Annual Report of the Engineering Research Institute
Faculty of Engineering, University of Tokyo
Vol. 51

Received June 22, 1992

A Sound Source Separation System using Spectral Features Integrated by the Dempster's Law of Combination

Kunio KASHINO † and Hidehiko TANAKA ‡

Synopsis

Sound source separation is a very important technique for processing performed music or spoken language on computers, but to realize an effective system, it still requires many technical breakthroughs. One of the problems is to integrate various bits of information which helps separate or identify sound sources. In this paper, a new approach toward sound source separation system is described, which includes evaluation and integration of spectral features. The feature evaluation is based on the rules inferred from the results of psychoacoustic experiments, and the feature integration is performed by the Dempster's law of combination. A sample operation of the prototype system, which shows the efficacy of the proposed approach, is also presented.

1 Introduction

It is hardly ever the case that the sound reaching our ears comes from a single source. Usually the sound arises from several different sources, but we have little difficulty in separating the sound and assigning each part to an appropriate source[6]. Our mechanism to do this has not yet been sufficiently clarified, but if we could realize this ability on computers, it would play an important role in processing acoustic information such as music and spoken language.

Roughly, there have been two types of approaches in the literature. One of them is the approach based on the localization of a target sound source. For example, a microphone array system has been developed to enhance one sound source selectively among several sources[2]. The other approach is the one using the characteristics of a target source. For example, adaptive comb filters have been used to select harmonic frequency components which often belong to the sound from the same source[7].

These approaches are based on a single cue, while human seems to make use of several kinds of cues for sound source separation. The observation of human processing shows that multiple cue mechanism is essential to achieve the excellent flexibility. Therefore we think how to evaluate and integrate multiple cues is an important problem for developing flexible and robust processing on computers.

†Graduate Student, Faculty of Engineering, The University of Tokyo
‡Professor, Faculty of Engineering, The University of Tokyo

(財)新世代コンピュータ技術開発機構の委託研究
「複数種類の楽器演奏を対象とする自動採譜システムの調査研究」の報告

In this paper, we describe a new approach to sound source separation which exploits multiple cues integrated by the Dempster's law of combination.

2 Preliminary Experiments

Among several kinds of cues for human to separate concurrent sounds, we focused on two cues: (1)a harmonic mistuning and (2)an onset asynchrony of the frequency components, as the first step. If a frequency component is mistuned sufficiently in an otherwise harmonic complex tone, the mistuned component tends to be heard as a separate sound, standing out from the complex tone[3]. And if there is an onset asynchrony in a frequency component of a complex tone, the component tends to be perceived as a separate pure tone[6]. In order to establish rules for evaluating these two features, we performed two experiments. In this section, a brief outline of them is described.

The first experiment (Pre.Exp.1) concerns the probability of auditory separation caused by harmonic mistuning[5]. As stimuli, we used complex tones, each of which consisted of six frequency components with equal amplitudes. One of the components was designated as the target component, which was mistuned by 0.5 % step, while the rest of the components were completely in harmonic relation with the fundamental frequency of 200 Hz. The duration of the stimuli was 1600 ms. We also examined co-frequency modulation conditions. Five subjects were required to hear these stimuli in a random order and to report whether the target component was heard as a separate pure tone or not. It was measured 9 times for each subject. The stimulus design of this experiment is shown in Table 1.

Table 1: The stimulus design of the Pre.Exp.1

Target Component	2nd, 4th, or 6th harmonic component
Mistuning	Upto $\pm$ 3.5 % (0.5% step)
Co-Frequency Modulation	No modulation, 1 %, 2%, 5%

An example of the results of this experiment is shown in Figure 1(a). The effect of the frequency of a target was small, so only the result when the target was the component of 400 Hz is shown for simplicity. No significant difference was found by ANOVA between Co-FM conditions. We adopt the linear approximation of the results of this experiment as the rule for evaluating inharmonicity.

$$C_f = \begin{cases} a\,|1-m|, & \text{if } |1-m| < 1/a\ , \\ 1, & \text{else.} \end{cases} \tag{1}$$

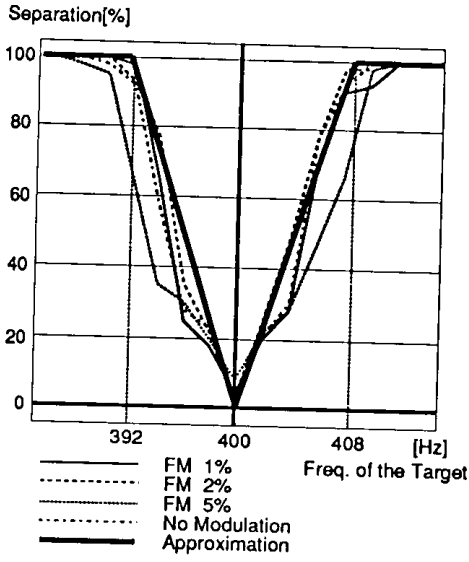
where C_f is the probability of auditory separation, m is a normalized frequency of a mistuned component, and a is a constant. The thick line in Figure 1 denotes the Equation 1, when $a = 50$.

Another experiment (Pre.Exp.2) was arranged to evaluate the probability of human auditory separation caused by an onset asynchrony of frequency components. As stimuli, we used the complex tones consisted of two harmonic frequency components with equal amplitudes. The relative onset times of two components, the rise times of each component, and the fundamental frequency were varied as parameters as is shown in Table 2. The duration of the

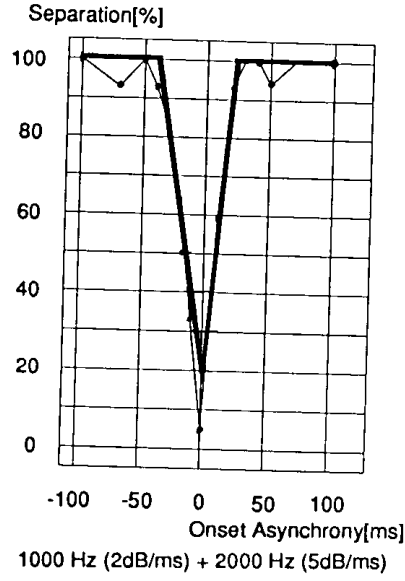
stimuli was 1000 ms. Three subjects were presented with these stimuli in a random order and asked whether a sine tone was heard separately or not. The measurement was repeated 8 times for each subject.

Table 2: The stimulus design of the Pre.Exp.2

Onset Asynchrony	-100ms, -70ms, -50ms, -40ms, -30ms, -20ms, -10ms, 0ms, 10ms, 20ms, 30ms, 40ms, 50ms, 70ms, 100ms
Onset Gradient	5dB/ms, 2dB/ms, 1dB/ms
Fundamental Frequency	200 Hz, 1000 Hz



(a) Auditory separation caused by a harmonic mistuning



(b) Auditory separation caused by an onset asynchrony

Thick lines denote the linear approximations ((a)Equation 1 and (b)Equation 2) of the results.

Figure 1: Examples of the results of the preliminary experiments

An example of the results of this experiment is depicted in Figure 1(b). All of the results cannot be shown here for lack of space, and this is the result when the stimulus is the tone which consists of a 1000 Hz component with 2dB/ms onset gradient and a 2000 Hz component with 5dB/ms onset gradient. We adopt the linear approximation of the results of this experiment (including the results not shown here) as the rule for evaluation of an onset asynchrony.

$$C_f = \begin{cases} \dfrac{1-b}{t_p}\, t + b, & \text{if } t < t_p\ , \\ 1, & \text{else.} \end{cases} \tag{2}$$

where C_f is the probability of auditory separation, t (≥ 0) is the delay of the onset time of

the later component, b is a constant, and t_p is a parameter written as follows:

$$t_p = \frac{p}{f} + \frac{q}{g} + r, \quad p, q, r : const. \tag{3}$$

where f is the frequency of the earlier component, g is the onset gradient of the earlier component.

In the prototype system, the values of the constants are set to as follows:

$$\begin{cases} b = 0.2, \\ p = 4000, \quad q = 50[\mathrm{dB}], \quad r = 10[\mathrm{ms}] \end{cases} \tag{4}$$

which are well matched to the result of this experiment.

3 System Description

The block diagram of the proposed system is shown in Figure 2. Input of this system is assumed to be a monaural acoustic signal, which is the mixture of sounds of several kinds of instruments. Output of this system is MIDI data which includes several MIDI channels, each of which is assigned to one instrument. A graphical display is also available to monitor the results.

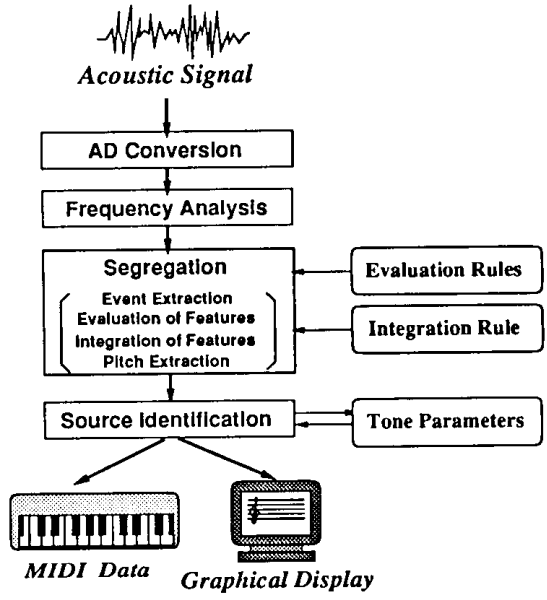


Figure 2: The block diagram of the proposed system

After AD conversion of the input signal, frequency analysis is done to get a spectrogram. In the segregation block, at first, strings of local peaks in the spectrogram are extracted. We

call them "events". Here acoustic processing is converted to symbol processing. Next, the evaluation rules described above are applied to all the events. Then, the integration rule is applied, and we get integrated probability of auditory separation. We adopt the Dempster's law of combination as the integration rule, which can be written as follows:

$$m(A_k) = \frac{\sum_{A_{1i} \cap A_{2j} = A_k} m_1(A_{1i}) m_2(A_{2j})}{1 - \sum_{A_{1i} \cap A_{2j} = \phi} m_1(A_{1i}) m_2(A_{2j})} \quad (A_k \neq \phi) \tag{5}$$

where $m(A_k)$ is the integrated probability, m_1 and m_2 are the basic probabilities inferred from independent bits of evidence, and A_{1i}, A_{2j} $(i, j = 0, 1, 2 \cdots)$ are focal elements[4].

By regarding this integrated probability as the distance between the events, all events are clustered. Each cluster corresponds to a note of a certain tone which human tends to hear as one. Here the fundamental frequency of each cluster is calculated.

Next step is the classification of tones. This process corresponds to the identification of instruments. This is performed by clustering of notes. This clustering is based on the tonal characteristics of each note, such as harmonic structure and power envelope. Note that this process does not necessarily require the registration of the models of target tones in advance, if we don't need to identify the *names* of the instruments. Based on the results of instrument identification, MIDI data is generated, with each instrument assigned to an appropriate MIDI channel. The result of the segregation and instrument identification is also displayed on the screen of a terminal in a simplified score-like format.

4 Experimental Results

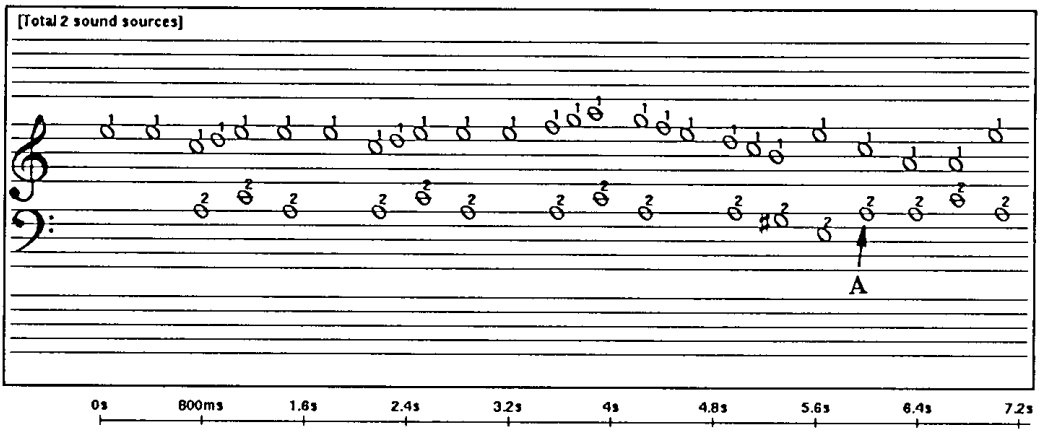
In this section, we describe an example of the operation of the prototype system. We input a part of Vivaldi's violin concerto (L'estro Armonico Op.3, No.6) played by a PCM sound module. Originally it has four parts, but only two of them, the solo part and the cembalo part were performed by flute tone and cembalo tone respectively. Some parameters in the system were manually adjusted.

The final output of the system is shown in Figure 3. Among 42 notes, one note (indicated by Arrow A) was recognized as one octave higher note by mistake, but all other notes were correctly identified.

Sometimes the system fails to output correct results. For example, in the case that two or more instruments simultaneously play the notes which have the same fundamental frequency, the system tends to fail in separation. And the system cannot recognize the sounds of percussion instruments because they have no harmonic relations. To avoid these errors, it will be necessary to introduce more cues to the current version of the system which utilizes only two cues.

5 Conclusion

In this paper, a new approach to sound source separation system is described. Our approach includes feature evaluation based on the rules derived from psychoacoustic experiments, and



The numbers beside each note stand for the serial number of the identified sources.

Figure 3: A result of sound source separation

feature integration by the Dempster's law of combination. Experimental results show the efficiency of the present approach. It will help to improve the performance of the system that we introduce other cues to the system, such as localization of sounds, temporal integration, and amplitude/frequency co-modulation. Applying knowledge processing will be also future work.

References

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

[2] 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).

[3] 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).

[4] Ishizuka, M.: Inference Methods Based on Extended Dempster & Shafer's Theory for Problems with Uncertainty/Fuzziness, *New Generation Computing*, **1**(2), (1983).

[5] Kashino, K. and Tanaka, H.: A study on sound source segregation: Co-FM and Harmonic relation, *Proc. of IEICE Fall Conference*, A-107, (1991). *(In Japanese)* .

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

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