Organization of Hierarchical Perceptual Sounds :

Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism

Kunio Kashino*, **Kazuhiro Nakadai**, **Tomoyoshi Kinoshita** and **Hidehiko Tanaka**

H.Tanaka Lab. Bldg#13, Department of Electrical Engineering,
Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-Ku, Tokyo 113 Japan.
kashino@MTL.t.u-tokyo.ac.jp

Abstract

We propose a process model for hierarchical perceptual sound organization, which recognizes *perceptual sounds* included in incoming sound signals. We consider perceptual sound organization as a scene analysis problem in the auditory domain. Our model consists of multiple processing modules and a hypothesis network for quantitative integration of multiple sources of information. When input information for each processing module is available, the module rises to process it and asynchronously writes output information to the hypothesis network. On the hypothesis network, individual information is integrated and an optimal internal model of perceptual sounds is automatically constructed. Based on the model, a *music scene analysis* system has been developed for acoustic signals of ensemble music, which recognizes rhythm, chords, and source-separated musical notes. Experimental results show that our method has permitted autonomous, stable and effective information integration to construct the internal model of hierarchical perceptual sounds.

1 Introduction

Over the past years, a number of approaches have been taken on machine vision: both theoretical and experimental efforts on feature extraction, shape restoration, stereo vision, knowledge-based vision and other techniques have been accumulated. On the other hand, research on machine audition, or computer systems to understand acoustic information, has been so far focused mainly on spoken language understanding. However, one of the requirements to an intelligent system is to possess the ability of recognition of various events in a given environment. Specifically, understanding not only visual information or speech but also various acoustic information would play an essential role for an intelligent system which works in the real world.

On recognition or understanding of non-speech acoustic signals, several pioneering works can been found in the literature. For example, environmental sound recognition systems and auditory stream segregation systems have been developed [Oppenheim and Nawab, 1992; Lesser *et al.*, 1993; Nakatani *et al.*, 1994], as well as music transcription systems and music sound source separation systems [Roads, 1985; Mellinger, 1991; Kashino and Tanaka, 1993; Brown and Cooke, 1994]. Here we consider two aspects: flexibility of processing and hierarchy of perceptual sounds.

First, we note that the flexibility of existing systems has been rather limited when compared with human auditory abilities. For example, automatic music transcription systems which can deal with given ensemble music played by multiple music instruments have not yet realized, although several studies have been conducted [Mont-Reynaud, 1985; Chafe *et al.*, 1985].

Regarding flexibility of auditory functions in humans, recent progress in physiological and psychological acoustics has offered significant information. Especially, the property of information integration in the human auditory system has been highlighted, as demonstrated in the "auditory restoration" phenomena [Handel, 1989]. To achieve flexibility, machine audition systems must have this property, since sound source separation, a sub problem of sound understanding, is an inverse problem in general formalization and cannot be properly solved without such information as memories of sound or models of the external world, as well as given sensory data.

Using the blackboard architecture, information integration for sound understanding has already been realized [Oppenheim and Nawab, 1992; Lesser *et al.*, 1993; Cooke *et al.*, 1993]. However, it is still necessary to consider a quantitative and theoretical background in information integration.

Second, we should consider the basic problem of sound understanding, "what is a single sound", noting the distinction between a *perceptual sound* and a physical sound. A perceptual sound in our terminology is a cluster of acoustic energy which humans hear as one sound, while a physical sound means an actual vibration of media. For example, when one listens to ensemble music of several instruments through one loudspeaker, there is a single physical sound source while we hear multiple perceptual sounds. As discussed in the following sections, an essential property of perceptual sound is its hierarchical structure.

*Currently at NTT Basic Research Laboratories.

With these points as background, we provide a novel process model of hierarchical perceptual sound organization with a quantitative information integration mechanism. Our model is based on probability theory and characterized by its autonomous behavior and theoretically proved stability.

2 Problem Description

2.1 Perceptual Sound Organization

An essential problem of perceptual sound organization is a clustering of acoustic energy to create such clusters that humans hear as one sound entity. Here it is important to note that humans recognize various sounds in a hierarchical structure in order to properly grasp and understand the external world. That is, a perceptual sound is structured in both spatial and temporal hierarchy. For example, when one waits for a person to meet standing in a busy street, the waiting person sometimes hears a whole traffic noise as one entity, while sometimes hears a noise of one specific car as one entity. If he or she directs attention to the specific car's sound, an engine noise of the car or a frictional sound from the road surface and the tires of the car can be heard separately as one entity.

Figure 1 shows an example of snapshot of perceptual sounds for music. Note that there is not only spatial structure as shown in this figure but also temporal clusters of perceptual sounds, typically melodies or chord progression, though the temporal structure of perceptual sounds has not been depicted in Figure 1 for simplicity of the figure.

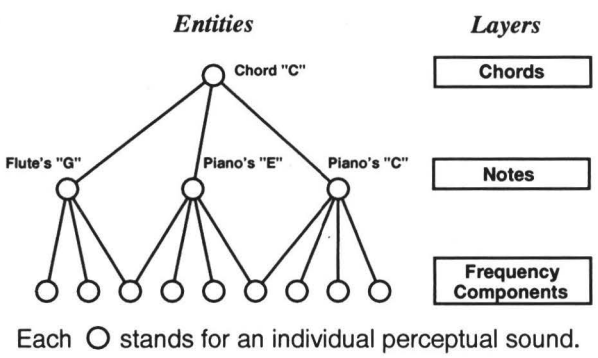


Figure 1: **An example of snapshot of perceptual sounds**

The problem of perceptual sound organization can be decomposed into the following sub problems:

1. Extraction of frequency components with an acoustic energy representation.
2. Clustering of frequency components into perceptual sounds.
3. Recognition of relations between the clustered perceptual sounds and building a hierarchical and symbolic representation of acoustic entities.

Note that we consider the problem as extraction of *symbolic* representation from flat energy data, while most approaches toward "auditory scene analysis" have so far considered their problem as restoration of target sound signals[Nakatani *et al.*, 1994; Brown and Cooke, 1992]. In the computer vision field, the scene analysis problem has been considered as extration of symbolic representation from bitmap images and clearly distinguished from the image restoration problem which addresses recovery of target images from noise or intrusions.

2.2 Music Scene Analysis

Here we have chosen music as an example of applicable domain of perceptual sound organization. We use the term music scene analysis in the sense of perceptual sound organization in music. Specifically, music scene analysis refers to recognition of frequency components, notes, chords and rhythm of performed music.

In the following sections, we first introduce general configuration of the music scene analysis system. We then focus our discussion on hierarchical integration of multiple sources of information, which is an essential problem in perceptual sound organization. Then behavior of the system and results of the performance evaluation are provided, followed by discussions and conclusions.

3 System Description

Figure 2 illustrates our process model OPTIMA (Organized Processing toward Intelligent Music Scene Analysis). Input of the model is assumed to be monaural music signals. The model creates hypotheses of frequency components, musical notes, chords, and rhythm. As a consequence of probability propagation of hypotheses, the optimal (here we use the term "optimal" in the sense of "maximum likelihood") set of hypotheses is obtained and outputted as a score-like display, MIDI (Musical Instrument Digital Interface) data, or re-synthesized source-separated sound signals.

OPTIMA consists of three blocks: (A) preprocessing block, (B) main processing block, and (C) knowledge sources. In the preprocessing block, first the frequency analysis is performed and a sound spectrogram is obtained. An example of sound spectrograms is shown in Figure 3.

With this acoustic energy representation, frequency components are extracted. This process corresponds to the first sub problem discussed in the previous section. In the case of complicated spectrum patterns, it is difficult to recognize onset time and offset time solely by bottom-up information. Thus the system creates several terminal point candidates for each extracted component, which are displayed in Figure 4 as white circles.

With Rosenthal's rhythm recognition method [Rosenthal, 1992] and Desain's quantization method [Desain and Honing, 1989], rhythm information is extracted for precise extraction of frequency components and recognition of onset/offset time. Based on the integration of beat probabilities and termination probabilities of terminal point candidates, the candidates were fixed their status: continuous or terminated, and consequently *processing scopes* are formed. Here a processing scope is a group of frequency components whose onset times are

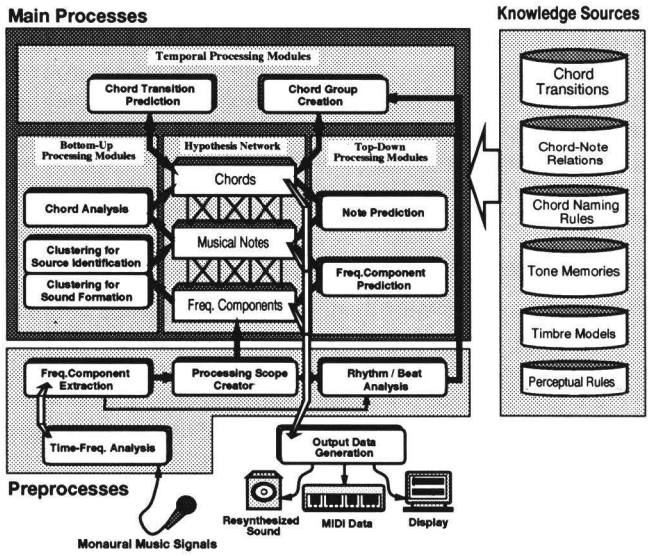
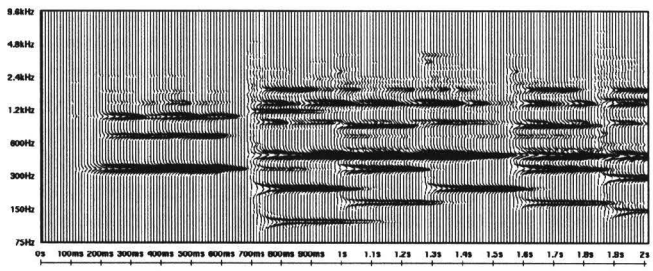


Figure 2: **Configuration of the process model**



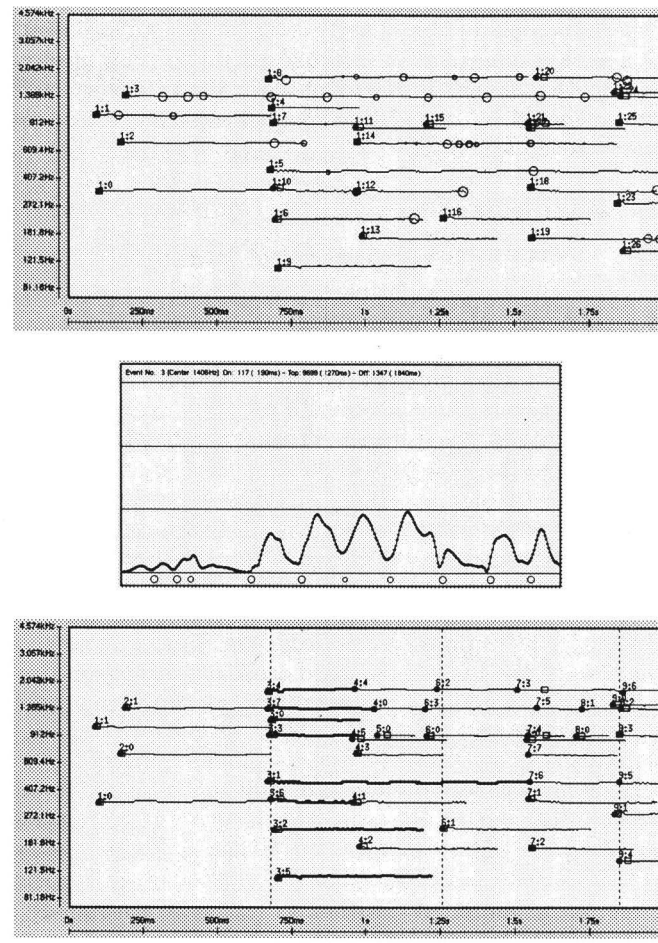
Ordinate: frequency, abscissa: time.
(Source: the beginning of a two part chamber ensemble "Auld Lang Syne", performed by a piano and a flute)

Figure 3: **An example of spectrograms**

close. The processing scope is utilized as a basic time clock for succeeding main processes of OPTIMA, as discussed later. Examples of the formed processing scopes are shown in Figure 4 (Bottom panel).

When each processing scope is created in the preprocessing block, it is passed to the main processing block, as shown in Figure 2. The main block has a hypothesis network with three layers corresponding to levels of abstraction: (1) frequency components, (2) musical notes and (3) chords. Each layer encodes multiple hypotheses. That is, OPTIMA holds an internal model of the external acoustic entities as a probability distribution in the hierarchical hypothesis space.

Multiple processing modules are arranged around the hypothesis network. The modules are categorized into three blocks: (a) bottom-up processing modules to transfer information from a lower level to a higher level, (b) top-down processing modules to transfer information from a higher level to a lower level, and (c) temporal processing modules to transfer information along the time axis. The processing modules consult knowledge sources



Top : Extracted frequency components (displayed as lines) with terminal point candidates (white circles). Radius of each circle corresponds to the estimated probability of termination. Ordinate: frequency, abscissa: time.

Middle : Terminal point candidates for the component "1:3" in the top panel with time-power plane display, showing the difficulty of finding where a component terminates or starts only by bottom-up information. Ordinate: power, abscissa: time.

Bottom : Processing scopes with the label "Scope-Id:Component-Id", formed with rhythm information. Vertical dotted lines show rhythm information extracted by the system. As an example, Scope No.3 is highlighted. Ordinate: frequency, abscissa: time.

(Source: the beginning of a two part chamber ensemble "Auld Lang Syne", performed by a piano and a flute)

Figure 4: **Examples of frequency components and processing scopes**

if necessary. The following sections discuss the information integration at the hypothesis network and behavior of each processing module.

4 Information Integration by the Hypothesis Network

For information integration in the hypothesis network, we require a method to propagate impacts of new information through the network. We employ Pearl's Bayesian network method [Pearl, 1986], which can fuse and propagate new information represented by probabilities through the network using two separate links (λ-link and π-link) if the network is a singly connected (*e.g.* tree-structured) graph.

Figure 5 shows our application of the hypothesis network. As shown in the previous section, the network has three layers: (1) C(Component)-level, (2) N(Note)-level, and (3) S(Chord)-level. The link between the C-level node and the N-level node is the S(Single)-Link, which corresponds to one processing scope. The link between the S-level and the N-level becomes the M(Multiple)-Link, as a consequence of temporal integration: multiple notes along time axis may form a single chord. The S-level nodes are connected along time by the T(Temporal)-Link, which encodes chord progression.

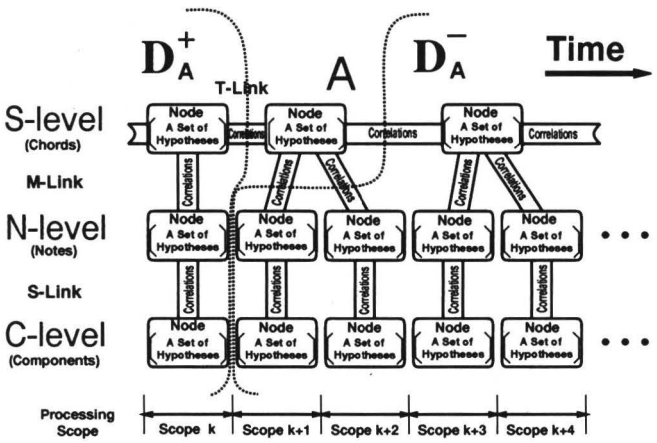


Figure 5: **Topology of the hypothesis network**

To discuss information integration scheme, assume we wish to find the belief (BEL) induced on the Node A in Figure 5, for example. Letting D_A^- stand for the data contained in the tree rooted at A and D_A^+ for the data contained in the rest of the network, we have

$$BEL(A) = P(A|D_A^+, D_A^-) \tag{1}$$

where A is a probability vector: $A = (a_1, a_2, \cdots, a_M)$. Using Bayes' theorem and assuming independence of hypotheses

$$P(D_A^+, D_A^-|a_j) = P(D_A^+|a_j)\ P(D_A^-|a_j), \tag{2}$$

we have

$$P(A|D_A^+, D_A^-) = \alpha P(D_A^-|A)\ P(A|D_A^+), \tag{3}$$

where α is a normalization constant.

Substituting as $\lambda(A) = P(D_A^-|A)$ and $\pi(A) = P(A|D_A^+)$, Equation (3) can be written as

$$BEL(A) = \alpha\ \lambda(A)\ \pi(A). \tag{4}$$

Given conditional probabilities P(Child|Parent) between any two adjacent nodes, $\lambda(A)$ can be derived from λ(Children of A) and $\pi(A)$ from π(Parent of A) [Pearl, 1986]. This derivation is considered as propagation of diagnostic (λ) or causal (π) support to A.

A minimum set of processing modules required in each node of the network is shown in Figure 6. B-Holder holds the belief (BEL) and passes new information as λ and π messages to the adjacent B-Holders. In our OPTIMA model, B-Holders are embedded in the hypothesis network and not explicitly drawn in Figure 2. H-Creator creates the hypotheses with initial probabilities. H-Correlator is for evaluating conditional probabilities $P(\text{Node}_1|\text{Node}_2)$, where Node_2 is a parent of Node_1, which are required in the information propagation process.

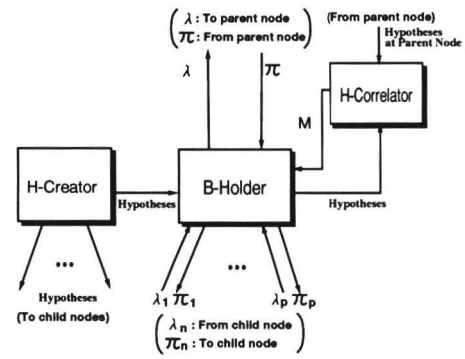


Figure 6: **Processing modules for each node of hypothesis network**

Note that the local computations required by the updating scheme are efficient: the order of computational requirement is (1) linear to the number of nodes and (2) square to the number of hypotheses in each node. In addition, not only instabilities or indefinite relaxations have been avoided by two-parameter system (π and λ), but also the order of provision of information does not affect the status of the network (probability values) after the propagation process. These properties of the hypothesis network support integration of multiple sources of information derived from autonomous processing modules. The following section shows how the processing modules work to create instances of the hypothesis network.

5 System Behavior

Based on the OPTIMA process model, a music scene analysis system has been implemented. The total amount of codes is approximately 60,000 lines (1.6 MByte) in C, except for the graphical user interface codes. Each processing module communicates with other modules through the TCP/IP socket interface, which enables us to install any modules in remote computers.

In our implementation, the frequency analysis module and the frequency component prediction module have been installed on a parallel computer (Fujitsu AP1000) to achieve high processing speed, while the other part of the system was developed on workstations. This section discusses configuration of knowledge sources and the behavior of processing modules in the main processing block in Figure 2.

5.1 Knowledge sources

Six types of knowledge sources are utilized in OPTIMA. The **chord transition dictionary** holds statistical information of chord progression, under the N-gram assumption (typically we use N=3); that is, we currently assume that the length of Markov chain of chords is three, for simplicity. Since each S-level node has N-gram hypotheses, one can note that the independence condition stated by Equation (2) is satisfied even in S-level nodes. We have constructed this dictionary based on statistical analysis of 206 traditional songs (all western tonal music), which are popular in Japan and other countries.

In the **chord-note relation** database, probabilities of notes which can be played under a given chord are stored. This information is also obtained by statistical analysis of the 2365 chords. A part of the stored data is shown in Table 1.

The **chord naming rules**, based on a music theory, are used to recognize chord when hypotheses of played notes are given.

Table 1: **Examples of the chord-note relation knowledge**

Conditional probabilities P(Note|Chord) obtained by statistical analysis of printed music.

Note	Given chord (3 examples)		
	A	A7	Am
A	.983 ± .023	.852 ± .074	1.00 ± .000
A#	.000 ± .000	.023 ± .031	.030 ± .058
B	.150 ± .064	.364 ± .101	.091 ± .098
C	.008 ± .016	.023 ± .031	.848 ± .122
C#	.850 ± .064	.818 ± .081	.000 ± .000
D	.025 ± .028	.057 ± .048	.182 ± .132
D#	.067 ± .045	.023 ± .031	.030 ± .058
E	.842 ± .065	.545 ± .104	.909 ± .098
F	.017 ± .023	.023 ± .031	.000 ± .000
F#	.125 ± .059	.148 ± .074	.000 ± .000
G	.025 ± .028	.773 ± .088	.121 ± .111
G#	.075 ± .047	.045 ± .044	.000 ± .000

± : 95 % confidence interval

The **tone memory** is a repository of frequency components data of a single note played by various musical instruments. Currently it maintains notes played by five instruments (clarinet, flute, piano, trumpet, and violin) at different expressions (forte, medium, piano), frequency range, and durations. We recorded those sound samples at a professional music studio.

The **timbre models** are formed in the feature space of the timbre. We first selected 43 parameters for musical timbre, such as onset gradient of the frequency components and deviations of frequency modulations, and then reduced the number of parameters to eleven by the principal component analysis. This eleven-dimension feature space, where at least timbres of above mentioned five instruments are completely separated with each other, is used as a timbre model information.

Finally, the **perceptual rules** describes the human auditory characteristics of sound separation[Bregman, 1990]. Currently, the harmonicity rules and the onset timing rules are employed[Kashino and Tanaka, 1993].

5.2 Bottom-up processing modules

There are two bottom-up processing modules in OPTIMA: **NHC** (Note Hypothesis Creator) and **CHC** (Chord Hypothesis Creator). **NHC** is a H-Creator for the note layer, and performs the clustering for sound formation and the clustering for source identification to create note hypotheses. It uses the perceptual rules for the clustering for sound formation, and the timbre models for discrimination analysis of timbres to identify the sound source of each note. **CHC** is a H-Creator for the chord layer, which creates chord hypotheses when note hypotheses are given. It refers to chord naming rules in the knowledge sources.

5.3 Top-down processing modules

FCP (Frequency Component Predictor) and **NP** (Note Predictor) are the top-down processing modules. **FCP** is a H-Correlator between the note layer and the frequency component layer, and evaluates conditional probabilities between hypotheses of the two layers, consulting tone memories. **NP** is a H-Correlator between the chord layer and the note layer, to provide a matrix of conditional probabilities between those two layers. **NP** uses the stored knowledge of chord-note relations.

5.4 Temporal processing modules

There are also temporal processing modules: **CTP** (Chord Transition Predictor) and **CGC** (Chord Group Creator). **CTP** is a H-Correlator between the two adjacent chord layers, which estimates the transition probability of two N-grams (not the transition probability of two chords), using the chord transition knowledge source. **CGC** decides the M-Link between the chord layers and the note layers. In each processing scope, **CGC** receives chord hypotheses and note hypotheses. Based on rhythm information extracted in the preprocessing stage, it tries to find how many successive scopes correspond to one node in the chord layer, to create M-Link instances. Thus the M-Link structure is formed dynamically as the processing progresses.

6 Evaluation

We have performed a series of evaluation tests on the system: frequency component level tests, note level tests,

chord level tests, and tests using sample song performances. In this section, a part of the results will be presented.

6.1 Note Level Benchmark Tests

An example of the experimental results for the N-level evaluation is displayed in Figure 7, which shows the effect of information integration to the note recognition rates. In Figure 7, tests have been performed in two ways: perceptual sound organization (1) without any information integration and (2) with information integration at the N-level. In the former case, the best note hypothesis produced by the bottom-up processing (**NHC**) is just viewed as the answer on the system, while in the latter case the tone memory information given by **FCP** is integrated. In both cases, we used two kinds of random note patterns: a two simultaneous note pattern and a three simultaneous note pattern. Both patterns were composed by a computer and performed by a MIDI sampler using digitized acoustic signals (16bit, 44.1kHz) of natural musical instruments (clarinet, flute, piano, trumpet, and violin). The recognition rate was defined as

$$R = 100 \cdot \left(\frac{right - wrong}{total} \cdot \frac{1}{2} + \frac{1}{2} \right) \quad [\%], \qquad (5)$$

where *right* is the number of correctly identified and correctly source-separated notes, *wrong* is the number of spuriously recognized (surplus) notes and incorrectly identified notes, and *total* is the number of notes in the input. Since it is sometimes difficult to distinguish surplus notes from incorrectly identified notes, both are included together in *wrong*. Scale factor 1/2 is for normalizing R: when the number of output notes is the same as the number of input notes, R becomes 0 [%] if all the notes are incorrectly identified and 100 [%] if all the notes are correctly identified by this normalization. The results in Figure 7 indicate that integration of tone memory information has significantly improved the note recognition rates of the system.

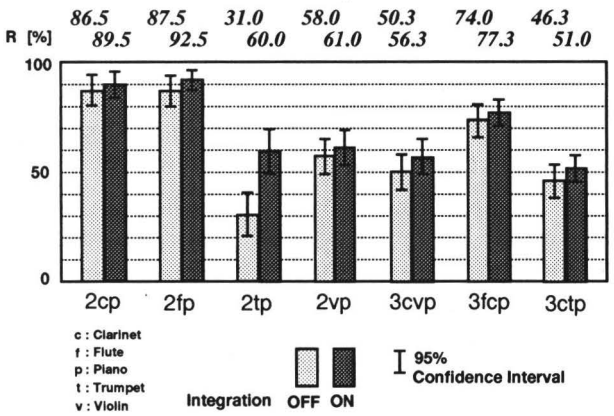


Figure 7: **Results of benchmark tests for note recognition**

6.2 Chord Level Benchmark Tests

Another example of the experimental results shows the efficacy of S-level information integration for the chord recognition rates (Figure 8). In this test, we chose a sample song with chord transition of 18 chords. Based on this chord transition pattern, test note groups were composed. To these 18 test note groups, noise (random addition or removal of the note) was added in four ways: (Exp.1) one noise note in one chord among 18 chords, (Exp.2) two noise notes in one chord among 18 chords, (Exp.3) one noise note in each of 18 chords, (Exp.4) two noise notes in each of 18 chords. Figure 8 displays significant improvement of chord recognition rates by our information integration scheme.

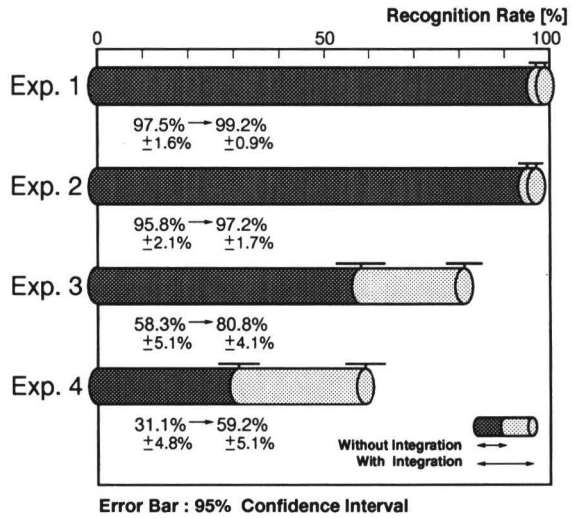


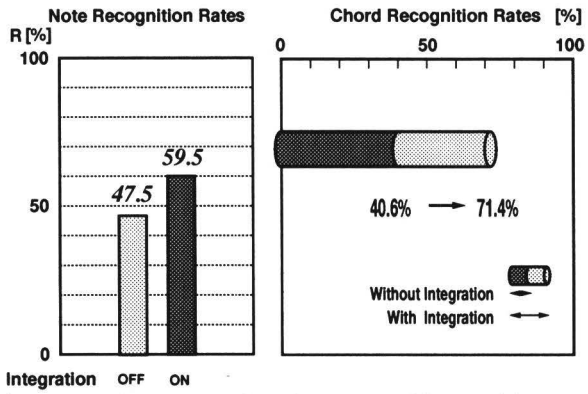
Figure 8: **Results of benchmark tests for chord recognition**

6.3 Evaluation Using a Sample Music

In addition to the benchmark tests by artificial test data, we have evaluated the system using music sound signals. Figure 9 shows the note and chord recognition rates for a sample song: a three part chamber ensemble of "Auld Lang Syne" performed by a sampler using acoustic signals of a flute, clarinet and piano. Figure 9 clearly shows that information integration is effective not only in a test data but also in a music performance.

7 Related Work

Based on the physiological and psychological findings such as the ones Bregman has summarized [Bregman, 1990], Brown and Cooke developed a computational auditory scene analysis system [Brown and Cooke, 1992]. However, it was basically a bottom-up based system, and effective integration of information was not considered. From a viewpoint of information integration, Lesser *et al.* proposed IPUS, an acoustic signal understanding system based on the blackboard architecture[Lesser *et al.*, 1993], and recently Cooke *et al.* have also considered a blackboard-based auditory scene analysis system[Cooke *et al.*, 1993]. The blackboard architecture used in those systems requires global control knowledge and tends to



(Source: three part chamber ensemble "Auld Lang Syne", performed by a sampler using a flute, a clarinet and a piano sound signal)

Figure 9: **Note and chord recognition rates for a sample music**

result in a system with complex control rules. By contrast, our model only needs the local computations and consequently supports a simple control strategy with theoretically proved stability. Recently Nakatani *et al.* reported their studies based on a multi-agent scheme [Nakatani *et al.*, 1994]. Our model can be viewed as a quantitative version of a multi-agent approach which uses probability theory.

8 Conclusion

We have proposed a method of hierarchical organization of perceptual sound, and described a configuration and behavior of the process model. Based on the model, a music scene analysis system has been developed. Specifically, our employment of a hypothesis network has permitted autonomous, stable and efficient integration of multiple sources of information.

The experimental results show that the integration of chord information and tone memory information significantly improves the recognition accuracy for perceptual sounds, in comparison with a conventional bottom-up based processing. Here we have focused on the mechanism of information integration and left out detailed discussions on optimality of the output of each processing module. We are planning to clarify theoretical limits of the accuracy of each processing module, and to conduct further experiments to evaluate systematically the advantages and disadvantages of information integration mechanism of the proposed model.

References

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

[Brown and Cooke, 1992] Brown G. J. and Cooke M. A Computational Model of Auditory Scene Analysis. In *Proceedings of International Conference on Spoken Language Processing*, pages 523–526, 1992.

[Brown and Cooke, 1994] Brown G. J. and Cooke M. Perceptual Grouping of Musical Sounds: A Computational Model. *Journal of New Music Research*, 23(1):107–132, 1994.

[Chafe *et al.*, 1985] Chafe C., Kashima J., Mont-Reynaud B., and Smith J. Techniques for Note Identification in Polyphonic Music. In *Proceedings of the 1985 International Computer Music Conference*, pages 399–405, 1985.

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

[Desain and Honing, 1989] Desain P. and Honing H. Quantization of Musical Time: A Connectionist Approach. *Computer Music Journal*, 13(3):56–66, 1989.

[Handel, 1989] Handel S. *Listening.* MIT Press, 1989.

[Kashino and Tanaka, 1993] Kashino K. and Tanaka H. A Sound Source Separation System with the Ability of Automatic Tone Modeling. In *Proceedings of the 1993 International Computer Music Conference*, pages 248–255, 1993.

[Lesser *et al.*, 1993] Lesser V., Nawab S. H., Gallastegi I. and Klassner F. IPUS: An Architecture for Integrated Signal Processing and Signal Interpretation in Complex Environments. In *Proceedings of the 11th National Conference on Artificial Intelligence*, pages 249–255, 1993.

[Mellinger, 1991] Mellinger D. K. *Event Formation and Separation of Musical Sound.* Ph.D. Thesis, Department of Music, Stanford University, 1991.

[Mont-Reynaud, 1985] Mont-Reynaud B. Problem-Solving Strategies in a Music Transcription System. In *Proceedings of the 1985 International Joint Conference on Artificial Intelligence*, pages 916–918, 1985.

[Nakatani *et al.*, 1994] Nakatani T., Okuno H. G., and Kawabata T. Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agent System. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 100–107, 1994.

[Oppenheim and Nawab, 1992] Oppenheim A. V. and Nawab S. H. *(eds.). Symbolic and Knowledge-Based Signal Processing.* Prentice Hall, 1992.

[Pearl, 1986] Pearl J. Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence*, 29(3):241–288, 1986.

[Roads, 1985] Roads C. Research in Music and Artificial Intelligence. *ACM Computing Surveys*, 17(2):163–190, 1985.

[Rosenthal, 1992] Rosenthal D. *Machine Rhythm: Computer Emulation of Human Rhythm Perception.* PhD. Thesis, Department of Computer Science, Massatussetts Institute of Technology, 1992.