A Sound Source Separation System with the Ability of Automatic Tone Modeling

Kunio Kashino *and* Hidehiko Tanaka

Bldg.# 13, Department of Electrical Engineering,
Faculty of Engineering, The University of Tokyo,
7-3-1, Hongo, Bunkyo-Ku, Tokyo, 113, Japan.

kashino@MTL.t.u-tokyo.ac.jp tanaka@MTL.t.u-tokyo.ac.jp

Abstract

We describe the configuration, implementation and evaluation of a sound source separation system which has the ability of automatic tone modeling. Input is assumed to be a monaural audio signal containing the sounds of several kinds of musical instruments. The system separates the input by instrument and generates MIDI data with each MIDI channel assigned to one kind of instrument. The output data is also displayed on the screen in a score-like format. A prototype system has been implemented and evaluated by a benchmark test. It is shown that the performance of the proposed system is better than that of a conventional bottom-up based system.

1 Introduction

Sound source separation refers to extracting information which originates in one sound source under the condition that multiple sound sources are simultaneously present. This is a key technique to realize an automatic transcription system for ensemble music, a speech separation system, and an auditory scene analysis [Bregman, 1990a] system.

Here a physical sound source and a perceptual sound are distinguished, and the concept of perceptual sound source separation is proposed. A perceptual sound stands for the sound which humans hear as one sound, while a physical sound means an actual sound source itself. For example, when one hits a key on piano keyboard, usually a hammer strikes more than one piece of wire at once. In this case, there are multiple physical sound sources, though one usually hears the sound as one sound. Another example is to listen to an ensemble of instruments through one loudspeaker. In this case, there is a single physical sound source while we hear separate sound sources.

Over the past years, several approaches have been taken on sound source separation; most of them have aimed at separation of a physical sound. A broad classification of them would be (1) the approach using the information of sound localization [Mitchell *et al.*, 1971] [Flanagan *et al.*, 1985] [Nagata *et al.*, 1991] and (2) the one based on template matching. Both approaches are inflexible: the former is sensitive to the location of sources or microphones, and cannot be applied to a monaural source; the latter approach requires the advance registration of templates.

This paper addresses a process model of perceptual sound source separation to realize flexible processing which has compatibility with human auditory characteristics. The flexibility has two aspects: the process model is (1) sound localization-free and (2) template registration-free. That is, the input of the system we describe is assumed to be a monaural sound signal, and the system runs without advance registration of the sound models.

Sound enhancement by means of harmonic selection [Parsons, 1976] [Nehorai *et al.*, 1986] and some bottom-up approaches [Brown, 1992] in the literature can be thought as examples of perceptual sound source separation. However, treatment of the cues for perceptual sound separation was, so far, limited to qualitative one. Thus we pay attention to *quantitative* treatment of the cues.

If the complete template set of musical instruments were available, we not require a registration-free approach. However, one musical sound differs by instrument, and player, and varies significantly according to the performance conditions. Moreover, there are a number of "new" musical sounds generated electronically. Therefore we think that a template registration-free approach plays an important role in achieving flexibility of the processing.

In the following section, we introduce the perceptual sound source separation problem. After that, important modules in the proposed system will be discussed in detail, followed by the experimental results for evaluation of the system. Finally some remarks and future direction of the work will

conclude this paper.

2 Perceptual Sound Source Separation

In this section, perceptual sound source separation is formalized. Here, perceptual sound source separation is viewed from a standpoint of creating a system which simulates human perceptual system *in performance*, not in mechanisms.

Let $S(t)$ denote a sound signal, which can be represented by a set of parameters $P(t)$:

$$P(t) = \{p_1(t),\ p_2(t),\ \cdots,\ p_N(t)\}\ , \qquad (1)$$

where $p_k(t)$ are parameters. Suppose the parameter set $P_T(t)$ represents monaural sound signal $S_T(t)$, where $S_T(t)$ is a mixture of sounds which arise from M sound sources. Then sound source separation can be viewed as a problem of obtaining a parameter set $P_i(t)$ which represents the $i-th$ sound source $S_i(t)$ from $P_T(t)$.

For example, assume that waveform of a sound signal is employed as the parameter. Then sound source separation is to determine the waveform of each sound source $s_i(t)$ from the mixture waveform $y(t)$, where

$$\sum_{i=1}^{M} s_i(t) = y(t)\ . \qquad (2)$$

Mathematically, it is obvious that Equation (2) is underdetermined. Here, frequency components are used as parameters which represents a sound signal. That is, assume that $S(t)$ can be represented using a set of frequency components $F_j(t)$:

$$F(t) = \{F_1(t),\ F_2(t),\ \cdots,\ F_L(t)\}\ , \qquad (3)$$

where $F_j(t)$ can be written as

$$F_j(t) = \{p_j(t),\ f_j(t),\ \varphi_j(t)\}\ , \qquad (4)$$

where $p_j(t)$ and $f_j(t)$ stand for the power and frequency of the spectral peak; $\varphi_j(t)$ is the parameter which corresponds to bandwidth of the spectral peak. We should note that some sound sources are insufficiently represented by Equation (4), such as white noise, and are not considered here. A sound source separation can be formulated as follows:

Problem 1 : Extraction of frequency components $F_j(t)$ on a sound spectrogram derived from a composite sound signal
(Extraction of frequency components),

Problem 2 : Clustering of $F_j(t)$ into a certain number of groups according to a certain criteria
(Clustering of frequency components).

In problem 2, each cluster corresponds to an individual sound source. Frequently one extracted frequency component is actually a mixture of overlapping multiple frequency components which originate in different sources. Thus the clustering should be a duplex grouping, which refers to the clustering which allows a frequency component to be a member of multiple clusters at the same time. This corresponds to decomposition of overlapping harmonics.

Problem 2 can be devided into two sub problems. One is clustering to group frequency components which humans tend to hear as one sound. We call this the "sound formation clustering" problem, and the results of this clustering we call a "sound cluster". Another sub problem is to group sound clusters which originate in the same sound source. We call this problem "source identification clustering", and the results of this clustering "source cluster".

In terms of perceptual sound source separation, the results of these clustering should be compatible with human auditory characteristics. Therefore it is important to investigate what humans use as clues in realizing auditory separation, not only for psychological validity but also from engineering point of view. In the field of psychological acoustics, it is claimed that the cues or clues which promote segregation or fusion of frequency components include the ones listed in Table 1 [Moore *et al.*, 1986] [Moore, 1989] [McAdams, 1989] [Bregman, 1990a] [Bregman *et al.*, 1990b] [Hartmann *et al.*, 1990] [Darwin *et al.*, 1992]. Among them, we have so far implemented four cues:

1. Harmonic mistuning of frequency components,
2. Onset asynchrony of frequency components,
3. Memory of timbres, and
4. Old-plus-new heuristic.

The automatic tone modeling has been realized in terms of the old-plus-new heuristic, as discussed in Section 7.

3 General Description of the System

Figure 1 gives a block diagram of the perceptual sound source separation system described in this paper.

Input of this system is assumed to be a monaural audio signal, which is a mixture of sounds of several kinds of musical instruments. Output of this system is MIDI data which includes several MIDI channels, each of which is assigned to one kind of instrument.

Table 1: Cues for Human Auditory Separation

Simultaneous Spectral Features
- (Seg.) Harmonic Mistuning of Freq. Comp.
- (Seg.) Onset Asynchrony of Freq. Comp.
- (Seg.) Offset Asynchrony of Freq. Comp.
- (Fus.) Common FM of Freq. Comp.
- (Fus.) Common AM of Freq. Comp.
- (Seg./Fus.) Old-Plus-New Heristic

Others
- (Seg./Fus.) Localization
- (Seg./Fus.) Sequential Grouping
- (Seg./Fus.) Memory of Timbres

Note:
Seg. : Promoting Segregation
Fus. : Promoting Fusion
Freq.Comp. : Frequency Components

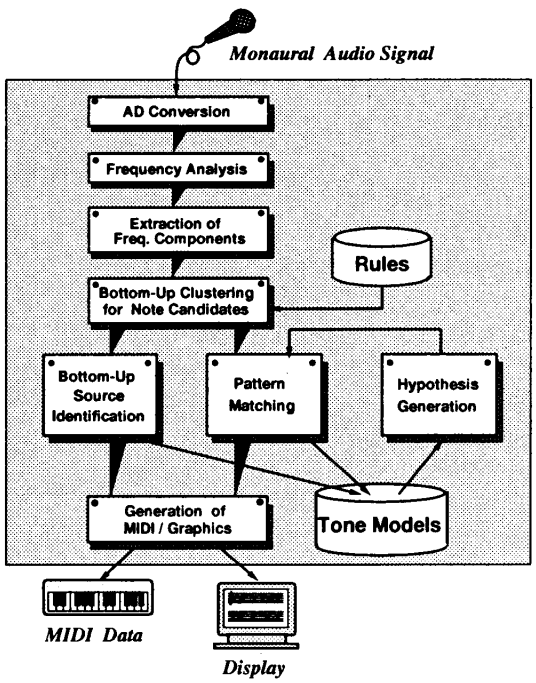

Figure 1: Configuration of the System

Graphics is also generated in a simplified score-like format for visual monitoring of the results.

The following sections describe the important modules of the system.

4 Extraction of Frequency Components

The input audio signal is first digitized at 16bit/48KHz, and then frequency analysis is performed by a bank of bandpass filters. Each filter is a 2-order IIR (Infinite Impulse Response) type, and the center frequency of each filter is set linearly in a *log* frequency scale.

The next step is the extraction of frequency components (**Problem 1**). One of the simple methods may consist of peak-picking along the frequency axis with a threshold, and peak-tracking (concatenation) along time axis. In reality, since noise or adjacent peaks will interfere, it is difficult to achieve practical accuracy by a simple threshold method. Thus we have developed the pinching plane method in peak picking and tracking.

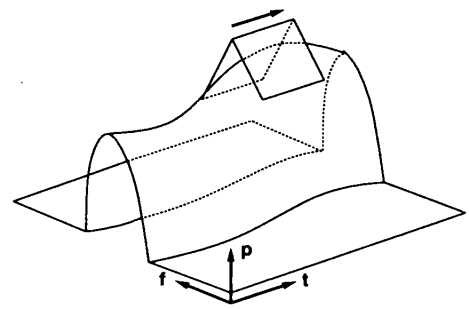

Figure 2: Extraction of Frequency Components Using Pinching Planes

As illustrated in Figure 2, this method uses two planes pinching a spectral peak in order to decide the direction of concatenation of peaks. The planes are the regression planes of the peak, calculated by a least-squares fitting.

The pinching plane z can be written as

$$z = at + bf + c, \quad a, b, c : const., \qquad (5)$$

where t and f are the time and frequency. Letting z_{ij} denote the power at (t_i, f_j) on the spectrogram, the normal vector $(a, b, -1)$ can be calculated by

the following equation:

$$\begin{pmatrix} n\sum_{i=1}^{m} t_i^2 & \sum_{i=1}^{m}\sum_{j=1}^{n} t_i f_j & n\sum_{i=1}^{m} t_i \\ \sum_{i=1}^{m}\sum_{j=1}^{n} t_i f_j & m\sum_{j=1}^{n} f_j^2 & m\sum_{j=1}^{n} f_j \\ n\sum_{i=1}^{m} t_i & m\sum_{j=1}^{n} f_j & mn \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{m}\sum_{j=1}^{n} t_i z_{ij} \\ \sum_{i=1}^{m}\sum_{j=1}^{n} f_j z_{ij} \\ \sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij} \end{pmatrix}, \qquad (6)$$

where m and n are width and height of the planes.

Thus we get two normal vectors $\vec{n}_1$ and $\vec{n}_2$ of pinching planes α and β, respectively. The direction vector $\vec{l}$ and angle φ formed by α and β are given by

$$\vec{l} = \frac{\vec{n}_1 \times \vec{n}_2}{|\vec{n}_1 \times \vec{n}_2|}, \text{ and} \qquad (7)$$

$$\cos\varphi = \frac{\vec{n}_1 \cdot \vec{n}_2}{|\vec{n}_1||\vec{n}_2|}. \qquad (8)$$

This φ corresponds to φ_j in Equation (4).

In practice, two parameters have been introduced: θ_p, a threshold for peak detection, and θ_e ($\theta_p \leq \theta_e$), a threshold for effective peaks. When a peak which is greater than θ_e is found, the calculation described above starts, (forward as well as backward along the time axis), finding peaks and concatenating them, then stops when the peaks become weaker than θ_p. By this method, the accuracy of extraction of frequency components has been improved as compared with a simple threshold method.

5 Bottom–Up Processing

5.1 Clustering for Sound Formation

The next step is a bottom-up clustering of frequency components. As shown in Figure 3, the task is to group frequency components using their features, which is a part of **Problem 2**. The result of this clustering roughly corresponds to a sound human tends to hear as one.

To perform this clustering, we have employed the evaluation–integration model of bottom–up clustering, as illutrated in Figure 4. This model consists of an independent evaluation of features and their integration.

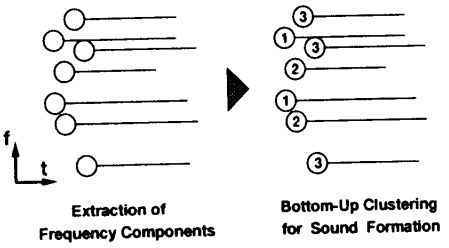

Figure 3: Bottom-Up Clustering for Sound Formation

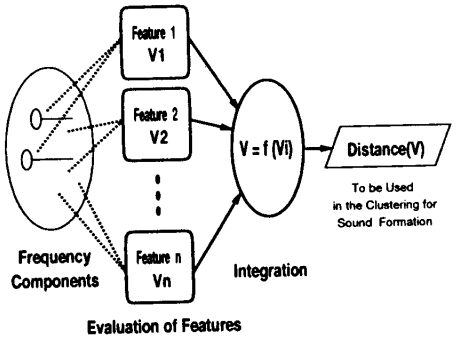

Figure 4: Evaluation–Integration Model of Bottom–Up Clustering

In the current implementation, (1) harmonic mistuning and (2) onset asynchrony between every two frequency components are evaluated in terms of probability of auditory separation. We performed psychacoustic experiments, and as an approximation of their results, two functions for feature evaluation have been derived * .

The probability of auditory separation by harmonic mistuning c_h is evaluated by

$$c_h(u) = \begin{cases} \frac{1}{p}\,u & u < p\,, \\ 1 & \text{otherwise,} \end{cases} \qquad (9)$$

where u (≥ 0) [%] is an amount of mistuning from harmonic relations. The value of parameter p is found to be 2.60 [%] by a least-square fitting of the results of psychoacoustic experiments.

The probability of auditory separation by onset asynchrony c_o is evaluated by

$$c_o(t) = \begin{cases} \frac{40}{S_p}\,t & t < (S_p/40)\,, \\ 1 & \text{otherwise.} \end{cases} \qquad (10)$$

*The psychoacoustic experiments and derivation of approximate functions based on statistical analyses will be described in a separate paper.

where $t(\geq 0)$ [s] is an amount of onset asynchrony and S_p [dB·s] is given by

$$S_p = \frac{a}{f} + \frac{b}{g} + c \qquad (11)$$

where f and g are the frequency in [Hz] and an average onset gradient in [dB/ms] of the earlier frequency component. Using multiple regression analysis, the values of parameters have been chosen to be:

$$\begin{cases} a &= 250\,, \\ b &= 1.11\,, \\ c &= 0.317\,. \end{cases} \qquad (12)$$

Integration of the evaluated values is performed by Equation (13):

$$m = 1 - (1 - c_o)(1 - c_h) \qquad (13)$$

which is a reduced form of Dempster's law of combination, where m is the integrated probability of auditory separation [Kashino *et al.*, 1992].

In the clustering, this m is used as a distance measure. At first, frequency components are roughly grouped by their onset time. In each group:

1. The component of lowest frequency is designated as the center of cluster C_1 ;
2. Scanning the components from low frequency to high, the component whose m value is greater than θ_m is found, and designated as the center of another cluster C_2 ;
3. By scanning the rest of components, the component that all m values (against exsisting centers of clusters) are greater than θ_m is found and designated as the center of cluster C_k ;
4. Scanning continues until no more centers of clusters can be found ;
5. Each remaining component is grouped in clusters if the m values against their centers are less than θ_m ;

and we get sound clusters.

5.2 Clustering for Source Identification

In the case that a sound cluster consists of one note, it is possible to identify sound sources by global characteristics of the sound clusters. This corresponds to another part of **Problem 2**. In the current implementation, the hierarchical and aggromerative method of clustering has been employed. Distance D_s, which is used in the clustering, is defined as

$$D_s = c_1\, f_p + c_2\, f_q + c_3\, t_a + c_4\, t_s \qquad (14)$$

where

f_p : Peak power ratio of the second harmonic to the fundamental component

f_q : Peak power ratio of the third harmonic to the fundamental component

t_a : Attack time

t_s : Sustain time

and each c_k is a coefficient for each parameter.

6 Tone Model - Based Processing

As mentioned in the previous section, each sound cluster roughly corresponds to one sound. However, there are some cases where a sound cluster contains more than one note. For example, it is difficult to separate two simultaneous notes which are an octave apart in pitch solely by bottom-up processing. Tone model-based processing tries to deal with such complicated situations. This process requires stored tone models; and the automatic acquisition of models will be discussed in next section.

One unit of the input of the tone model based processing is called a "processing scope". A processing scope basically consists of one sound cluster, but in the case that multiple sound clusters share at least one frequency component as a member, these sound clusters are included in one processing scope. As tone models, we use such T_k as

$$\begin{aligned} T_k &= \{\vec{a}_{ij}\}, \\ \vec{a}_{ij} &= \left(\frac{p_{ij}}{p_m},\ \frac{f_{ij}}{f_m} \right), \end{aligned} \qquad (15)$$

where p_m is the maximum value of p_{ij} in T_k, and f_m is an average frequency of the fundamental frequency component of T_k. That is, tone model T_k is a matrix of a two-dimensional vector which has normalized power and normalized frequency as elements. Each row of the matrix corresponds to a frequency component, and each column is proportional to time.

In the tone model based processing, "mixture hypotheses" are generated from T_k and matched with a processing scope one by one, to find the mixture hypothesis with the closest fit. In hypothesis generation, the number of maximum simultaneous notes, time / phase shift, and amplitude adjustment are treated as follows:

- *Number of Maximum Simultaneous Notes*
  Currently limited to be three.
- *Time Shift of Tone Models*
  Each tone model is relatively shifted along time axis from −20ms to 20ms by 10ms step, and a mixture hypothesis itself is also shifted from

−20ms to 20ms by 10ms step against the processing scope.

- *Relative Phase of Overlapping Harmonics*
  In case of overlapping harmonics, the relative phase of each frequency component should be taken into account. Letting p_1, p_2, and p_3 be amplitude of frequency components C_1, C_2, and C_3, the amplitude of overlapping component p_a is given by

$$p_a = \{ (p_1 + p_2 \cos\varphi_{12} + p_3 \cos\varphi_{13})^2 + (p_2 \sin\varphi_{12} + p_3 \sin\varphi_{13})^2 \}^{\frac{1}{2}} \quad (16)$$

  where φ_{ij} is a relative phase between C_i and C_j. In the current implementation, the phase of each component is relatively shifted from −90 degrees to 90 degrees by 15 degree steps.

- *Amplitude Adjustment of Tone Models*
  The amplitude of the tone model is adjusted so that the amplitude of the lowest frequency component in the tone model fits the amplitude of the corresponding component in the processing scope.

In matching, distance D_t is defined as

$$D_t = \sum_{i=1}^{F} \sum_{j=1}^{T} |p'_{ij} - p_{ij}| \cdot f'_{ij} \quad (17)$$

where F is the number of frequency components in the processing scope (two corresponding components between a hypothesis and a processing scope are counted as one), T is the number of time samples in the hypothesis, p_{ij} and f_{ij} are the power and frequency of the component in the hypothesis, and p'_{ij} is the power of the component in the processing scope.

Through matching, the hypothesis which gives the minimum D_t value is regarded as the result of sound source separation and identification. In terms of a perceptual sound source separation, this can be viewed as solving **Problem 2** based on "memory of timbres" as mentioned in Section 2.

7 Automatic Tone Modeling

Tone model-based processing requires stored tone models, but is effective in recognizing chords. To realize a registration-free system with high accuracy, we discuss automatic acquisition of tone models.

The perceptual basis of automatic tone modeling is a "old-plus-new heuristic" [Bregman, 1990a], which refers to a hypothesis that a complex sound is interpreted as a combination of *old* (already encountered) sounds as much as possible, and the remainder of the interpretation is perceived as a *new* sound.

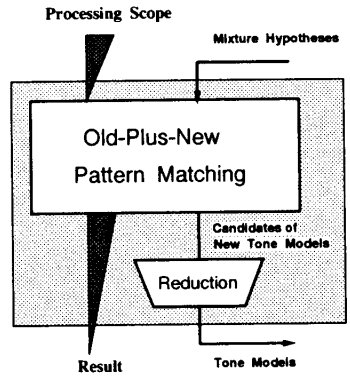

Figure 5: Pattern Matching Module with the Automatic Tone Modeling Mechanism

The process model of this mechanism is shown in Figure 5. As described in the previous section, the matching module calculates the distance D_t in Equation (17) for each hypothesis. In this calculation, the correspondence of frequency components between the processing scope and a hypothesis should be considered. As shown in Figure 6, there are five cases:

Case 1: Frequency components correspond; the component in the hypothesis is a non-overlapped one
(Hit, Non-overlap)

Case 2: Frequency components correspond; the component in the hypothesis is an overlapped one
(Hit, Overlap),

Case 3: No corresponding components in the processing scope; the component in the hypothesis is a non-overlapped one
(No hit, Non-overlap),

Case 4: No corresponding components in the processing scope; the component in the hypothesis is an overlapped one
(No hit, Overlap), and

Case 5: No corresponding components in the hypothesis.

Considering these cases, we extend the distance D_t in Equation (17) as

$$D_t^* = \frac{1}{\sqrt{H(G - H_o)}} D_{tz} \quad (18)$$

where G is the number of frequency components in the hypothesis, H is the number of "hits", H_o is the number of "hit and overlap" components, and

$$D_{tz} = \sum_{i=0}^{G} \sum_{j=1}^{T} |p'_{ij} - p_{ij}| \cdot f_{ij} \cdot z \,, \quad (19)$$

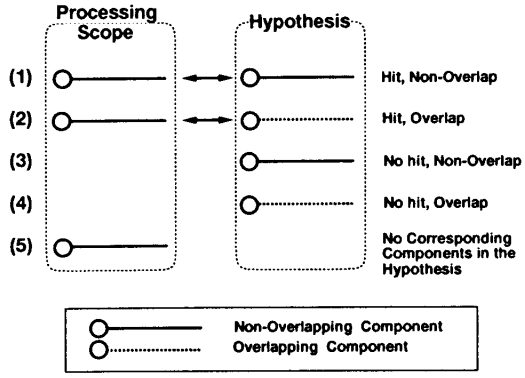

Figure 6: Correspondence of Frequency Components

where z is defined as 0 for the component of "Hit, Overlap", otherwise 1, and the value of p'_{ij} is regarded as 0 in **Case 3** and **Case 4**.

In the matching module, the hypothesis which minimizes D_t^* is found; if the minimum D_t^* value is greater than a threshold, the remainder of the interpretation (that is, such components in the processing scope as **Case 5**) are grouped as a new model candidate.

Thus new tone model candidates are generated. The number of candidates is reduced in two ways: one is a "max set" reduction, meaning that if a candidate is judged as a subset of already exsisting tone models, that candidate will not become a new tone model. Another reduction is a kind of equivalence clustering, similar to the one described in section 5.2 .

In the current implementation, the number of new model candidates generated in each processing scope is limited to one or less (no new model candidates or one). Another heuristic in use is that new model candidates are assumed not to share any overlapping harmonics with the hypothesis.

8 Evaluation of the Model

8.1 Benchmark Test

The process model has been evaluated by a benchmark test. The test sound signal was a random note pattern (125 chords) generated by a PCM sound module, which is controlled by a computer. The note range was limited from 65 to 79 in MIDI note number. An example of the random note pattern is shown in Figure 7.

The test has been performed in three ways: (1) only using bottom-up processing without advance registration of tone models, (2) using automatic tone modeling without advance registration of tone models, and (3) using advance registration of tone models. In the second case, tone models obtained by the bottom-up source identification module were removed manually except for one model, to evaluate the function of automatic tone modeling. The number of simultaneous notes was chosen to be 2 (in piano tone and flute tone) and 3 (2 in piano tone, and 1 in flute tone). Duration of each note was fixed to 720 ms.

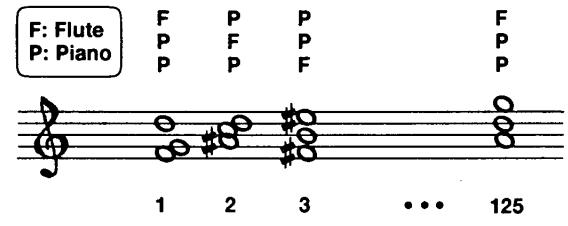

Figure 7: Random Note Pattern (3 Simul. Notes)

Recognition rate R [%] is defined as

$$R = 100 \cdot \left(\frac{right - wrong}{total} \cdot \frac{1}{2} + \frac{1}{2} \right) \qquad (20)$$

where *right* is the number of correctly separated and identified notes, *wrong* is the number of spurious (surplus) notes in the output 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.

8.2 Result

The result of a benchmark test is shown in Figure 8. In the case of two simultaneous notes, the recognition rate for method two is significantly better than that of the bottom-up processing, and is comparable to that of the processing based on manual (advance) registration of tone models. This clearly shows that the automatic tone modeling works effectively. In the case of three simultaneous notes, recognition rate of the automatic tone modeling system deteriorates to the approximately same value as that of the bottom-up processing. The main reasons for this phenomenon can be considered as (1) errors in clustering for sound formation, (2) inaccuracy in distance measure given by Equation (18), and (3) an effect of overlapping harmonics: we expect the recognition rate will be improved by further study on these points.

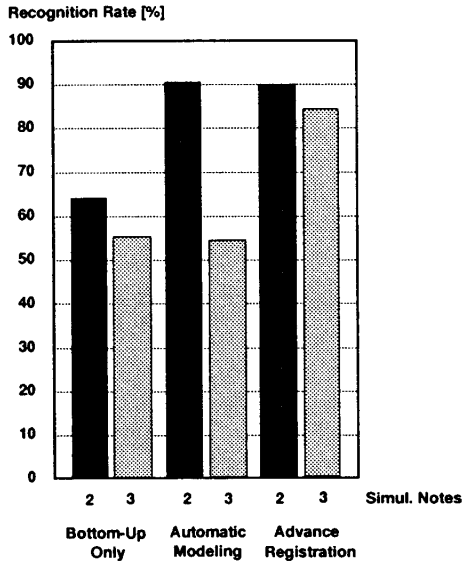

Figure 8: Results of a Benchmark Test

9 Conclusion

We described a configuration, implementation and evaluation of a perceptual sound source separation system. In the system, four cues which humans use for sound separation have been implemented. Specifically, our implementation of the old-plus-new heuristic has permitted automatic tone modeling. It is shown by experimental results that the tone modeling is succeeded when the number of simultaneous notes is two, and consequently the recognition rate has been improved in comparison with the system based on only bottom-up processing. This paper is our first report of an automatic tone modeling approach, and formal benchmark tests and detailed analysis of their results will be reported later. Future work will include effective cooperation of bottom-up processing and model-based processing, a study on stability of automatic tone modeling, further experiments on treatment of overlapping harmonics, and adaptive control of the threshold to extract new tone model candidates.

Acknowledgement

The authors would like to thank Kazuhiro Nakadai for valuable suggestions and for his help in implementation of the system described in this paper.

References

[Bregman, 1990a] Bregman, A. S.: *Auditory Scene Analysis*, MIT Press, (1990).

[Bregman *et al.*, 1990b] Bregman, A. S., Levitan, R. and Liao, C.: Fusion of auditory components: Effects of the frequency of amplitude modulation, *Perception and Psychophysics*, **47**(1), (1990).

[Brown, 1992] Brown, G. J.: *Computational Auditory Scene Analysis : A Representational Approach*, Doctoral Dissertation, Department of Computer Science, University of Sheffield, (1992).

[Darwin *et al.*, 1992] Darwin, C. J., and Ciocca, V.: Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component, *J. Acoust. Soc. Am.*, **91**(6), (1992).

[Flanagan *et al.*, 1985] Flanagan, J. L., Johnston, J.D., Zahn, R. and Elko, G. W.: Computer-steered microphone arrays for sound transduction in large room, *J. Acoust. Soc. Am.* , **78**(5), (1985).

[Hartmann *et al.*, 1990] Hartmann, W. M., McAdams, S. and Smith, B. K.: Hearing a mistuned harmonic in an otherwise periodic complex tone, *J. Acoust. Soc. Am.*, **88**(4), (1990).

[Kashino *et al.*, 1992] Kashino, K., and Tanaka, H.: A Sound Source Separation System using Spectral Features Integrated by the Dempster's Law of Combination, *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, Vol.**51** (1992).

[McAdams, 1989] McAdams, S.: Segregation of Concurrent sounds. I:Effects of frequency modulation coherence, *J. Acoust. Soc. Am.*, **86**(6), (1989).

[Mitchell *et al.*, 1971] Mitchell O. M. E., Ross C. A. and Yates G. H. : Signal Processing for a Cocktail Party Effect, *J. Acoust. Soc. Am.* , **50**(2), (1971).

[Moore *et al.*, 1985] Moore, B. C. J., Peters, R.W., and Glasberg, B. R.: Thresholds for the detection of inharmonicity in complex tones *J. Acoust. Soc. Am.*, **77**(5), (1985).

[Moore *et al.*, 1986] Moore, B. C. J., Glasberg, B. R.: Thresholds for hearing mistuned partials as separate tones in harmonic complexes, *J. Acoust. Soc. Am.*, **80**(2), (1986).

[Moore, 1989] Moore, B. C. J.: *An Introduction to the Psychology of Hearing, Third Ed.*, Academic Press, (1989).

[Nagata *et al.*, 1991] Nagata, Y., Abe, M., and Kido, K.: A study on estimating waveforms of sound sources using many sensors, *J. Acoust. Soc. Jpn.*, **47**(4), (1991) (*In Japanese*).

[Nehorai *et al.*, 1986] Nehorai, A. and Porat, B.: Adaptive Comb Filtering for Harmonic Signal Enhancement, *IEEE Trans. on ASSP*, **34**(5), (1986).

[Parsons, 1976] Parsons, T. W.: Separation of speech from interfering speech by means of harmonic selection, *J. Acoust. Soc. Am.* , **60**(4), (1976).