech recognition, for example, is beyond our problem.

audition more powerful than human audition, similar to the relationship of an airplane's ability to that of a bird.

The rest of this paper is organized as follows: Section 2 explains the research issues for the challenge, in particular, its relevance and significance to AI. Section 3 presents its feasibility by showing the result of preliminary experiments on understanding two simultaneous speeches with/without interfering sounds. Section 4 discusses the detailed plan of our challenge problem. Concluding remarks are given in Section 5.

2 Research Issues for Understanding Three Simultaneous Speeches

In this section, we explain the reasons why we focus on three simultaneous speeches, not two simultaneous speeches and how significant the challenge is to AI researches. We also discuss several research issues involved in realizing understanding three simultaneous speeches. The main research areas related to the challenge are automatic speech recognition (ASR), signal processing, speech understanding, computational auditory scene analysis, and psychoacoustics.

2.1 Automatic Speech Recognition (ASR)

At present, one of the hottest topics of ASR research is how to make ASR systems more robust so that they can perform well outside *laboratory conditions* [Hansen *et al.*, 1994]. Conventional approaches for robust ASR are speech enhancement and many techniques for speech enhancement such as noise reduction and speaker adaptation have been developed [Hansen *et al.*, 1994; Minami and Furui, 1995].

One possible approach is to enhance a speech by employing noise reduction techniques. Once a speech is enhanced, it can be subtracted from a mixture of sounds in waveform. By repeating this procedure to the residue (remaining sounds), it seems possible to extract most speeches from a mixture of sounds.

This approach, however, works only up to two simultaneous speeches. The reason is as follows; Most conventional noise reduction techniques assume that the signal-to-noise ratio (SNR) of speech is 0 dB or better. The SNR of speech in a mixture of two simultaneous speeches is approximately 0 dB and thus noise reduction techniques can be applied to two simultaneous speeches. However, new techniques need to be developed for understanding three simultaneous speeches.

2.2 Signal Processing

Speech separation is more aggressive approach than noise reduction. Adaptive filters are used for speech separation [Ramalingam, 1994]. Spatial information on the sound source plays an important role in separating a speech from a mixture of sounds. This mechanism is called *localization*, which is performed by using a dummy head microphone (called *binaural sounds*) [Blauert, 1983; Bodden, 1993] or by using microphone arrays [Hansen *et al.*, 1994; Stadler and Rabinowitz, 1993]. For a pair of microphones, localization can be obtained better from binaural sounds than from stereo sounds.

Adaptive window technique for localizing two simultaneous voices by using two microphones is also developed for speech enhancement in real-time [Banks, 1993]. Procedures for enhancing the intelligibility of a target speaker (talker) in the presence of a simultaneous talker is developed by using harmonic selection and cepstral filtering [Stubbs and Summerfield, 1991]. Classification tasks within an automated two-speech separation system are performed by neural net [Roger *et al.*, 1989].

Most of these systems can separate a speech from a mixture of two simultaneous speeches. A speech separation system is developed by using harmonic structure and directional information and can extract one speech from a mixture of more than two overlapping speeches [Luo and Denbigh, 1994].

Since the spectrum of speech separated by speech separation techniques developed so far is distorted, they cannot be applied continuously to the remaining signals. We need to develop new techniques for the challenge. In addition, a segregated speech cannot be used as a input to automatic speech recognition systems due to spectral distortion. We also need to develop an interfacing technique between speech separation and ASR.

2.3 Computational Auditory Scene Analysis (CASA)

Speech enhancement technologies developed so far focus on only one speech and treat other speeches or sounds as noise. CASA takes an opposite approach. First, it deals with the problems of handling mixture of sounds to develop methods and technologies. Then it applies these to develop ASR systems that work in a real-world environment. The main research topic of CASA is *sound stream segregation*, a process that segregates sound streams that have consistent acoustic attributes from a mixture of sounds.

In extracting acoustic attributes, some systems assume the humans auditory model of primary processing and simulate the processing of cocklear mechanism [Brown, 1992; Slaney *et al.*, 1994]. Brown and Cooke designed and implemented the system that builds various auditory maps for input sounds and integrates them to segregate speech from input sounds [Brown, 1992; Brown and Cooke, 1992]. An auditory map represents acoustic attributes such as onset, offset, AM and FM modulations, and formants. Since the integration process becomes complicated when treating a mixture of sounds under the real-world environments, the blackboard architecture is used to simplify this integration process [Cooke *et al.*, 1993].

To design a more flexible and expandable system, control mechanisms are needed. IPUS (*Integrated Processing and Understanding Signals*) [Lesser *et al.*, 1993] integrates signal processing and signal interpretation into the blackboard system. IPUS has various interpretation knowledge sources which understand actual sounds such

as hair driers, footsteps, telephone rings, fire alarms, and waterfalls [Nawab, and Lesser, 1992].

Nakatani *et al.* took a multi-agent approach to sound stream segregation which extracts individual sound stream from a mixture of sounds by agents each of which traces harmonic structure with directional information [Nakatani *et al.*, 1994]. They use the Fourier transformation instead of the auditory model because the former is easy to implement and its properties are well analyzed.

2.4 Psychoacoustics

Psychoacoustic people have studied the human auditory mechanism extensively as auditory scene analysis [Bregman, 1990], but computer modeling has not been exploited yet. Emerging computational auditory scene analysis research focuses on computer modeling and has prompted interdisciplinary studies with psychoacoustic and AI and signal processing communities. In addition, our challenge problem has fostered psychoacoustic studies on how many simultaneous speeches human can listen to. Kashino *et al* claimed that human could listen to at most two things simultaneously by performing various experiments [Kashino and Hirahara, 1996]. If this is true, the challenge will attempt to make computer audition superior to human's capability of listening.

3 Preliminary Experiments

In this section, we demonstrate the feasibility of the challenge by describing the preliminary experiments on understanding two simultaneous speeches (up-to-date information of our AAAI-96 paper [Okuno *et al.*, 1996]). This problem is attacked by speech stream segregation, one of the main research topics of computational auditory scene analysis. The whole system consists of two components, speech stream segregation and speech recognition, as is shown in Figure 1.

First speech streams are extracted from a mixture of speeches, and then each speech stream is recognized by conventional automatic speech recognition system.

3.1 Speech Stream Segregation

Human voice consists of harmonic sounds such as vowel and voiced consonants, and non-harmonic sounds such as unvoiced consonants. By assuming the structure of "Vowel (V) + Consonant (C) + Vowel (V)" of speech, speech stream segregation is realized by the following two subprocesses:

(1) extracting and grouping harmonic stream fragments (*harmonic structure extraction*), and

(2) restoring non-harmonic parts by residue (*residue substitution*).

Rough flow of the computation is depicted in Figure 2.

Harmonic structures are extracted from a binaural input by the Bi-HBSS (Binaural Harmonics-Based Stream Segregation) system [Nakatani *et al.*, 1995; Nakatani *et al.*, 1996]. Bi-HBSS uses a harmonic structure and the

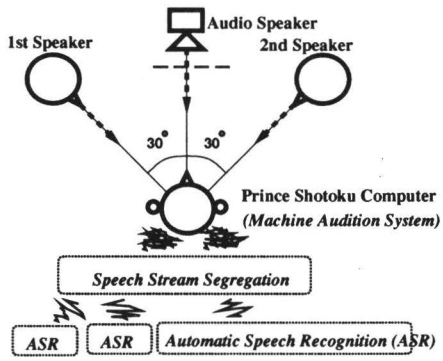

Figure 1: System architecture and sound sources for experiments on understanding two simultaneous speeches.

direction of sound source as cues of segregation. Bi-HBSS adopts a pair of HBSSes [Nakatani *et al.*, 1994] for the right and left channel to extract harmonic stream fragments. It determines the fundamental frequency (F_0) of a harmonic stream fragment by coordinating the pair of HBSSes. The direction of sound source is identified by calculating the interaural time difference (ITD) and interaural intensity difference (IID) of a pair of harmonic stream fragments of the same F_0 extracted by the pair of HBSS. Harmonic stream fragments are grouped by the direction of the sound source.

The residue obtained by subtracting harmonic structures from an input sound is substituted for non-harmonic parts of a group. If a group ends with non-harmonic parts, the residue is substituted for 150 msec. The idea of residue substitution is similar to the psychophysical observation known as *auditory induction* [Green *et al.*, 1995; Warren, 1970]. It is a phenomena that human listeners can perceptually restore a missing sound component if it is very brief and masked by appropriate sounds.

3.2 Automatic Speech Recognition

The automatic speech recognition system, HMM-LR [Kita *et al.*, 1990], is used to recognize speech streams. HMM-LR is based on hidden Markov model of each phonetic transition, in spite of Rodney Brooks' challenge problem. The parameters of HMM-LR are trained by a set of 5,240 words uttered by five speakers.

Since the spectrum of speech streams segregated by the speech stream segregation is distorted due to binaural input, binauralized training data is used to recover from the degradation of the performance of recognition [Okuno *et al.*, 1996].

3.3 Performance Evaluation

The performance of automatic speech recognition is usually measured by the *cumulative accuracy up to the 10th candidate* (or simply *cumulative accuracy*) of word recognition, since ASR returns the first about 10 candidates of each word. Such candidates are further selected by successive speech understanding systems. Therefore,

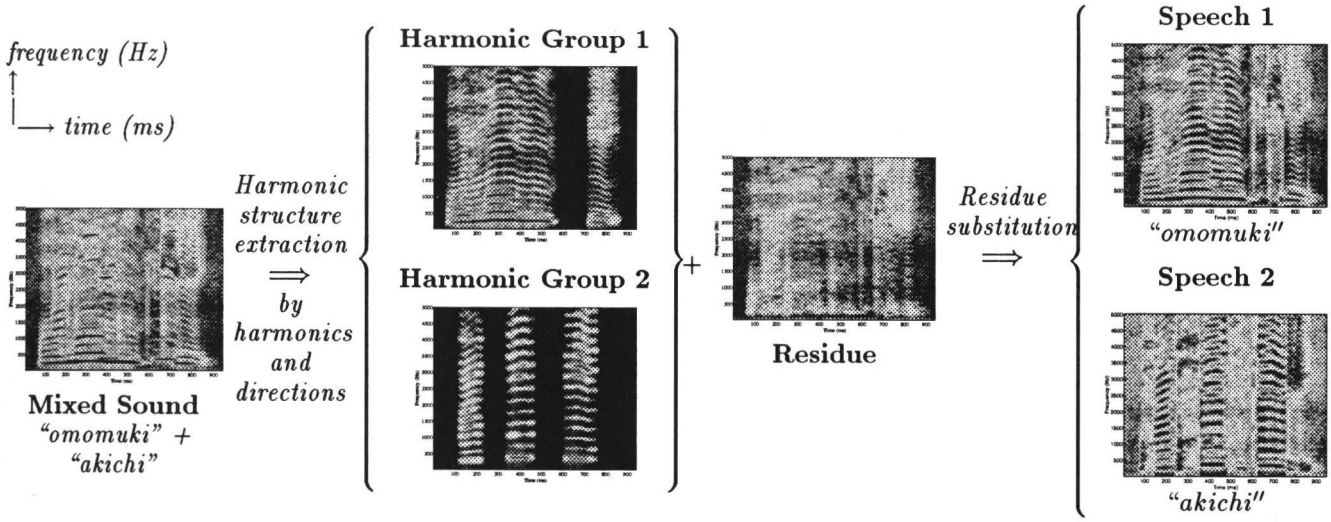

Figure 2: **Speech stream segregation**: From binaural sounds consisting of two Japanese words "*omomuki*" and "*akichi*", two harmonic groups are extracted by tracing harmonic structures and the direction of sound sources. The residue is also generated by subtracting harmonic groups from input. Then, the residue is substituted for missing non-harmonic parts (black parts in the spectrogram) of each group, and thus speech streams are generated.

Table 1: Three sets of benchmark sounds

No.	Speaker 1	Speaker 2	Audio Speaker
Double	Woman 1	Woman 2	—
Triple	Woman 1	Woman 2	Sound 1
Triple'	Woman 1	Woman 2	Sound 2

we adopted the same measurement with open tests. By open tests, we mean that the training and benchmark (testing) data are disjoint.

We used three sets of 500 benchmark sounds; one set of 500 two-sound mixtures and two sets of 500 three-sound mixtures (Table 1). The first sound is uttered by the first speaker at 30° to the left from the center, and the second sound is uttered after 150 msec by the second speaker at 30° to the right from the center (Figure 1). To recognize the first speech in the mixed sound directly by HMM-LR, the utterance of the second speaker is delayed by 150 msec.

The third sound with F_0 of 250 Hz is an intermittent harmonic sound from the center. It starts before the first speaker and repeats to last for 1 sec with 50 msec of pause. The average power ratios of the first and second sounds to the third sound in benchmarks **Triple** and **Triple'** are 1.7 dB and -1.3 dB, respectively.

The *error rate caused by interfering sounds* is defined as follows. Let the cumulative accuracy of recognition of original data up to the 10th candidate be $\mathcal{CA}_{org}$, and let the cumulative accuracy of recognition of (non-binaural) mixed sounds up to the 10th candidate be $\mathcal{CA}_{mix}$. The error rate caused by interfering sounds, $\mathcal{E}$, is calculated as $\mathcal{E} = \mathcal{CA}_{org} - \mathcal{CA}_{mix}$.

To evaluate the performance of speech stream segregation, *error reduction rate* is defined. Let the cumulative accuracy of recognition up to the 10th candidate be $\mathcal{CA}_{seg}$. The error reduction rate, $\mathcal{R}_{seg}$, is calculated as follows:

$$\mathcal{R}_{seg} = \frac{\mathcal{CA}_{seg} - \mathcal{CA}_{mix}}{\mathcal{CA}_{org} - \mathcal{CA}_{mix}} \times 100 = \frac{\mathcal{CA}_{seg} - \mathcal{CA}_{mix}}{\mathcal{E}} \times 100.$$

Table 2: Error rates in the word recognition caused by an interfering speaker without/with third sound.

Benchmark	Speaker 1	Speaker 2
Double	76.19%	95.50%
Triple	94.99%	95.70%
Triple'	94.99%	95.90%

The original cumulative accuracies of word recognition uttered by single speakers, Woman 1, and Woman 2, are 94.99%, and 96.10%, respectively. The error rate by interfering sounds is shown in Table 2.

Error reduction rates by speech stream segregation for the three benchmark sets are shown in Figure 3.3. The **Ideal** shows the upper limits of error reduction, which are calculated for the case in which the utterances of a single speaker are recognized after speech stream segregation. For **Double**, 77% of errors caused by an interfering speaker were reduced by speech stream segregation. By additional noise, the SNR of each speech is decreased further (by about 1 dB and 2 dB for **Triple** and **Triple'**, respectively), but, 55% and 49% of errors are reduced respectively.

Since this performance was attained without using any features specific to human voices, we believe that understanding three simultaneous speeches is a short-term research problem.

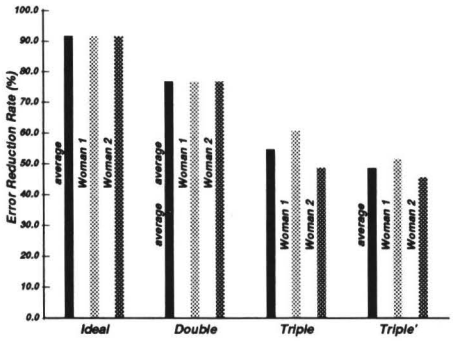

Figure 3: Error reduction rates for two speeches

4 Detailed Plan for the Challenge

The research issues depend on an approach taken by a challenger. Some possible approaches are listed below:

- Either speech separation system or speech stream segregation may be exploited.
- Speech separation or speech stream segregation may run either incrementally or in batch.
- Multiple speech enhancement or separation systems may run concurrently to extract all speeches or one system may extract all speeches.
- Speech segregation/separation system may be either used as a front-end to ASR or integrated with ASR.
- Top-down or hybrid approaches needed for continuous speech recognition or understanding may be employed, although word recognition is requested by the challenge.

We only give a general guideline on the benchmarks and evaluation criteria in this paper. Further information will be made available at the URL of `http://www.nue.org/CASA97/`.

4.1 Benchmark Sounds

The common platform for the challenge is quite important in order to share and transfer the methodology and technology developed by each challenger. Monaural data of speech used for the challenge should should be widely available. The current candidates are as follows:

- The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus for English speeches. It contains a total of 3,600 sentences by 360 speakers uttering the same 10 sentences. (http://www.ldc.upenn.edu/)
- The continuous speech corpus developed by the Acoustic Society of Japanese for Japanese speeches. It contains a total of 9,600 sentences by 64 speakers uttering some of 503 sentences, which were recorded by ATR.

Since these corpus are copyrighted, only the combination of word utterances will be made available. One benchmark set contains a total of 200 combination of words; three speakers, arbitrary combination of men and women, utter a different word simultaneously.

The acoustic field is made simple enough to produce benchmark sounds easily. A sound source (speaker) should be placed from 1.4 meters to 2 meters from the microphone on the floor in a free-field without reverberation. Several combinations of speaker positions selected from 0°, 30°, 45°, 60°, 90°, 120°, 135°, 150°, and 180° may be strongly recommended. The challenge does not assume that any speaker move during speaking. However, challengers may attack the problem of moving speakers.

A mixture of sounds may be either recorded or generated artificially. The number of microphones should be less than or equal to 3. If a challenger wants to use a binaural input, it may be generated artificially by using Head-Related Transfer Function (HRTF), which specifies the spectral transformation of a binaural sound. The data of HRTF for the KEMAR dummy head microphone is available from MIT [Gardner and Martin, 1994]. For this HRTF, a sound source should be placed 1.4 meters from the dummy head microphone. In our preliminary experiments, all sound sources were placed at the distance of 2 meters from the dummy head microphone.

4.2 Measurement of Evaluation

The measurement of performance evaluation is error reduction rate as well as cumulative accuracy of up to 10th candidate, both of which are defined in the previous section. The first measurement may be important because it is rather independent for automatic speech recognition systems used.

The first stage of the challenge investigates the performance of word recognition.

4.3 Tentative Schedule

- The challenge problem will be presented at IJCAI-97 as well as IJCAI-97 workshop on Computational Auditory Scene Analysis.
- The guideline on the benchmarks will be made available by the end of 1997.
- Intermediate progress reports will be presented at AAAI-98 or an appropriate conference in 1998.
- Final progress reports will be submitted in Jan., 1999 and will be presented at IJCAI-99.

5 Conclusions

In this paper, we proposed *understanding three simultaneous speeches* as a new challenge and standard AI problem. It provides rich research issues for a wide range of AI including automatic speech recognition, speech understanding, and CASA, as well as psychoacoustics. We expect that research on the challenge would play an important role in realizing the "Prince Shotoku Computer" or powerful computer audition systems.

References

[Banks, 1993] D. Banks. Localization and separation of simultaneous voices with two microphones. In *IEE Proceedings I*, Vol.140, No.4, pp.229-34, 1993.

[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.

[Bregman, 1990] A.S. Bregman. *Auditory Scene Analysis – the Perceptual Organization of Sound.* MIT Press, 1990.

[Brown, 1992] G.J. Brown. Computational auditory scene analysis: A representational approach. Ph.D diss., Dept. of Computer Science, University of Sheffield, 1992.

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

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

[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.

[Gardner and Martin, 1994] B. Gardner, and K. Martin. HRTF Measurements of a KEMAR Dummy-Head Microphone. *MIT Media Lab Perceptual Computing – Technical Report*, #280, May 1994. http://sound.media.mit.edu/KEMAR.html

[Green *et al.*, 1995] P.D. Green, M.P. Cooke, and M.D. Crawford. Auditory Scene Analysis and Hidden Markov Model Recognition of Speech in Noise. In *Proc. of 1995 International Conference on Acoustics, Speech and Signal Processing*, vol.1:401–404, IEEE, 1995.

[Hansen *et al.*, 1994] J.H.L. Hansen, R.J. Mammone, and S. Young. Editorial for the special issue on robust speech processing". *IEEE Transactions on Speech and Audio Processing* 2(4):549–550, 1994.

[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. In *Proc. 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 *Proc. 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. In *Proc. of 1995 International Conference on Acoustics, Speech and Signal Processing*, vol.1:129–132, IEEE, 1995.

[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 *Proc. of 12th National Conference on Artificial Intelligence*, 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 *Proc. of 14th International Joint Conference on Artificial Intelligence*, vol.1:165–172, IJCAI, 1995.

[Nakatani *et al.*, 1996] T. Nakatani, M. Goto, and H.G. Okuno. Localization by harmonic structure and its application to harmonic sound stream segregation. In *Proc. of 1996 International Conference on Acoustics, Speech and Signal Processing*, IEEE, 1996.

[Nawab, and Lesser, 1992] S.H. Nawab and V. Lesser. Integrated Processing and Understanding of Signals. In Oppenheim, A.V. and Nawab, S.H. (Eds.) *Symbolic and Knowledge-Based Signal Processing*, 251–285. Prentice-Hall, 1992.

[Nawab *et al*, 1995] S.H. Nawab, C.Y. Espy-Wilson, R. Mani, and N.N. Bitar. Knowledge-Based analysis of speech mixed with sporadic environmental sounds. In [Rosenthal and Okuno, 1997].

[Okuno *et al.*, 1995] H.G. Okuno, T. Nakatani, and T. Kawabata. Cocktail-Party Effect with Computational Auditory Scene Analysis — Preliminary Report —. In *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 *Proc. of 13th National Conference on Artificial Intelligence*, pp.1082–1089, 1996.

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

[Roger *et al.*, 1989] C. Rogers, D. Chien, M. Featherston, and K. Min. Neural network enhancement for a two speaker separation system. In *Proc. of 1989 International Conference on Acoustics, Speech and Signal Processing*, pp.357-60, IEEE, 1989.

[Rosenthal and Okuno, 1997] D. Rosenthal and H.G. Okuno (Eds.). *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates. Forthcoming.

[Selman *et al.*, 1996] B. Selman, R.A. Brooks, T. Dean, E. Horovitz, T.M. Mitchell, and N.J. Nilsson. Challenge Problems for Artificial Intelligence, In *Proc. of 13th National Conference on Artificial Intelligence*, pp.1340–1345, 1996.

[Slaney *et al.*, 1994] M. Slaney, D. Naar, and R.F. Lyon. Auditory Model Inversion For Sound Separation. In *Proc. of 1994 International Conference on Acoustics, Speech, and Signal Processing*, vol.2:77–80, IEEE, 1994.

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

[Stubbs and Summerfield, 1991] R.J. Stubbs and Q. Summerfield. Effects of signal-to-noise ratio, signal periodicity, and degree of hearing impairment on the performance of voice-separation algorithms, *J. of Acoustical Society of America*, 89(3):1383-93, 1991.

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