Sound Ontology for Computational Auditory Scene Analysis

Tomohiro Nakatani† and **Hiroshi G. Okuno**

NTT Basic Research Laboratories

Nippon Telegraph and Telephone Corporation

3-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-0198, JAPAN

okuno@nue.org

Abstract

This paper proposes that sound ontology should be used both as a common vocabulary for sound representation and as a common terminology for integrating various sound stream segregation systems. Since research on computational auditory scene analysis (CASA) focuses on recognizing and understanding various kinds of sounds, sound stream segregation which extracts each sound stream from a mixture of sounds is essential for CASA. Even if sound stream segregation systems use a harmonic structure of sound as a cue of segregation, it is not easy to integrate such systems because the definition of a harmonic structure differs or the precision of extracted harmonic structures differs. Therefore, sound ontology is needed as a common knowledge representation of sounds.

Another problem is to interface sound stream segregation systems with applications such as automatic speech recognition systems. Since the requirement of the quality of segregated sound streams depends on applications, sound stream segregation systems must provide a flexible interface. Therefore, sound ontology is needed to fulfill the requirements imposed by them. In addition, the hierarchical structure of sound ontology provides a means of controlling top-down and bottom-up processing of sound stream segregation.

Introduction

Sound is gathering attention as important media for multi-medal communications, but is less utilized as input media than characters or images. One reason is the lack of a general approach to recognize auditory events from a mixture of sounds. Usually, people hear a mixture of sounds, and people with normal hearing can segregate sounds from the mixture and focus on a particular voice or sound in a noisy environment. This capability is known as the *cocktail party effect* (Cherry 1953). Perceptual segregation of sounds, called *auditory scene analysis*, has been studied by psychoacoustic and psychophysical researchers for more than forty years. Although many observations have been analyzed and reported (Bregman 1990), it is only recently that researchers have begun to use computer modeling of auditory scene analysis.

This emerging research area is called *computational auditory scene analysis (CASA)* (Brown and Cooke 1992; Cooke *et al.* 1993; Nakatani, Okuno, and Kawabata 1994a; Rosenthal and Okuno 1998), and its goal is the understanding of an arbitrary sound mixture including non-speech sounds and music. Computers need to be able to decide which parts of a mixed acoustic signal are relevant to a particular purpose – which part should be interpreted as speech, for example, and which should be interpreted as a door closing, an air conditioner humming, or another person interrupting. CASA focuses on the computer modeling and implementation for the understanding of acoustic events

One of its main research topics of CASA is sound stream segregation. In particular, CASA focuses on a general model and mechanism of segregating various kinds of sounds, not limited to specific kinds of sounds, from a mixture of sounds.

Sound stream segregation can be used as a front-end for automatic speech recognition in real-world environments (Okuno, Nakatani, and Kawabata 1996). As seen in the cocktail-party effect, humans have the ability to selectively attend to a 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.

Speech enhancement is essential to enable automatic speech recognition to work in such environments. Conventional approaches to speech enhancement are classified as noise reduction, speaker adaptation, and other robustness techniques (Minami and Furui 1995). Speech stream segregation is a novel approach to speech enhancement, and works as the front-end system for automatic speech recognition just as hearing aids for hearing impaired people.

Of course, speech stream segregation as a front-end for ASR is the first step toward more robust ASR. The

†Current address: NTT Multimedia Business Department, *nak@mbd.mbc.ntt.co.jp*

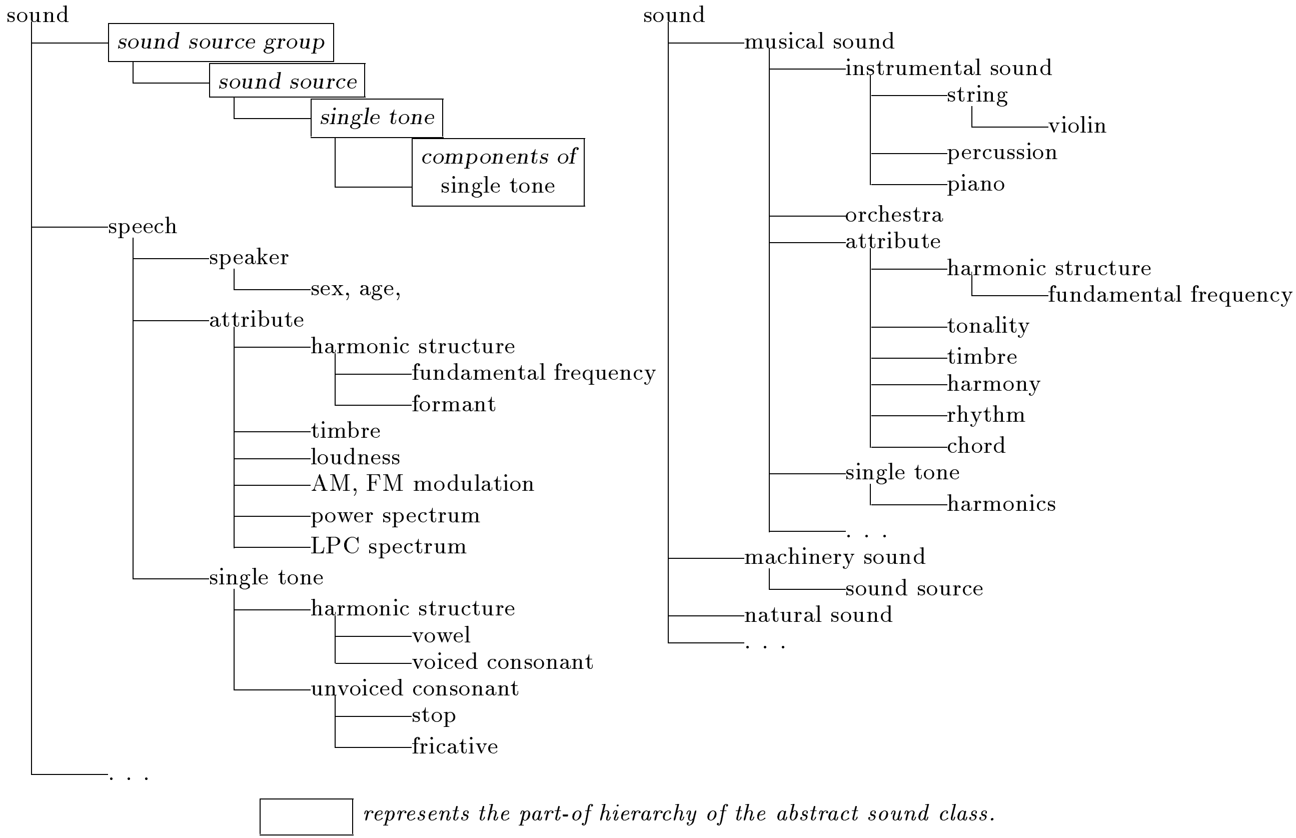


Figure 1: Example of sound ontology.

next step may be to integrate speech stream segregation and ASR by exploiting top-down and bottom-up processings.

Sound ontology is important in musical and speech stream segregation on the following aspects:

1. integrating sound stream segregation systems,
2. interfacing sound stream segregation systems with applications, and
3. integrating bottom-up and top-down processings in sound stream segregation.

For example, speech stream segregation system and musical stream segregation system are combined to develop a system that can recognize both speeches and music from a mixture of voiced announcements and background music. Although the both systems use a harmonic structure of sounds as a cue of segregation, harmonic structures extracted by one system can not be utilized by the other system because the definition of a harmonic structure and the precision of extracted harmonic structure differ. Therefore, sound ontology is needed to share the similar information by the both systems. In addition, sound ontology is expected to play an important role in making the system expandable and scalable.

The second example is one of AI challenges, that is, the problem of listening to three simultaneous speeches which is proposed as a challenging problem for AI and CASA (Okuno, Nakatani, and Kawabata 1997). This *CASA challenge* requires sound stream segregation systems to be interfaced with automatic speech recognition systems. In order to attain better performance of speech recognition, extracted speech streams should fulfill the requirements on the input of automatic speech recognition systems. Some automatic speech recognition systems use power spectrum as a cue of recognition, while others use LPC spectrum. Therefore, speech stream segregation systems should improve the property of extracted speech streams that is required by a particular automatic speech recognition system. If one sound stream segregation system is designed to be interfaced with these two kinds of automatic speech recognition systems, sound ontology may be used as a design model for such a system.

The third example is that speech stream segregation should incorporate two kinds of processings, primitive segregation of speech stream and schema-based segregation of speech streams. Primitive segregation may be considered as bottom-up processing by means of lower level properties of sound such harmonic structures, timbre, loudness, AM or FM modulation. Schema-based segregation may be considered as top-down processing by means of learned constraints. It includes a memory based segregation that uses a memory of specific speaker's voices, and a semantic based segregation that uses contents of speeches. The important issue is how to integrate the both processing and sound ontology may be used as a driving power of integration.

The rest of the paper is organized as follows. Section 2 presents a sound ontology and Section 3 discusses its usage with respect to the above-mentioned three issues. Section 4 presents the ontology-based integration of speech stream and musical stream segregation systems. Section 5 and 6 discuss related work and concluding remarks.

Sound Ontology

Sound ontology is composed of *sound classes*, *definitions of individual sound attributes*, and *their relationships*. It is defined hierarchically by using the following two attributes:

- **Part-of hierarchy** — a hierarchy based on the inclusion relation in sound
- **Is-a hierarchy** — a hierarchy based on the abstraction level in sound

Some part of sound ontology concerning speech and music is shown in Fig. 1. Other parts may be defined on demand of segregating corresponding sounds. In this paper, we focus on speech and musical stream segregation.

A box depicts a Part-of hierarchy of basic sound classes for speech and music, which is composed of four layers of sound classes. A *sound source* in the figure is a temporal sequence of sounds generated by a single sound source. A *sound source group* is a set of sound sources that share some common characteristics as music. A *single tone* is a sound that continues without any durations of silence, and it has some low-level attributes such as harmonic structure. In each layer, an upper class is composed of lower classes which are components that share some common characteristics. For example, a harmonic stream is composed of frequency components that have harmonic relationships.

The Is-a hierarchy can be constructed using any abstraction level. For example, voice, female voice, the voice of a particular woman, and the woman's nasal voice form an Is-a hierarchy. (Not specified in Fig. 1.)

With sound ontology, each class has some attributes, such as fundamental frequency, rhythm, and timbre. A lower class in the Is-a hierarchy inherits the attributes of its upper classes by default unless another special specification is given. In other words, an abstract sound class has attributes that are common to more concrete sound classes. In addition, an actually generated sound, such as uttered speech, is treated as an instance of a sound class. In this representation, segregating a sound stream means generating an instance of a sound class and extracting its attributes from an input sound mixture.

Proposed Usage of Sound Ontology

Ontology-Based Integration

Sound stream segregation should run incrementally, not in batch, because it is usually combined with applications and is not a stand-alone system. For incremental processing, we exploit not only sharing of information but also sharing of processing. Therefore, for integration of existing segregation systems they are first decomposed into processing modules by using the sound ontology as a common specification.

In this paper, we take an example of integrating speech and musical stream segregation. The rough procedure of modularization is as follows. First, existing segregation systems are divided into processing modules, each of which segregates a class of sound in sound ontology, such as, harmonic structure, voiced segment, and musical note. Then, these modules are combined to segregating streams according to their identified sound types.

Obviously, such an integration requires the procedure of interaction between different kinds of modules. To specify the relationships between sounds, a *relation class* is defined between two sound classes. It is represented by a pair of sound classes, such as "[speech, musical note]". A relation class has the same two hierarchical structures as sound classes defined as follows: if and only if both classes of a relation class are at a level higher than those of another relation class in the Is-a hierarchy (or in the Part-of hierarchy), the former is at a higher level. In the Is-a hierarchy, processing modules and interaction modules are inherited from a upper level to a lower level unless other modules are specified at some lower class.

Ontology-based sound stream segregation is executed by generating instances of sound classes and extracting their attributes. This process can be divided into four tasks:

1. find new sound streams in a sound mixture,
2. identify classes of individual sound streams,
3. extract attributes of streams according to their sound classes, and
4. extract interaction between streams according to their relation classes.

Since classes of sound streams are usually not given in advance, sound streams are treated initially as instances of abstract classes. These streams are refined to more concrete classes as more detailed attributes are extracted. At the same time, these streams are identified as components of stream groups at a higher level of the Part-of hierarchy, such as the musical stream. As a result, the attributes of these streams are extracted more precisely by using operations specific to concrete classes and by using attributes extracted for group streams.

Sound Ontology Based Interfacing

The CASA challenge requires speech stream segregation systems to interface with automatic speech recognition systems. First, speech stream segregation system that extracts each speech stream from a mixture of sounds. Then each extracted speech stream is recognized by a conventional automatic speech recognition system.

As such an approach, Binaural Harmonic Stream Segregation (Bi-HBSS) was used as a speech stream segregation (Okuno, Nakatani, and Kawabata 1996). By taking the structure of "Vowel (V) + Consonant (C) + Vowel (V)" of speech into consideration, a speech stream was extracted by the following three successive subprocesses:

1. Harmonic stream fragment extraction,
2. Harmonic grouping, and
3. Residue substitution.

The first two subprocesses reconstruct the harmonic parts of speech and calculates the residue by subtracting all extracted harmonic parts for the input. Since any major attributes for extracting non-harmonic parts have not been known yet, it is reasonable to substitute the residue for non-harmonic parts. Since Bi-HBSS takes binaural sounds (a pair of sounds recorded by a dummy head microphone) as inputs, it uses the direction of the sound source for segregation. They reported that the recognition performance with the Hidden Markov Model based automatic speech recognition system called HMM-LR (Kita, Kawabata, and Shikano 1990) is better when the residue of all directions is substituted than when the residue of the sound source direction (Okuno, Nakatani, and Kawabata 1996).

This interface is not, however, valid for another automatic speech recognition system. Our experiment with one of the popular commercial automatic speech recognition systems, HTK (Young *et al.* 1996), shows that the interface won't work well and that if the residue of the sound source direction is used by residue substitution the recognition performance is improved. That is, the reason why Bi-HBSS does not work well with HTM is that the cues of recognition used by HMM-LR and HTK differs. HMM-LR uses only power spectrum and ignores input signals of weak power, while HTK uses LPC spectrum and the power of input is automatically normalized. Therefore, weak harmonic structures included in the residue that are usually ignored by HMM-LR causes HTK's poor recognition performance.

Thus, speech stream segregation systems should generate an output appropriate to successive processing. Since the system architecture of sound stream segregation is constructed on the basis of sound ontology, adaptive output generation is also realized by the same architecture.

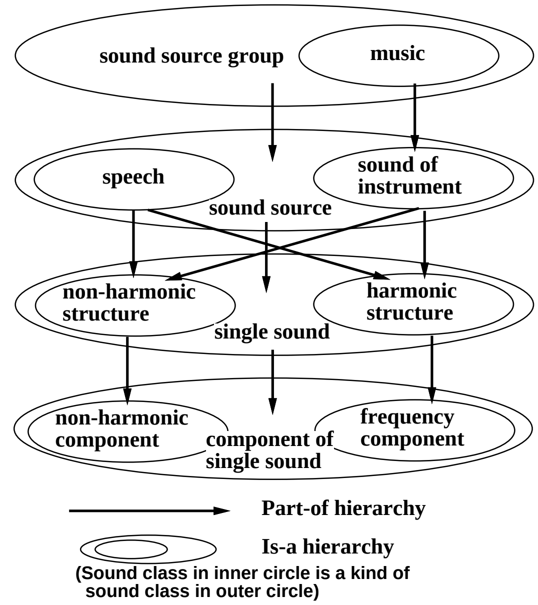


Figure 2: Hierarchical structure of sound ontology.

Integration of Bottom-up and Top-down Processing

The representation of single tone components also has analysis conditions under which the system extracts the attributes of individual components. This is because the quantitative meaning of sound attributes may differ by analysis means, such as FFT, auditory filter bank, LPC analysis, and by analysis parameters.

For example, interface agents use a common abstract terminology "harmonic structure" in exchanging information on voiced segments and on musical notes because voiced segments and musical notes both are a kind of harmonic structure based on sound ontology. Of course, voiced segments and musical notes do not entirely have the same attributes. In fact, the time patterns of voiced segments are much more complex than those of musical notes, so that some harmonic components appear and/or disappear according to their phoneme transitions, while the piano exhibits quite marked deviations from harmonicity during its attack period. Moreover, many frequency components of a musical note are often impossible to discriminate from those of other notes and thus are treated as a set of harmonic sounds, because they continuously overlap each other in a chord. These attributes should be handled somewhat differently in each system. Therefore, the specific sound classes of a harmonic structure (i.e., a voiced segment or a musical note) also have attributes specific to the classes.

Thus, sound ontology enables segregation systems to share information on extracted sound streams by providing common representation of sounds and correspondence between different analysis conditions.

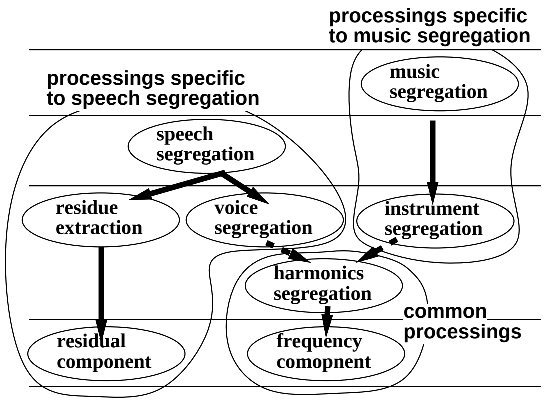


Figure 3: Common and specific processing modules for speech and musical stream segregation.

Ontology-Based Speech and Musical Stream Segregation

Sound segregation processing modules

Speech and musical stream segregation systems are integrated on the basis of sound ontology. As mentioned before, two systems, Bi-HBSS and OPTIMA (Kashino *et al.* 1995) are decomposed into primitive processing modules as follows:

- Processing modules for speech stream segregation:
  1. voice segregation,
  2. unvoiced consonants (residual signal after all harmonic sounds are subtracted from input signal) extraction,
  3. sequential grouping of voice, and
  4. unvoiced consonant interpolation.
- Processing modules for musical stream segregation:
  1. note extraction,
  2. identification of sound sources (instruments),
  3. rhythm extraction,
  4. code extraction, and
  5. knowledge sources that store musical information statistics.

The relationship between processing modules are shown in Fig. 3. Some are common, while others are specific to speech or musical stream segregation.

Three interfacing modules are designed to integrate the modules; discriminator of voice and musical notes, primitive fundamental frequency tracer, and mediator of single tone tracers.

Discriminator of voice and musical note

Many signal processing algorithms have been presented to distinguish sound classes, such as discriminant analysis and the subspace method (Kashino *et al.* 1995). For simplicity, we adopt a heuristics on the fundamental frequency pattern of a harmonic stream. A Harmonic sound is recognized as a musical note if the standard deviation of the fundamental frequency, denoted σ_{f_0}, for n millisecond from the beginning of the stream satisfies the following inequality:

$$\sigma_{f_0}/\bar{f}_0 < c,$$

where n and c are constants, and $\bar{f}_0$ is the average fundamental frequency. Otherwise, the sound is treated as a part of speech.

Primitive fundamental frequency tracer

A primitive fundamental frequency tracer is redesigned, although such a tracer is implicitly embedded in Bi-HBSS or Optima. The primitive fundamental frequency tracer extracts a harmonic structure at each time frame as follows:

1. the fundamental frequency at the next time frame is predicted by linearly extending its transition pattern, and
2. a harmonic structure whose fundamental frequency is in the region neighboring the predicted one in the next frame is tracked as the fundamental frequency.

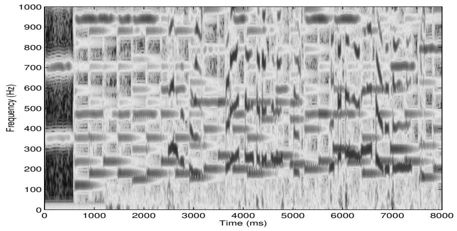
Since musical notes have a specific fundamental frequency pattern, that is, musical scale, it can be used as a constraint on musical stream. If a sound being traced is identified as a musical stream, the primitive fundamental frequency tracer is replaced by musical fundamental frequency tracer. First, fundamental frequency is predicted by calculating its average fundamental frequency, and the search region for the fundamental frequency is restricted to a narrower region than the default. As a result of using stricter constraints, more precise and less ambiguous fundamental frequencies of musical notes can be extracted.

Mediator of single sound tracers

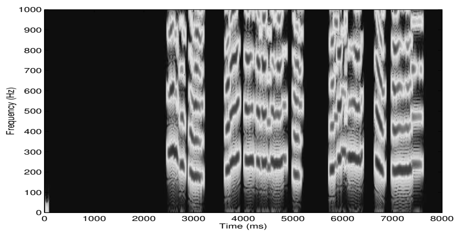
Two types of interaction modules between single sound tracers are designed in order to assign individual harmonic components to a musical or speech stream segregation system: one module in the relation class "[harmonic structure, harmonic structure]", and another in the relation class "[musical note, musical note]".

The interaction module in the relation class "[harmonic structure, harmonic structure]" defines default interaction, because "[harmonic structure, harmonic structure]" is the parent relation class of "[musical note, musical note]" and so on. This interaction module decompses overlapping frequency components into streams in the same way as Bi-HBSS (Nakatani, Okuno, and Kawabata 1995b).

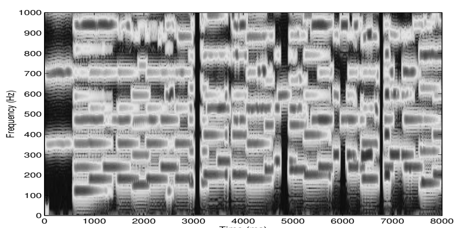
The other module in the relation class "[musical note, musical note]"leaves overlapping frequency components shared between sinle sound tracers, because such decomposing is quite difficult (Kashino *et al.* 1995).



(a) Input mixture of music (Auld Lang Syne) and narration (female voice)



(b) Segregated harmonic stream corresponding to narration (female voice)



(c) Segregated harmonic stream corresponding to music (Auld Lang Syne)

Figure 4: Spectrograms of input mixture and segregated harmonic streams.

The other relation classes such as "[speech, speech]" are not explicitly specified, but inherit the interaction modules of the default relation class "[harmonic structure, harmonic structure]".

Results of speech and musical segregation

Evaluation is performed with a mixture of the musical piece "Auld Lang Syne" played on flute and piano (sound synthesized by a sampler) and a narration by a woman uttering a Japanese sentence. The spectrograms of the input and extracted two harmonic streams are shown in Fig. 4. Although the prototype system takes monaural sounds as input instead of binaural sounds, its performance of segregation is better in terms of pitch errors and spectral distortion than that of Bi-HBSS. Of course, when fundamental frequencies of voice and music cross each other, two sounds are not segregated well. This problem is unavoidable if harmonics is the only clue for segregation. As mentioned before, Bi-HBSS overcomes the problem by using directional information.

Some additional modules such as an unvoiced consonant interpolation module and rhythm and chord extraction modules are under development to improve the performance of segregation of all parts of voice and music.

Related work

Nawab *et al.* proposed "unified terminology" as universal representation of speech and sporadic environmental sounds, and developed a spoken digit recognition system using the IPUS architecture, a variant of blackboard architecture (Nawab *et al* 1995). The idea of combining processing modules based on unified terminology is quite similar to our ontology-based sound stream segregation. Since a module is implemented as a separate knowledge source, the processing is performed in batch and incremental processing is difficult. We think that HEARSAY-II like usage of blackboard architecture which each knowledge source has a limited capability would require a sound ontology

Minsky suggested a *musical CYC project*, a musical common sense database, to promote researches on understanding music (Minsky and Laske 1992). However, a collection of various musical common sense databases may be more easily to construt than a monolith huge database, and we expect that a music ontology will play an important role in combining musical common sense databases.

Conclusion

Sound ontology is presented as a new framework for integrating existing sound stream segregation systems, interfacing sound stream segregation systems with applications, and integration of top-down and bottom-up processings. That is, sound ontology specifies a common representation of sounds and a common specification of sound processing to combine individual sound stream segregation systems. We believe that sound ontology is a key to an expansible CASA system because it provides a systematic and comprehensive principle of integrating segregation technologies.

Future work includes design of more universal sound ontology, full-scale implementation of speech and musical segregation systems, and attacking of the CASA challenge. Last but not least, controling bottom-up processing with top-down processing along with a sound ontology is an important and exciting future work.

Acknowledgments

We thank Drs. Kunio Kashino, Takeshi Kawabata, Hiroshi Murase, and Ken'ichiro Ishii of NTT Basic Research Labs, Dr. Masataka Goto of Electro Technology Lab, and Dr. Hiroaki Kitano of Sony CSL for their valuable discussions.

References

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

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

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

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

Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H. 1995. Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism, In *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, vol.1:173–178, IJCAI.

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

Lesser, V., Nawab, S.H., Gallastegi, I., and Klassner, F. 1993. IPUS: An Architecture for Integrated Signal Processing and Signal Interpretation in Complex Environments. In *Proceedings of Eleventh National Conference on Artificial Intelligence (AAAI-93)*, 249–255, AAAI.

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

Minsky, M., and Laske, O. 1992. Forward: Conversation with Marvin Minsky, In Balaban, M., Ebcioğlu, K., and Laske, O. eds. *Understanding Music with AI: Perspectives on Music Cognition*, ix–xxx, AAAI Press/MIT Press.

Nakatani, T., Okuno, H.G., and Kawabata, T. 1994a. Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agent System. In *Proceedings of 12th National Conference on Artificial Intelligence (AAAI-94)*, 100–107, AAAI.

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

Nawab, S.H., Espy-Wilson, C.Y., Mani, R., and Bitar, N.N. 1995. Knowledge-Based analysis of speech mixed with sporadic environmental sounds. In Rosenthal and Okuno, eds. *Working Notes of IJCAI-95 Workshop on Computational Auditory Scene Analysis*, 76–83.

Okuno, H.G., Nakatani, T., and Kawabata, T. 1996. Interfacing Sound Stream Segregation to Speech Recognition Systems — Preliminary Results of Listening to Several Things at the Same Time. In *Proceedings of 13th National Conference on Artificial Intelligence (AAAI-96)*, 1082–1089.

Okuno, H.G., Nakatani, T., and Kawabata, T. 1997. Understanding Three Simultaneous Speakers. In *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, Vol.1:30–35.

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

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

Young, S., Jansen, J., Odell, J., Ollanson, D., and Woodland, P. 1996. *the HTK Book for HTK V2.0*. Entropic Combridge Reseach Lab. Inc.