A study on the selective segregation of the target instrument sound from the mixed sound

Masaaki Kubo (110040)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 14, 2003

Keywords: Cooktail party effects, Computational Auditory Scene Analysis, Selective segregation, Auditory segregation model.

1 Introduction

In order to realize a model of segregating the target sound from mixed sound, there are two issues : (1) selecting the target sound from mixed sound and (2) separating the target sound from the overlapped frequency component by other sound sources. Although various current models[1, 2, 3] have been proposed to segregate, it cannot solve both two issues completely.

This paper proposes a model for selectively segregating the target instrument sound from mixed musical instrument sound using acoustical features of musical instrument as knowledge.

2 Selective segregation Model

When we are tried to listen for the target sound from mixed sound if we know the target sound as well before listening it, we can easily segregate it. From this experiential fact, it is thought that one can use some information about the target sound. Thus, in this paper, it is assumed as follows:

1. The target sound surely exists in somewhere in mixed sound, and

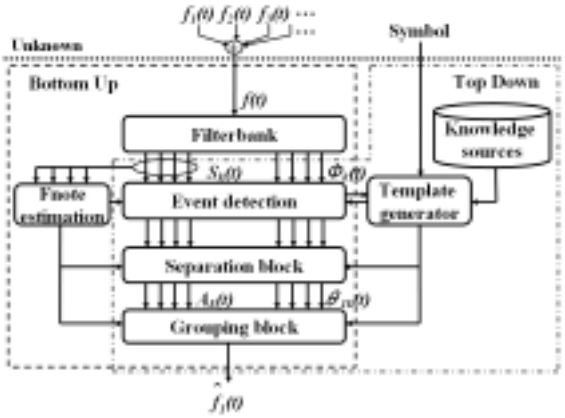

Figure 1: Selective Segregation Model

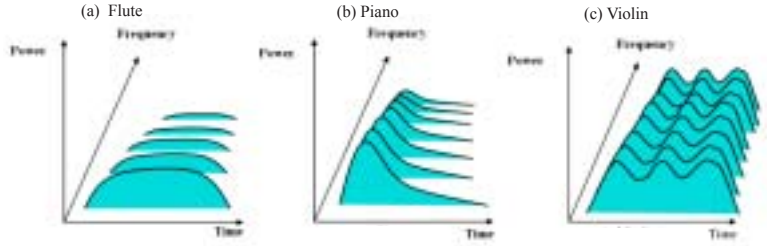

Figure 2: Typical shapes of Template

2. It has the acoustic feature of the target sound as knowledge.

The selective segregation model based on these assumptions is shown in Fig. 1. The inputs of the model are the observed mixed signal $f(t)$ and the target musical instrument name (Symbol). Moreover, the acoustical feature of musical instrument sound is given for source of knowledge from the assumptions 1 and 2, and these are used as information about the target sound. This model is composed of a top-down processing for selecting the position of target sound in mixed sounds and a bottom-up processing for separating the frequency component of the target from frequency component of others. These processings are shown by the dashed line in Fig. 1.

First, the observed mixed signal $f(t)$ is decomposed into its instantaneous amplitude $S_k(t)$ and instantaneous phase $\phi_k(t)$ using the constant bandwidth filterbank[3].

Next, the fundamental frequency of the musical instrument sound contained in mixed sound is estimated in the Fnote estimation block. Here, some peaks are extracted from the auto-correlation function in terms of frequency region at each time of $S_k(t)$, and the candidates of fundamental frequencies are estimated. This block can estimate each fundamental frequency from the mixed sound in which musical instrument sound overlaps in time region.

On one hand, a feature template is generated using the template generator, depending on the Symbol of the target musical instrument. The generated template is the shape on time-frequency region, based on the fundamental frequency, duration, and general acoustical feature (power, spectrum form, etc.) of musical instrument sound. The shape of the standard template for Flute, Piano, and Violin are shown in Fig. 2.

In event detection block, the target sound candidate is extracted from the mixed sound using the obtained information in the above.

1. The fundamental frequency candidates are extracted from the histogram of the candidate in the frequency region of the estimated fundamental frequency.

2. Onset and Offset of each musical instrument sound are estimated from the instantaneous amplitude of the analysis filter corresponding to each fundamental frequency candidate.

3. The harmonic component of each fundamental frequency candidates, obtained in procedure 1, is only extracted from $S_k(t)$ using duration, obtained procedure 2. Then, it becomes each target candidate.

4. The candidate with the highest correlation between the feature template and the target sound candidates is selected as the target sound.

Then, an overlapped frequency component is separated in the separation block. Here, this separation block also uses four psychoacoustically heuristic regularities proposed by Bregman constraints, of an Auditory segregation model [3]. Finally, the target sound $(A_k(t), \theta_{1k}(t))$ is reconstructed using the filterbank. As mentioned in the above, the model is realizing both processing selection of the target sound and separation of overlapped frequency.

3 Simulations

3.1 Segregation of a musical instrument sound from mixed sound

First, the segregating simulations were performed to evaluate the following conditions that used the musical instrument single sound.

(a) Target in the background noise.

(b) Target added to the mixed three sounds.

Five types of mixed signal $f(t)$ were used as simulation stimuli, where the SNRs of $f(t)$ ranged from -10 to 20 dB in 10-dB steps. We used two measures to evaluate the segregation performance of the proposed method.

$$\mathrm{SNR} = 10\log_{10}\frac{\int_0^T f_1(t)^2 dt}{\int_0^T (f_1(t) - \hat{f}_1(t))^2 dt}\,[\mathrm{dB}], \tag{1}$$

and

$$\mathrm{Precision} = \frac{1}{T}\int_0^T \left(10\log_{10}\frac{\sum_{k=1}^K \tilde{A}_k(t)^2 dt}{\sum_{k=1}^K (\tilde{A}_k(t) - A_k(t))^2}\right) dt\,[\mathrm{dB}], \tag{2}$$

where the SNR means that is ratio of signal $f_1(t)$ to noise $f_1(t) - \hat{f}_1(t)$ in the concurrent time region. Precision means that is the temporal average of the segregated error in terms of the instantaneous amplitude $A_k(t)$.

In order to show the advantages of the model, we compare with the results of two simulations : (Top-Down) segregating only the harmonic component of the target sound and (Bottom-Up) segregating the target sound when feature template is not used for the separation block.

The segregation result at the case of setting the target sound to the flute on condition (a) is shown in Fig. 3. Consequently, we can confirm the advantage on the SNR and the Precision of the proposed model as compare with the result of only the top-down processing and the bottom-up processing. Next, The segregation result at the case of setting the target sound to the piano on condition (b) is shown in Fig. 4. A mixed musical instrument sound is four kinds, Flute (A4), Violin (C4), Piano (G3), and Horn (Eb2). For example, when the SNR of mixed signal was 0 dB, it was possible to improve the SNR by about 2.04 dB and the Precision by about 5.46 dB as segregation accuracy, comparing the only top-down processing.

Results show that proposed model could select the target musical instrument sound and could segregate with high accuracy from the mixed signal which mixed noise and other musical instrument sound.

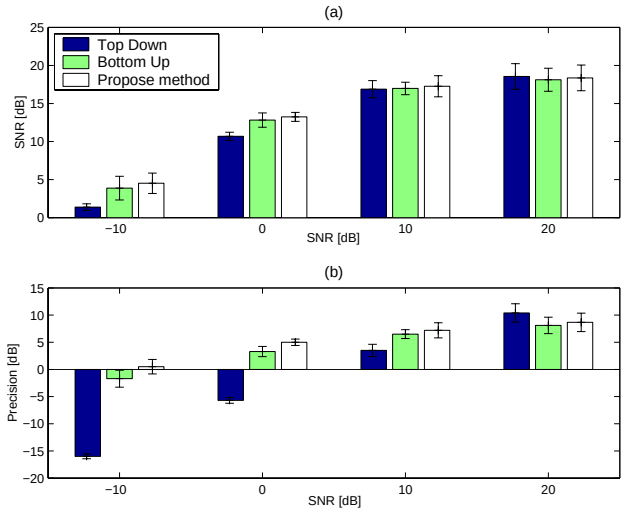

Figure 3: Segregation accuracy for condition (a)[Flute] : (a) SNR, (b) Precision

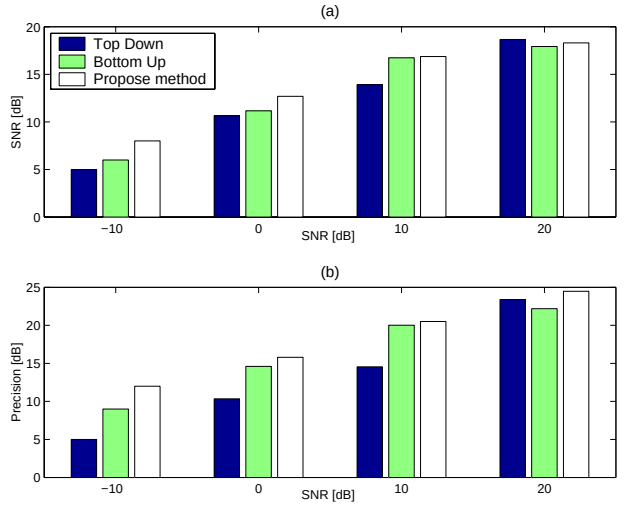

Figure 4: Segregation accuracy for condition (b)[Piano] : (a) SNR, (b) Precision

3.2 Segregation of musical performance sound from mixed sound

This simulation is segregating the performance sound by the target musical instrument from the situation of performing the melody from which

Table3.1. Musical instrument and Music title

	Musical instrument	Music title	Note(Melody)
$f_1(t)$	Piano	chu-rippu	6 Notes(CDECDE)
$f_2(t)$	Flute	kirakiraboshi	7 Notes(CCGGAAG)
$f_3(t)$	Violin	choucho	6 Notes(GEEFEE)
$f_4(t)$	White Noise	-	-

three kinds of musical instruments differ. Here, the tone and sequence of each single note which constitutes a performance are given to the model as new knowledge equivalent to musical score information. These mixed musical instrument performance sounds are shown in Table 1, and an example of segregation is shown in Fig. 5. The mixed signal $f(t)$ is mixed target piano performance $f_1(t)$ with other musical instrument performance and white noise, for its become SNR=0 dB. After segregating a single sound, which constitutes a performance from a mixed signal, the target performance sound $\hat{f}_1(t)$ is obtained Fig. 5 (f) by adding all single sounds. The segregation accuracy of the obtained performance sound was about 10 dB the SNR. Also, It was possible to improve the SNR by about 1.5 dB and the precision by about 2 dB as segregation accuracy, comparing the proposed model with only top-down processing. From the result, it was shown that it is applicable to segregation of musical instrument performance sound because this model applies new knowledge, such as musical score information.

4 Conclusion

This paper proposed a selective segregation model using the acoustical feature of target sound as knowledge. Three segregating simulations were carried out to evaluate the proposed model. These simulations were:

1. Segregating target sound from noise-added target sound,
2. Segregating target sound from mixed four instrument sounds, and
3. Segregating musical performance from mixed sound.

Results of simulations 1 and 2 showed that the proposed model is cannot select target sound but also segregate from mixed sound with high accuracy.

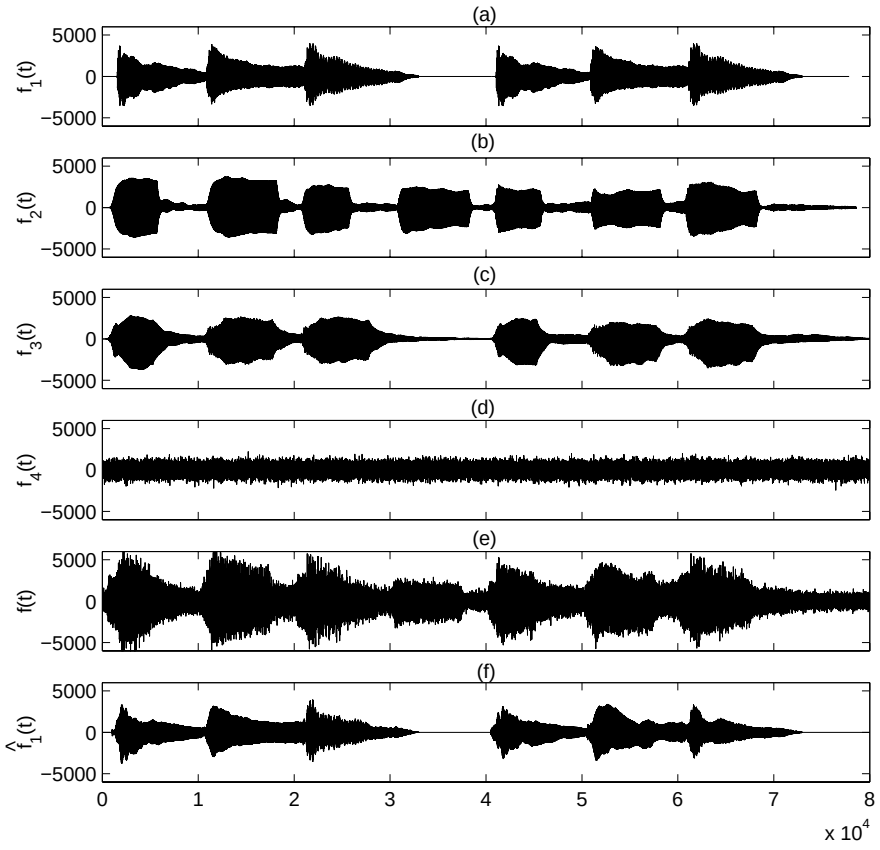

Figure 5: Segregation example : (a)Piano(Target musical performance sound)$f_1(t)$, (b) Flute$f_2(t)$, (c)Violin$f_3(t)$, (d)White noise$f_4(t)$, (e)Mixed signal$f(t)$, (f)Segregated target sound$\hat{f}_1(t)$

Moreover, the result of simulation 3 showed that the proposed model can segregate musical performance sound from three performance mixed sound using musical score information as knowledge.

References

[1] Ellis, D.P.W. , "Prediction-driven computational auditory scene analysis," Ph.D. thesis, MIT Media Lab, Massachusetts, 1996.

[2] Kashino, K., Tanaka, H., "A computational model of auditory segregation of two frequency component - evaluation and integration of multiple cues, " IEICE, J77-A(5), 731-740, 1994.

[3] UNOKI, M. and AKAGI, M., "A Method of Extraction the Harmonic Tone from Noisy Signal Based on Auditory Scene Analysis, " Vol.J82-A, No.10, pp.1497-1507, 1999.