Musical Sound Source Identification Based on Frequency Component Adaptation

Tomoyoshi Kinoshita and **Shuichi Sakai** and **Hidehiko Tanaka**
Graduate School of Information Engineering, The University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo, 113-8656, JAPAN
E-mail: {kino,sakai,tanaka}@mtl.t.u-tokyo.ac.jp
TEL: +81-3-5841-7413, FAX: +81-3-5800-6922

Abstract

In auditory scene analysis, sound source identification is an essential operation when extracting musical notes from acoustical signals composed of multiple sound sources. We have previously proposed a processing model OPTIMA for music scene analysis and implemented its experimental system. However, the system was not robust to signals with overlapped frequency components. In this paper, we present a new method that improves this problem by using overlap pattern of frequency components, and implemented as a processing module in OPTIMA. Weighted template-matching method is applied to identify sound sources repeatedly to each frequency component cluster. The weight is evaluated according to the significance of each feature of the signal. When multiple components are overlapped, our system adaptively modifies features of an input signal to a combination of overlapped components. Experimental results show that the system can identify sound sources of 66% to 75% of musical notes. It also showed about 10% improvement in accuracy, compared to the result without the proposed mechanism.

1 Introduction

As we can recognize the surrounding scene from acoustic signals received through our ears, human being and other animals are endowed with this function from birth. By contrast, it is still unable for computers or other machinery to realize their acoustic environment. To build a system that works in the real world, realization of this function is a large and essential step. In the field of auditory psychology, Bregman's work [Bregman, 1990] proposed some basic theory on this function — auditory scene analysis —. However, its realization on a computer is not examined in the work, and has attracted much interest of researchers.

Accordingly, several works have been done on auditory scene analysis by computers. Lesser *et al.* presented IPUS [Lesser *et al.*, 1995] based on blackboard architecture. In this system, agents work following a rule-based policy, thus flexibility and scalability is not achieved so much. Nakatani *et al.* employed a residue-driven architecture for sound segregation [Nakatani *et al.*, 1995]. The system has multiple agents, and two agents called tracer and eraser extract acoustic signal from one source. Ellis proposed a system based on a prediction-driven method [Ellis, 1996]. This system is sensitive to the context of the input stream. Our current method does not cope with context, but after note hypotheses are obtained, information integration mechanism works to reflect other clues including scene context.

All of these works segregate streams in acoustic signals and do not identify their sound source. As a system that identifies sound sources, Kashino *et al.*'s work, named OPTIMA [Kashino *et al.*, 1995] is the basis of the proposed system. OPTIMA has multiple processing modules that work independently. After the processes, each result is expressed as probabilistic information and integrated into the final result. In OPTIMA, sound source identification is performed in a similar way as the current system. It also uses physical features of frequency components, such as power ratio of harmonics and sharpness of onset. However, identification is done in a simple way based on principle component analysis and discriminant analysis. Even if multiple components overlap and features are affected, OPTIMA processes them as they are. Therefore, OPTIMA cannot identify sound sources with overlapped components, and this problem has waited for solutions.

Another Kashino's work [Kashino and Murase, 1997] uses wave-form template, and through template filtering and phase tracking, the template is arranged according to the input signal. Generally, signals from two individual instruments vary significantly even if both are the same kind of sound source. Kashino pointed that divergent phase between frequency components causes this phenomena, and introduced a phase tracking mechanism. However, when a template was not prepared as wave-form data but frequency components, this problem would be solved, since time-frequency analysis generally discards phase information of each component.

In this paper, we propose a novel system for sound

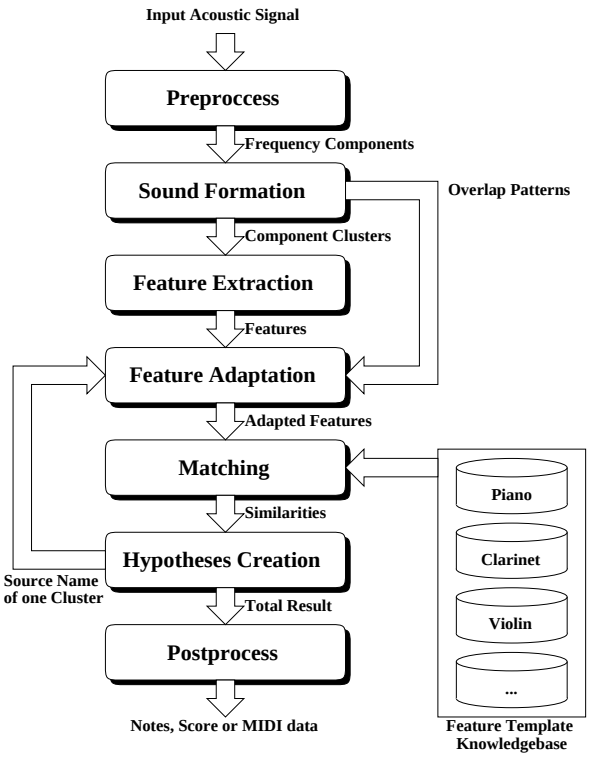


Figure 1: Configuration of proposed system.

Source	Feature #1	Feature #2	Feature #3	⋯
Piano	0.724	23.48	5.901	⋯
Piano	0.271	18.22	3.725	⋯
⋮	⋮	⋮	⋮	⋯
Clarinet	0.513	49.11	7.224	⋯
⋮	⋮	⋮	⋮	⋱

Table 1: Example of a feature template.

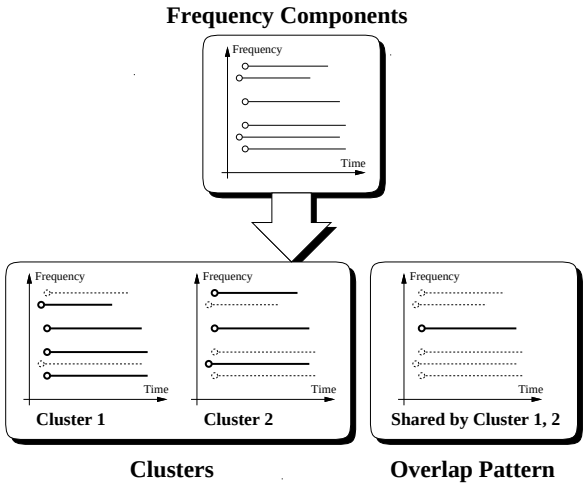


Figure 2: Clusterize into notes and obtain overlap pattern.

source identification. Our system uses an adaptive method for component's feature analysis and achieved much better identification accuracy than OPTIMA.

The system architecture is described in Section 2. And experimental results are shown in Section 3. Finally, conclusion is described in Section 4.

2 System Overview

The system has seven processing blocks and a knowledge-base of frequency components' features of sound sources. In this paper, we call this knowledge-base, a feature template. Figure 1 shows the overall diagram.

Frequency components are extracted from input acoustic signal in the **Preprocesses** stage. In the **Sound Formation** stage, components obtained in the previous stage are clusterized, where each cluster corresponds to a musical note. Overlap pattern is also extracted in this stage. The system computes features from each component in the **Feature Extraction** stage. Extracted features are adapted in the **Feature Adaptation** stage in accordance with the overlap pattern. In the **Matching** stage, the adapted features and templates in Feature Template Knowledge-base are compared and similarity is evaluated. The similarities are applied to the **Hypotheses Creation** stage, where note hypotheses are created. When all the clusters' sources are not identified, the process is restored to the Feature Adaptation stage. Finally, note data or scores are created in the **Postprocesses** stage. Details of each stage are described in the following subsections.

Feature Template has a set of records of features. Each record consists of a sound source name and a list of feature values (Table 1).

2.1 Preprocesses and Sound Formation

The system first performs time frequency analysis to obtain a sound spectrogram from the input and tone model signals. Then frequency components are extracted. In this process, we chose the IIR filter-bank method and the pinching planes method[Kashino *et al.*, 1995].

In the second stage, sound formation is performed. Through this process, frequency components are clusterized into musical notes. We used Kashino's method[Kashino *et al.*, 1995] for the task. In this method, distortion of harmony and common onset between two frequency components are extracted, and clustering is performed using the result. Additionally, we extracted patterns of frequency components' overlap. Here "overlap pattern" is defined as a set of combinations of components shared by multiple clusters (Figure 2).

- Power of each component
- Center frequency (power-weighted average of frequency)
  Here, center frequency is defined as follows:
  $$\sum pf \Big/ \sum p$$
  where p and f is the power and frequency of each frequency component in a power-spectrogram.
- Skewness and kurtosis of each component's power envelope
  Here, skewness is defined as follows:
  $$E[(p-\mu)^3]/\sigma^3$$
  and kurtosis is defined as follows:
  $$E[(p-\mu)^4]/\sigma^4 - 3$$
  where p is the power at each sample point of frequency components and μ is the mean of p.

Table 2: List of frequency component features (excerpt).

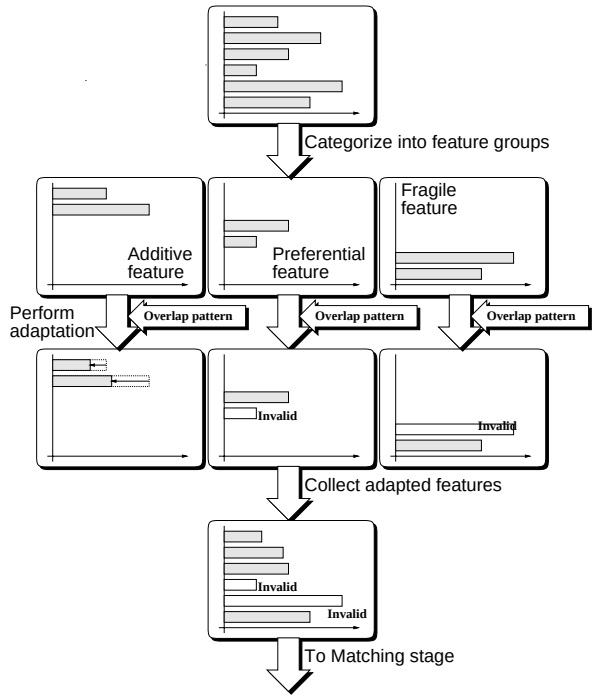


Figure 3: Adaptation mechanism.

2.2 Feature Extraction

Next, the system extracts features of frequency components. Here, feature is defined as physical parameters such as a shape of power envelope, a strength of an attack, or a power ratio of each component. Table 2 shows the list of the adopted features.

When the power of some harmonics is too small, features extracted from it is invalidated and would not be used to obtain similarity in the Matching stage.

2.3 Feature Adaptation

In real music, multiple notes are usually presented simultaneously and several frequency components are overlapped. This causes transformation of frequency components and their features. Therefore, we cannot immediately use the extracted features.

In this stage, features are arranged in accordance with the features' characteristics and the components' overlap pattern (Figure 3).

Categorize Features

We categorized features by their characteristics into three types as listed below:

1. **Additive features**
   When a frequency component overlaps with others, its feature is generally the sum of each component's feature.
   (*eg.* power of component)
2. **Preferential features**
   The feature becomes the maximum or minimum value of overlapping components.
   (*eg.* strength of attack)
3. **Fragile features**
   If multiple components overlap, the feature value is not significant any longer.
   (*eg.* skewness of component's envelope)

Adaptation

When multiple components overlap, the system recalculates each feature according to the feature's type, following the algorithms described below. In case a component is not shared by other clusters, nothing is done.

1. **Additive features**
   Adaptation is performed in the following algorithm:
   If the sound source of a cluster which shares this component is already determined
   Then
   Feature value is considered to be the sum of multiple feature values, and is subtracted by the mean of the sharing component's feature value in the template.
   Else
   The system cannot estimate the added feature value. Therefore, nothing is done.
2. **Preferential features**
   Adaptation is performed in the following algorithm:
   If the sound source of a cluster which shares this component is already determined
   Then
   If feature value is close to the sound source's feature value in the template

Then
The feature value is considered to belong to the processing note, and nothing is done.

Else
The feature value is considered not to belong to the processing note, and the component is invalidated.

Else
The system cannot decide whether the feature value belongs to the processing note or not. Therefore, nothing is done.

3. **Fragile features**
   In this case, the feature value is considered to be corrupted, and the component is invalidated.

If sound sources of some components are already determined, the recalculation is performed applying their feature templates.

2.4 Matching

Source identification is performed by weighted-matching method between sound source feature templates and adapted features.

Weight calculation

The system computes the mean and standard deviation from feature templates of each source in advance.

Then, the weight of each feature is obtained by the following equation; in this paper, we use superscript as the feature id and subscript as the sound source id:

$$w^i_{s,t} = P\left(|X| \leq \frac{|\mu^i_t - \mu^i_s|}{\sigma^i_s}\right) \tag{1}$$

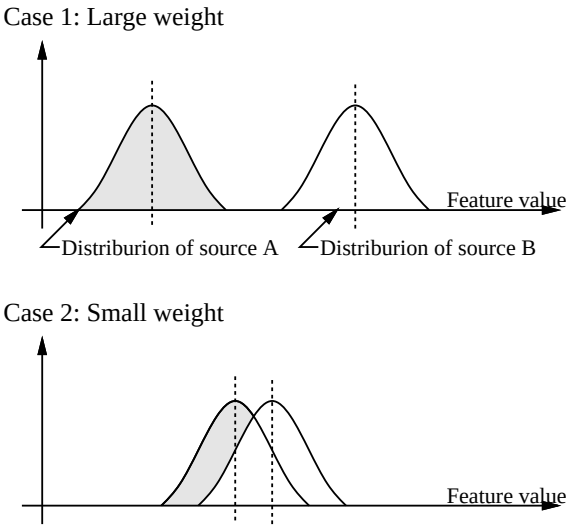
$$W^i_s = \sqrt{\frac{1}{|S|-1} \sum_{t \in S, t \neq s} {w^i_{s,t}}^2} \tag{2}$$

where S denotes the set of sound sources ($=\{Piano, Clarinet, \cdots\}$) and s denotes each sound source ($s \in S$). P is a probabiluty value in normal distribution ($P(|X| \leq z) = \int_{-z}^{z} (1/\sqrt{2}) \exp(-x^2/2) dx$).

Here, we define D^i_s as the distribution of the s-th source's i-th feature. Using the equation above, W^i_s becomes larger when D^i_s is isolated from other sound sources. For example, when D^i_s is sufficiently apart from other sources' D^i_s, the weight W^i_s becomes 1, and when all D^i_s have equivalent mean values, the weight becomes 0. Therefore, the i-th feature value with large W^i_s is assumed to be significant in the s-th source's identification (Figure 4). W^i_s is applied to similarity evaluation in the next stage.

Similarity Calculation

The system evaluates the similarity between features of the input signal and each template. The similarity represents the confidence of source identification.



Case1: If the feature value distributions of two sound sources are sufficiently apart, the feature is considered appropriate as a clue to identify sources.

Case2: When distributions are close, the feature is considered inappropriate.

Figure 4: Weighting based on the distribution of feature values.

Before the similarity evaluation, distance between the target input signal and features in the template database is calculated. The distance is obtained as follows:

$$d^i_s = P\left(|X| \geq \frac{|f^i - \mu^i_s|}{\sigma^i_s}\right) \tag{3}$$

where d^i_s, f^i, μ^i_t and σ^i_t denote the distance of the i-th feature between the input and the s-th source's template, the i-th feature value of the input signal, average and standard deviation of the i-th feature of the s-th source's template, respectively. P is a probability function as previously denoted.

Then, the similarity is calculated by the following equation:

$$S_s = \exp\left(\sum_i W^i_s \log d^i_s \Big/ \sum_i W^i_s\right) \tag{4}$$

where W^i_s and d^i_s are obtained beforehand.

In this equation, i represents each feature, but when a feature value is invalidated, the summation is skipped. The following list describes the case when a feature becomes invalid:

1. A frequency component is not extracted because its power is too small. The feature value is invalidated during the Frequency Component Extraction stage.

Source Name	Feature Name
Piano	temporal symmetry of first harmonics temporal symmetry of third harmonics skewness of first harmonics
Clarinet	kurtosis of first harmonics temporal symmetry of third harmonics power of second harmonics
Violin	attack strength of first harmonics attack strength of second harmonics kurtosis of first harmonics

Table 3: Top three large weighted features.

2. A component overlaps with another one and the feature is fragile.

2.5 Hypotheses Creation

After the matching process is performed, the system fixes the source name of the note with the lowest pitch among the clusters, and feeds it back to the Feature Adaptation stage. In the Feature Adaptation stage, adaptation is performed again. At this time, one cluster's source name is determined and its template is used in calculation.

When all clusters' sources are identified, the system creates note hypotheses. Each hypothesis has multiple notes, and each note has metrics such as onset time, duration time, pitch and source name.

2.6 Postprocesses

In OPTIMA architecture, note hypotheses obtained in the last stage and other probabilistic information are integrated afterwards[Kashino *et al.*, 1995]. Through this process, errors included in the hypotheses are expected to be corrected.

3 Evaluation

We evaluated the proposed system in two ways. First, feature's weight value is calculated and its validity is evaluated. Next, source identification accuracy for random note pattern is evaluated. In this paper, the effectiveness of the proposed method is proved by comparing results of recognition with and without feature adaptation and weight calculation.

3.1 Weight Calculation

First, Table 3 shows the experimental result of W_s^i calculation in the Matching Stage.

This result matches our intuition quite well. Piano's power has a sharp attack and a calm decline, and most of its power is distributed at the beginning of the envelope. Low value of Temporal symmetry or large skewness reflects asymmetry of Piano's power envelope. Clarinet has little power in even harmonics and each component has trapezoidal power distribution. Therefore, kurtosis and power of the second harmonics becomes lower than other sources, and temporal symmetry becomes larger.

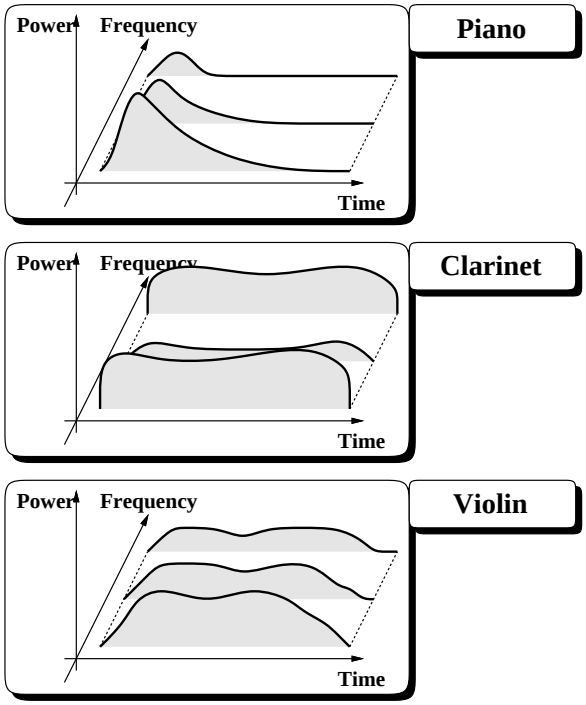


Figure 5: Typical shapes of sound source's frequency component.

Violin's onset is relatively gentle in the sources we used. Hence attack strength is a significant clue to identify whether a source is a violin or not. (Figure 5)

3.2 Experiment to Random Note Pattern

We prepared a random note pattern in advance to the evaluation. The random note pattern consists of pairs of notes. The sound source and pitch of each note are chosen at random. Random notes are categorized into three classes:

Class 1: The second harmonics of a note overlaps with the base component of another note. In this class, all components of the note overlaps with components of other notes. This is the most difficult case among the three classes.

Class 2: The second harmonics of a note and the third harmonics of another note overlap. These two notes have an interval of perfect fifth; frequencies of the notes' base components are $2:3$.

Class 3: Patterns categorized into neither Class 1 nor 2. Overlap of components rarely occurs.

Here, each class has 300 patterns, and a total of 900 note patterns were processed in the system. In this experiment, we used signal from different instruments for input and feature templates.

Figure 6 shows the result of the note creation. This is the recognition accuracy when source name misrecognition is ignored, which shows the upper-boundary of the source identification accuracy.

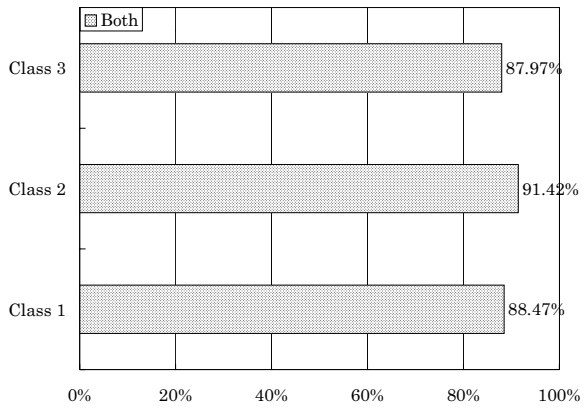


Figure 6: Result of note creation for random note patterns.

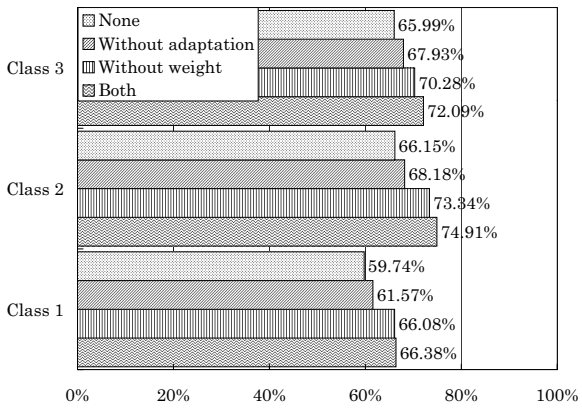


Figure 7: Result of whole process for random note patterns.

Figure 7 shows the source identification accuracy. These bars indicate the percentage of correctly-identified notes. Each bar shows the result when feature adaptation and weight calculation are performed or not, as follows:

Bar Title	Weight Calculation	Feature Adaptation
Both	Performed	Performed
Without adaptation	Performed	Not performed
Without weight	Not performed	Performed
None	Not performed	Not performed

Note that accuracy stands for the average of recall and precision.

This result includes failures in the note creation stage, and does not show independent accuracies of source identification. Figure 8 shows the result when the pitch of each note is given before source identification is performed.

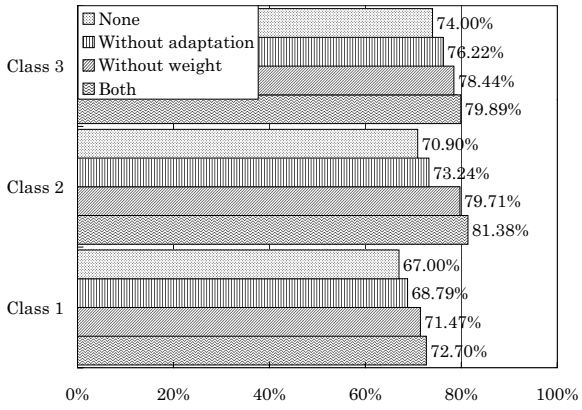


Figure 8: Result of source identification when the pitch is given beforehand.

4 Conclusions

We have presented a new method for musical sound source identification, which enables the identification for component-overlapped signals by feature adaptation.

The experimental result showed the effectiveness of this method. In the case of random note pattern, the recognition accuracy has improved from 70% to 81% in the most effective case; Class 2.

In real music, multiple notes have frequencies of integral ratio to express fine harmony, and component-overlaps like Classes 1 and 2 in this paper appear often. Therefore, the proposed method is effective also to the source identification of musical signals.

For practical use, the accuracy of the proposed system is insufficient and more improvement is waited for. In the adaptation stage, we used a simple algorithm for overlapped components, such as "Do nothing" and "Mark as invalid". We are going to examine more sophisticated ways which consider other informations like the difference between types of instruments.

We used data from different individual of the same instrument for input and feature templates of the experiment. Most of the influence is avoided through the calculation of mean of feature value, but in some cases the system failed to identify the source name because of this influence. Therefore, this problem should be solved.

Some errors appeared in the evaluation result are caused by component extraction failure. Its improvement is also our future work.

Acknowledgments

This research is supported by Grant-in-Aid for JSPS Fellows (No.09-07628) and JSPS Fellowships for Young Scientists.

We also would like to thank NTT Communication Science Laboratories for the permission to use the

NTTMSA-P1 acoustic signal database.

References

[Bregman, 1990] A. S. Bregman. Auditory scene analysis. *MIT Press*, 1990.

[Ellis, 1996] Daniel P. W. Ellis. *Prediction-driven Computational Auditory Scene Analysis*. PhD thesis, Department of Electrical Engineering and Computer Science, M.I.T., 1996.

[Kashino and Murase, 1997] Kunio Kashino and Hiroshi Murase. A music stream segregation system based on adaptive multi-agents. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, volume 2, pages 1126–1131, August 1997.

[Kashino *et al.*, 1995] Kunio Kashino, Kazuhiro Nakadai, Tomoyoshi Kinoshita, and Hidehiko Tanaka. Organization of hierarchical perceptual sounds. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Volume 1*, pages 158–164. Morgan Kaufmann Publishers, Inc., August 1995.

[Lesser *et al.*, 1995] Victor R. Lesser, S. Hamid Nawab, and Frank I. Klassner. IPUS: An architecture for the integrated processing and understanding of signals. *Artificial Intelligence*, 77:129–171, 1995.

[Nakatani *et al.*, 1995] Tomohiro Nakatani, Hiroshi G. Okuno, and Takeshi Kawabata. Residue-driven architecture for computational auditory scene analysis. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 1, pages 165–172, August 1995.