Journal of Signal Processing, Vol. 10, No. 6, pp. 419–431, November 2006

PAPER

A Model-Concept of the Selective Sound Segregation — A Prototype Model for Selective Segregation of Target Instrument Sound from the Mixed Sound of Various Instruments —

Masashi Unoki, Masaaki Kubo, Atsushi Haniu and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {unoki, kubomasa, a-haniu, akagi}@jaist.ac.jp

Journal of
Signal Processing

信号処理

Journal of Signal Processing, Vol. 10, No. 6, pp. 419–431, November 2006

PAPER

A Model-Concept of the Selective Sound Segregation — A Prototype Model for Selective Segregation of Target Instrument Sound from the Mixed Sound of Various Instruments —

Masashi Unoki, Masaaki Kubo, Atsushi Haniu and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {unoki, kubomasa, a-haniu, akagi}@jaist.ac.jp

Abstract We propose a novel model-concept of selective sound segregation based on Auditory Scene Analysis and then describe implementation of a prototype model for selectively segregating a target musical instrument sound from the mixed sound of various musical instruments. This model is extended from our previously proposed model of segregating two acoustic sources (Unoki and Akagi, Speech Communication, 27, 261-279, 1999). The extended model consists of two blocks: our previous model as bottom-up processing and a selective processing based on knowledge sources as top-down processing. A novel idea is to segregate a target sound from the mixed sound based on the top-down information as an interaction between bottom-up and top-down processing. To demonstrate the ability of the proposed model, we carried out three simulations: (i) segregation of the target sound from noisy sound (signal extraction); (ii) segregation of the target sound from four mixed sounds (concurrent separation); and (iii) segregation of the target performance sound from mixed sound (selective segregation). Simulation results showed that the proposed model could adequately selectively segregate not only the target instrument sound, but also the target performance sound, from the mixed sound of various instruments; this is not possible when using only bottom-up or top-down processing. The advantage provided by this model-concept led to significantly improved results. This model can be applied to selective speech-sound segregation, enabling its extension to computational modeling of the mechanisms of a human's selective hearing system.

Keywords: cocktail party effect, computational auditory scene analysis, selective sound segregation, musical instrument

1. Introduction

1.1 General issue

A human can easily selectively listen to a desired sound (a target sound) in a real environment that simultaneously contains various kind of sound such as conversation speech, instrument sounds, animal songs, noises, reflections, etc. Let us, for example, consider a general problem of selective listening of the target sound in the case of sound mixtures as shown in Fig. 1. Here, the sound of four speech signals, independently generated by four speakers as different words, three musical performances (independently played on a flute, piano, and violin) as background music, and some other background noises, are mixed together under unpredictable conditions. Here, we assume that the task is to try to selectively hear a target sound (e.g., a familiar Japanese word /kon-nichiwa/ ("Hello" in English) pronounced by a Japanese male speaker standing behind three other people) from among the mixed sound (a mixture of speech from various people amidst background music) under noisy conditions. In this task, we can easily selectively listen to the target sound if we know who the target is (or what the target is) and we know his voice as well (or we have previously listened to it). Of course, we can more easily selectively listen to it if we are fluent Japanese speakers and this word is familiar to us. However, we may not be able to selectively listen to the target sound if we do not know what it is and/or we have never heard it (e.g., if we cannot speak Japanese in this case)[1].

[1] In general, it is believed that we also use visual information such as lip reading, etc. to selectively understand the target

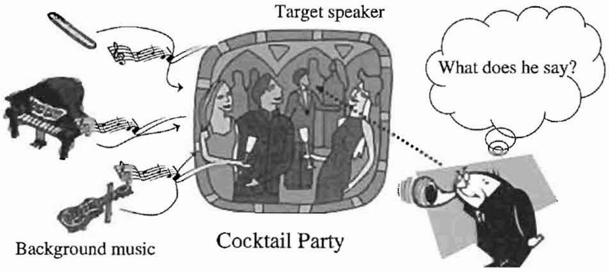

Fig. 1 The selective sound segregation problem known as the cocktail party problem: There is a mixed sound of three musical instruments being played (flute, piano, and violin), four speech signals being uttered by two male and two female speakers, and background noise. The task is to selectively segregate the speech sound uttered by the male standing behind three people from the mixed sound.

This ability seems to be strongly related to the availability of prior information (stored knowledge gained through experience) that allows us to identify the target sound within the mixed sound. In general, hearing a target sound in this type of situation (the task) depends on what is called the "cocktail party effect" [1]. The ability of the human auditory system seems to play a significant role in this effect. If this ability of selective listening to a target sound can be implemented as a computational model, it should improve the performance of preprocessors and enable robust speech-recognition systems and hearing-aid systems.

1.2 Previous approaches

The straightforward approach of signal processing for speech separation has been investigated by many researchers, who have proposed many separation methods. For example, to enable robust speech recognition [2], noise reduction or suppression techniques [3] and speech enhancement methods [4] can be used. Signal processing has been investigated based on signal estimation using a linear system [5, 6] and signal estimation based on a stochastic process for signal and noise [7]. In practice, though, it is difficult to construct a computational model that can process signals in a way related to selective separation (as hearing in this research field), because the signals exist in a concurrent time-frequency region and this problem is an ill-posed inverse problem. Therefore, we need to use reasonable constraints to solve the problem.

Recently, blind source separation (BSS) methods based on independent component analysis (ICA) have been proposed (e.g., [8,9]). In these methods, measuring independence among the source signals is used to separate source signals from the mixed signals. The methods can enable good source separation in artificial environments or when all source-signals satisfy the assumptions; in addition, the number of microphones used in these models should be greater than or equal to the number of source-signals. However, in these methods, it is difficult to selectively separate only the target signal, as is desired by the listener or model, from mixed signals because we do not know which is the desired source-signal among the multiple outputs. Moreover, this issue is also due to the permutation problem in the frequency domain BSS method with ICA. On the other hand, one factor contributing to the cocktail party effect (selective listening) is regarded as a function of an active scene analysis system, called auditory scene analysis (ASA) [10]. Bregman has reported that the human auditory system uses four psychoacoustically heuristic regularities related to acoustic events to solve the ASA problem [11] (see Appendix). One possible approach has been to develop sound separation models based on ASA to try to solve the problem by using Bregman's regularities. This is referred to as "computational auditory scene analysis (CASA)"[2]

In this research field, selective listening can be referred to as "selective segregation" in which the segregation consists of the following sequential processing stages: decomposition of acoustical features; separation of each set of features corresponding to each sound source; and then grouping each separated set of features to merge each sound source. The aim of this attempt is to obtain a model that enables us to solve the selective segregation problem by applying reasonable constraints (Bregman's regularities [11]; see Appendix) to sounds and the environment. There are two main types of CASA-motivated segregation models [12], based on either bottom-up (e.g., [13–19]) or top-down (e.g., [20–22]) processes[3].

Most models based on bottom-up processing can adequately separate or extract a target signal such as a harmonic complex tone (vowel) and artificial sinusoidal signals from the mixture in a concurrent time-frequency region. However, these are not good at grouping them in order to separate a natural speech

sound. However, in this example, we assume that clues regarding only the actual sound information can be used.

[2] CASA is a computational version of ASA and includes a function similar to this ability of the human auditory system.

[3] In particular, in the case of musical sound, this is also called "music scene analysis" [22,23], and models have been proposed for extracting significant information (musical sequences, rhythm, etc.) regarding a target sound from a mixed sound and to understand the target [20,22,24]. Models have also been proposed for identifying a target sound as form of music scene analysis [23,25].

signal such as consonants and long sentences. In contrast, models based on top-down processing can deal with realistic signals for selection and/or extraction as the grouping process. However, these models cannot completely separate the target components from the mixture so the extracted signal still includes residual signals such as noise and artifacts.

1.3 Motivation and contribution

We think that selective segregation is best achieved, however, through the interaction of bottom-up and top-down processes. That is, by (1) precisely selecting the position of the target sound in the mixed sound based on our knowledge and then (2) completely segregating the target sound from the other sounds in the concurrent time-frequency region. However, since top-down and bottom-up processes focus only on either (1) or (2), respectively, each alone cannot be used to realize a selective sound segregation model. Therefore, to realize a selective sound segregation model as shown in Fig. 1, we have to resolve two issues: (I) how to precisely select the target sound within a real environment (selection), and (II) how to completely separate the target sound from the mixed sound in which overlapped components exist in a concurrent time-frequency region (separation).

The ultimate goal of our work is to construct a selective sound segregation model that can be applied to any real world sound as a realistic problem. In this paper, as the first step, we consider a selective sound segregation problem for instrument sound mixtures using a single-channel method (monaural processing, without a direct cue) as a basic problem. We propose a novel model-concept of the selective segregation where a combination of top-down and bottom-up processing is used, and then describe implementation of a prototype model to selectively segregate a target instrument sound from the mixed sound of instruments. Our main aim in this paper is to demonstrate the ability to solve the above two issues by reasonably combining top-down and bottom-up processing.

This paper is organized as follows. In Sec. 2, we describe the sound segregation problem that we dealt with assumptions and our model-concept of the selective sound segregation. In Sec. 3, we describe the algorithm and model implementation. In Sec. 4, we discuss our simulation results. Section 5 gives our conclusions and perspectives regarding further work.

2. Selective Sound Segregation Model

2.1 Simplified segregation problem

Consider again the selective sound segregation problem shown in Fig. 1. This is a general problem so it should be simplified so that we can deal with it as a basic selective segregation problem. Thus, in this paper, we consider the problem of selective sound segregation in the case of mixed instrument sounds as shown in Fig. 1. The sound of three musical performances, independently played on a flute, piano, and violin, are mixed together. When we try to selectively listen to the piano sound from the mixed sound, we can easily succeed as we imagine the piano sound. Thus, our conceptual idea is that the bottom-up process decomposes all features from the mixed sound and presents us with all the answer candidates for solving the problem, and then the top-down processing selects a reasonable solution from all conditions and merges it according to our knowledge.

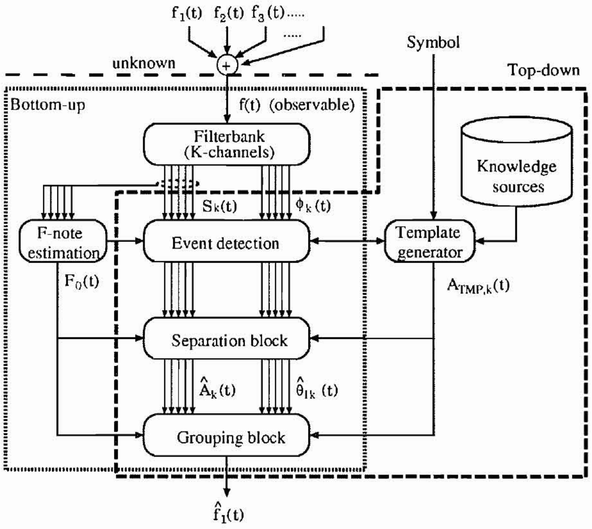

Fig. 2 Selective sound segregation model

2.2 Model concept

Figure 2 shows the proposed selective segregation model based on our concept. This model consists of the two types of processing: top-down processing to select the position of the target sound in the mixed sound (to resolve issue (1), as shown by the dashed-line in Fig. 2), and bottom-up processing to separate the target sound from the other sounds in the concurrent time-frequency region (to resolve issue (2), the dotted line in Fig. 2). The bottom-up processing is the same method proposed in [18,19], but it has been modified so that it can be combined with top-down processing.

2.3 Assumption and definition

In this model, the original signals ($f_1(t)$, $f_2(t)$, $f_3(t)$, and so on) are not known, nor is it known how many different sounds there are. The only model inputs are the observed mixed signal $f(t)$ (i.e., single microphone input) and a knowledge key such as the symbol for the target instrument name (here, this is for $f_1(t)$; e.g., "piano"). To investigate whether the proposed model using top-down information can segregate the target sound from the mixed sound, we assume that the exact target sound can exist anywhere in the mixed sound[4], and knowledge about the target sound can be represented through the acoustical features. Thus, the key enables the model to obtain information regarding the acoustical features of the target sound from the knowledge sources.

This model concept is based on the problems associated with segregating two acoustic sources. This fundamental problem is defined as follows [18, 19].

First, only the mixed signal $f(t)$, where $f(t) = f_1(t) + f_2(t)$, can be observed and $f(t)$ is then decomposed into its frequency components by a K-channel filterbank. The output of the k-th channel $X_k(t)$ is represented by

$$\begin{aligned} X_k(t) &= X_{1,k}(t) + X_{2,k}(t) & (1)\\ &= S_k(t)\exp(j\omega_k t + j\phi_k(t)) & (2)\end{aligned}$$

where $S_k(t)$ and $\phi_k(t)$ are the instantaneous amplitude and phase, respectively. If the outputs of the k-th channel, which correspond to $f_1(t)$ and $f_2(t)$, are assumed to be

$$\begin{aligned} X_{1,k}(t) &= A_k(t)\exp(j\omega_k t + j\theta_{1k}(t)) & (3)\\ X_{2,k}(t) &= B_k(t)\exp(j\omega_k t + j\theta_{2k}(t)) & (4)\end{aligned}$$

then the instantaneous amplitudes $A_k(t)$ and $B_k(t)$ can be determined as

$$\begin{aligned} A_k(t) &= \frac{S_k(t)\sin(\theta_{2k}(t) - \phi_k(t))}{\sin\theta_k(t)} & (5)\\ B_k(t) &= \frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{\sin\theta_k(t)} & (6)\end{aligned}$$

where $\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$, $\theta_k(t) \neq n\pi, n \in \mathbf{Z}$, and ω_k is the center frequency of the k-th channel.

We cannot uniquely determine $A_k(t), B_k(t), \theta_{1k}(t)$, and $\theta_{2k}(t)$, though, without some constraints. This is easily understood by considering the above equations. The problem, therefore, is the ill-posed inverse problem. To solve this problem, we previously proposed a basic model that uses constraints related to the four Bregman regularities [18, 19], as shown in Table 1 (see Appendix and [18, 19] for details).

The basic problem given in the above concerns two-sound segregation. Thus, in this paper, the problem is set so that $f_1(t)$ is the target sound selected by top-down processing and $f_2(t)$ is the other mixed sound (i.e., $f_2(t) + f_3(t) + \cdots + f_M(t)$). The problem (to selectively segregate $f_1(t)$ from $f(t)$) is then solved using the modified solution based on ASA [18, 19].

3. Model Implementation

The proposed model is implemented in six blocks: the filterbank, F0 (F-note) estimation, template generation, event detection, separation block, and grouping block (Fig. 2).

3.1 Filterbank

The filterbank decomposes the observed signal $f(t)$ into complex spectra $X_k(t)$. It is designed as a constant narrow-band filterbank using an FIR-type bandpass (gammatone) filter $g_k(t)$ as follows.

$$\begin{aligned} X_k(t) &= \int_0^t g_k(\tau) * f(t-\tau)d\tau & (7)\\ g_k(t) &= At^{N-1}\exp(-2\pi b_w t)\exp(-j\omega_k t) & (8)\end{aligned}$$

where $N = 4$, $b_w = 20$ (20-Hz bandwidth), $\omega_k = 2\pi k f_c$, $k = 1, 2, \cdots, K$, $K = 500$, $f_c = 10$ Hz, and the sampling frequency is 20-kHz (see [17] for details).

The instantaneous amplitude $S_k(t)$ and phase $\phi_k(t)$ are determined using the Hilbert transform technique with regard to $X_k(t)$ [18, 19].

3.2 F0 (F-note) estimation block

The F0 (fundamental frequency) estimation block determines the candidates for the note of the musical instrument sound by obtaining peaks in the autocorrelation function $r_t(\ell)$ in terms of channel number k (the frequency region) at each time t of $S_k(t)$s, as follows.

$$r_t(\ell) = \sum_{k=1}^{K_0} \overline{S_k(t)} \cdot \overline{S_{k+\ell}(t)} \qquad (9)$$

where $\ell = 1, 2, \cdots, K_0$, $K_0 = K/2$, and $\overline{S_k(t)}$ is the center-clipped $\log(S_k(t))$. Center-clipping is a process that replaces the values of $\log(S_k(t))$ with regard to channel number k with the mean value of $\log(S_k(t))$ when the original values are less than the mean value. This technique emphasizes the peaks in $r_t(\ell)$. Here, these peak frequencies in $r_t(\ell)$ correspond to the frequencies related to the F0 of each sound. Therefore, these can be regarded as the candidates to be the F0 of the target sound. In this paper, F0 is referred as the F-note for musical sound.

The histograms for each F-note candidate are then calculated according to the time axis in the time-frequency region. The M-best candidates with higher histogram values are passed to the event-detection

[4] In practical cases, it is uncertain whether the target signal exactly exists anywhere in the mixed sound. The aim of this work is to show it is possible to model selective sound segregation using prior top-down information such as knowledge of the target. Thus, this assumption is to test the model-concept. Segregation issues that are not consistent with this assumption, such as illusionary hearing, are beyond the scope of this paper.

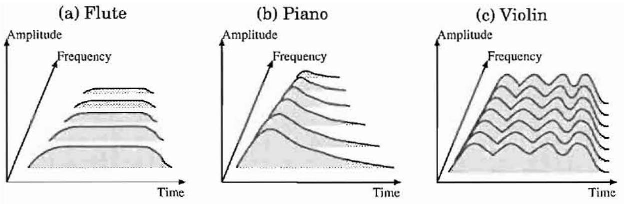

Fig. 3 Schematic shape of a standard template for the target instrument: (a) flute, (b) piano, and (c) violin

block and the final estimated F-note of the target sound, $F_0(t)$, is determined in this block. In this paper, $F_0(t)$ fluctuates in steps, and the temporal differentiation of $F_0(t)$ is zero in all segments. As a result, this paper assumes that $E_{0,R}(t) = 0$ in Table 1 (ii) for each segment. Most of the segments correspond to each F-note duration in the target instrument sound.

3.3 Template generation block

The template generator produces an acoustical template from the knowledge sources, depending on the target sound symbol. The generated template is composed of the shape of the instantaneous amplitude in the time-frequency region, based on the fundamental frequency (F-note), duration, and general acoustical features of the musical instrument sound. The schematic shapes of the standard template for flute, piano, and violin are shown in Fig. 3. The standard templates considered in this paper were set as the averaged instantaneous amplitude of the target sound under various conditions (normalized duration and normalized harmonic components based on the F-note, etc.). The template, used in the separation block, is then reshaped as a function of the segregation duration and F-note. This can be extended by analyzing all of the sounds, as was done in [22,26,27], to obtain a realistic generated template.

3.4 Event detection block

The event detection block uses a template of the target to determine the concurrent time-frequency region of the target sound. In this block, the F-note ($F_0(t)$) of the target is selected from the M-best candidates of the F-note while the block searches to check whether the extracted amplitude based on the harmonicity of each F-note candidate matches the generated template based on the target symbol (Fig. 3). The matching degree is determined as a measure of correlation between each estimated amplitude and the generated template in which the duration of the template is rearranged to equal the duration of the F-note candidate. This corresponds to constraint (iii) in Table 1. The estimated target event can then be obtained from the candidate with the highest correlation. The onset and offset of the target instrument sound, $T_{k,\mathrm{on}}$ and $T_{k,\mathrm{off}}$, are determined from the estimated instantaneous amplitude based on the harmonic components of the selected fundamental frequency. This corresponds to constraint (i) in Table 1.

3.5 Separation block

The separation block determines $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ from $S_k(t)$ and $\phi_k(t)$ using constraints (ii) and (iv) in the determined concurrent time-frequency region. Constraint (ii) in Table 1 is implemented such that $C_{k,R}(t)$ and $D_{k,R}(t)$ are linear ($R = 1$) polynomials, which reduces the computational cost of estimating $C_{k,R}(t)$ and $D_{k,R}(t)$. Under this assumption, $A_k(t)$ and $\theta_{1k}(t)$, which can be allowed to undergo a temporal change in region, constrain the second-order polynomials ($A_k(t) = \int C_{k,1}(t)dt + C'_{k,0}$ and $\theta_{1k}(t) = \int D_{k,1}(t) + D'_{k,0}$). Then, by substituting $dA_k(t)/dt = C_{k,R}(t)$ into Eq. (5), we end up with the linear differential equation of the input phase difference $\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$. By solving this equation, a general solution is determined by

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{S_k(t)\cos(\phi_k(t) - \theta_{1k}(t)) + C_k(t)}\right) \tag{10}$$

where $C_k(t) = -\int C_{k,R}(t)dt - C_{k,0} = -A_k(t)$ [18,19].

In the segment $T_h - T_{h-1}$ of each instrument duration, which can be determined by $E_{0,R}(t) = 0$, $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ are determined through the following steps. First, the estimated regions, $\hat{C}_{k,0}(t) - P_k(t) \le C_{k,1}(t) \le \hat{C}_{k,0}(t) + P_k(t)$ and $\hat{D}_{k,0}(t) - Q_k(t) \le D_{k,1}(t) \le \hat{D}_{k,0}(t) + Q_k(t)$, are determined by using the Kalman filter, where $\hat{C}_{k,0}(t)$ and $\hat{D}_{k,0}(t)$ are the estimated values and $P_k(t)$ and $Q_k(t)$ are the estimated errors. Next, the candidates of $C_{k,1}(t)$ at any $D_{k,1}(t)$ are selected by using spline interpolation in the estimated error region. $\hat{C}_{k,1}(t)$ is then determined by

$$\hat{C}_{k,1} = \underset{\hat{C}_{k,0}-P_k \le C_{k,1} \le \hat{C}_{k,0}+P_k}{\arg\max} \frac{<\hat{A}_k, A_{\mathrm{TMP},k}>}{||\hat{A}_k|| \cdot ||A_{\mathrm{TMP},k}||} \tag{11}$$

where $\hat{A}_k(t)$ is obtained through spline interpolation and $A_{\mathrm{TMP,k}}(t)$ is the reshaped standard template as shown in Fig. 3, as a function of the separation duration ($T_{k,\mathrm{on}}$ to $T_{k,\mathrm{off}}$) and F-note $F_0(t)$. Finally, $\hat{D}_{k,1}(t)$ is determined by

$$\hat{D}_{k,1} = \underset{\hat{D}_{k,0}-Q_k \le D_{k,1} \le \hat{D}_{k,0}+Q_k}{\arg\max} \frac{<\hat{A}_k, A_{\mathrm{TMP},k}>}{||\hat{A}_k|| \cdot ||A_{\mathrm{TMP},k}||} \tag{12}$$

Table 1 Constraints corresponding to Bregman's regularities (See Appendix for details)

Regularity (Bregman, 1993)	Constraint (Unoki and Akagi, 1999)	
(i) common onset/offset	synchronous of onset/offset	$\|T_{\rm S} - T_{k,\rm on}\| \leq \Delta T_{\rm S}, \|T_{\rm E} - T_{k,\rm off}\| \leq \Delta T_{\rm E}$
(ii) gradualness of change	piecewise-differentiable polynomial approximation	$dA_k(t)/dt = C_{k,R}(t), d\theta_{1k}(t)dt = D_{k,R}(t)$ $dF_0(t)/dt = E_{0,R}(t)$
(slowness)	(Kalman filtering)	
(smoothness)	(spline interpolation)	$\sigma_A = \int_{t_a}^{t_b}[A_k^{(R+1)}(t)]^2 dt \Rightarrow \min$ $\sigma_\theta = \int_{t_a}^{t_b}[\theta_{1k}^{(R+1)}(t)]^2 dt \Rightarrow \min$
(iii) harmonicity	multiples of the fundamental frequency	$n \times F_0(t), \qquad n = 1, 2, \cdots, N_{F_0}$
(iv) changes occurring in the acoustic event	correlation between the instantaneous amplitudes	$\frac{A_k(t)}{\|A_k(t)\|} \approx \frac{A_\ell(t)}{\|A_\ell(t)\|}, \qquad k \neq \ell$

The difference between our proposed model and the previous model is that we use a template of $A_{\rm TMP,k}(t)$ instead of the averaged $\hat{A}_k(t)$ [18, 19]. These equations mean we can determine a unique solution from among the candidates. Since $\theta_{1k}(t)$ and $\theta_k(t)$ are determined from $\hat{D}_{k,1}(t)$ and $\hat{C}_{k,1}(t)$, we can determine $A_k(t)$, $B_k(t)$, and $\theta_{2k}(t)$ from Eq. (5), Eq. (6), and $\theta_{2k}(t) = \theta_k(t) + \theta_{1k}(t)$, respectively.

3.6 Grouping block

The grouping block merges the instantaneous amplitudes $A_k(t)$s and phases $\theta_{1k}(t)$s in the concurrent time-frequency region of the target using constraints (i) and (iii) in Table 1 to reconstruct $X_{1k}(t)$ in Eq. (3). It then reconstructs them into the segregated signal $\hat{f}_1(t)$ using inverse processing of the filterbank All processing in the grouping block is done across the channel, following Eqs. (3), (5), (10)-(12) step-by-step, so the permutation problem does not occur in the proposed method.

3.7 Example

First, we assume that the target sound $f_1(t)$ in this example is a flute sound (A4) and the mixed sound $f(t)$ is a combination of piano (G3), flute (A4), horn (Eb2), and violin(C4). The observed signal $f(t)$ is then decomposed into $S_k(t)$ and $\phi_k(t)$ by the constant narrow-band filterbank in Eq. (7). Figure 4(a) shows the magnitude of the filterbank output, $S_k(t)$s, in which the frequency range of $S_k(t)$ is restricted to a range from 100 Hz to 1 kHz. Black parts show the harmonics of each musical sound. The harmonics of piano, flute, horn, and violin were located from 1200 to 3200, from 2000 to 4000, from 4200 to 5200, and from 100 to 7800 in the sample number, respectively.

Next, the F-note estimation block determines the candidates for the target sound note. Figure 4(b) shows the auto-correlation function at each time (1)-(3) from Fig. 4(a). Panels (b-1), (b-2), and (b-3) show the candidates (some of the peaks) for the F-note of each musical sound at each time (at the (1) 1000, (2) 2800, and (3) 4500 points). Figure 4(c) shows seven F-note candidates (seven peaks for each time). One of these peaks corresponds to the F-note of one of the musical sounds. In panel (b-1), the maximum peak is at about 270 Hz and corresponds to the F-note of the violin (C4) at the 1000 point in Fig. 4(a). In panel (b-2), the maximum peak is at about 450 Hz and corresponds to the F-note of the flute (A4) at the 2800 point in Fig. 4(a). In panel (b-3), the maximum peak is at about 160 Hz and corresponds to the F-note of the horn (Eb2) at the 4500 point in Fig. 4(a).

The template generation block produces the averaged instantaneous amplitude of the target (flute) as shown in Fig. 3(a). The event-detection block determines the concurrent time-frequency region of the target "flute" using the generated template. The F-note of the flute is selected from the M-best candidates ($M = 7$) for the F-note in Fig. 4(c) while this block determines whether the extracted amplitudes based on the harmonicity of each F-note candidate matches the generated template. $T_{k,\rm on}$ and $T_{k,\rm off}$ are then determined from the estimated instantaneous amplitude based on the harmonicity of the selected F-note. In this example, they are determined as being at about the 1200 and 3200 points, respectively. $A_{\rm TMP,k}(t)$ is the reshaped standard template, as shown in Fig. 3, that is a function of the separation duration ($T_{k,\rm on}$ to $T_{k,\rm off}$) and the estimated F-note $F_0(t)$.

Finally, the separation block determines $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ from $S_k(t)$ and $\phi_k(t)$ using Eqs. (10)-(12) in the determined concurrent time-frequency region. The grouping block merges the $A_k(t)$ and $\theta_{1k}(t)$ in the concurrent time-frequency region of the target (flute), and then reconstructs them into the segregated signal $\hat{f}_1(t)$.

4. Simulations

To show that the proposed model can selectively and precisely segregate the target instrument sound $f_1(t)$ from the observed sound $f(t)$, we carried out

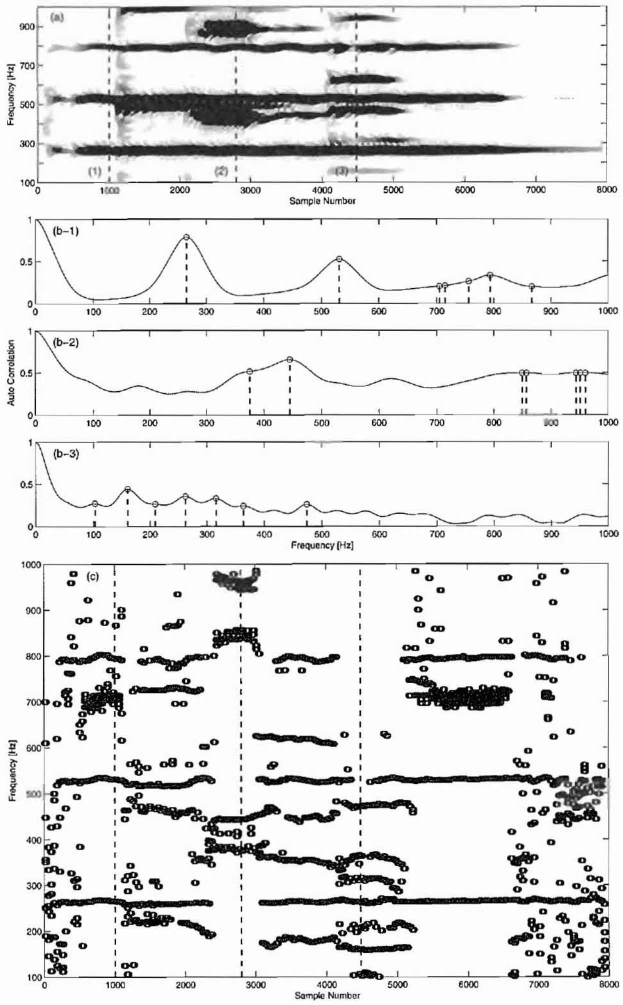

Fig. 4 F-note estimation: (a) sound spectrogram, (b) autocorrelation functions at each time (1)-(3), and (c) F-note candidates

three types of simulation: (i) segregation of the target sound $f_1(t)$ from noisy sound $f(t)$ (with added white noise) [signal extraction]; (ii) segregation of the target sound $f_1(t)$ (piano, flute, horn, or violin) from four mixed sounds (piano, flute, horn, and violin) [concurrent separation]; and (iii) segregation of the target performance sound $f_1(t)$ from mixed sound $f(t)$ [selective segregation]. The first two simulations correspond to typical engineering problems such as signal extraction and concurrent signal separation. Especially, in the second simulation, four other fundamental frequencies existed in the mixed sound, so we refer to this as "concurrent separation" here. The third simulation corresponds to a more general segregation problem, so we refer to this as "selective segregation". A mixed signal $f(t)$ was used as the simulation stimuli in each simulation, where the SNR of $f(t)$ ranged from -10 to 20 dB in 10-dB steps. These original signals were generated using a tone-generator (YAMAHA, MU-2000) [28].

To evaluate the segregation performance of our proposed method, we used the following two measures. These measures revealed whether the model precisely segregated the target from the mixed sound in terms of amplitude as well as waveform. Both measures show improvement if they become higher positive values.

$$\mathrm{SNR} = 10\log_{10}\frac{\int_0^T f_1(t)^2 dt}{\int_0^T \left(f_1(t)-\hat{f}_1(t)\right)^2 dt} \quad [\mathrm{dB}] \quad (13)$$

Precision

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

$$(14)$$

Moreover, to show the advantages of the proposed model, we compared the model performance when (a) using only top-down processing and (b) using bottom-up processing. Here, the proposed model in which the separation block is not active (i.e., we only extract the harmonic component of the target sound, and do not segregate it in each channel) is used for the top-down processing, whereas the previous model [18,19] is used for the bottom-up processing (i.e., we do not use any template).

4.1 Simulation 1

The results of the first simulations for flute (E3, A4, D4, or C5) are shown in Fig. 5, where $f(t)$ was the target flute sound mixed with white noise. Each bar height and error bar shows the averaged value and the standard deviation, respectively. All three methods led to almost the same degree of improvement when the SNR was high, because $A_k(t) \approx S_k(t)$ and $\theta_{1k}(t) \approx \phi_k(t)$, but the improvement with the proposed method was biggest of all. For example, when the SNR of the mixed signal was 0 dB, it was possible to improve the SNR by about 14.9 dB from $f(t)$, and to improve the SNR by about 2 dB and the precision by about 5 dB in terms of segregation accuracy (compared with the top-down processing).

This comparison shows the importance of separating each component from the overlapped components in each channel. These results confirm that the proposed model can segregate the target sound from the mixed sound as well as the two-sound segregation model proposed by [18,19].

4.2 Simulation 2

The results of the second simulations for flute (A4) are shown in Fig. 6, where $f(t)$ was the target flute (A4) sound mixed with piano (G3), violin (C4), and

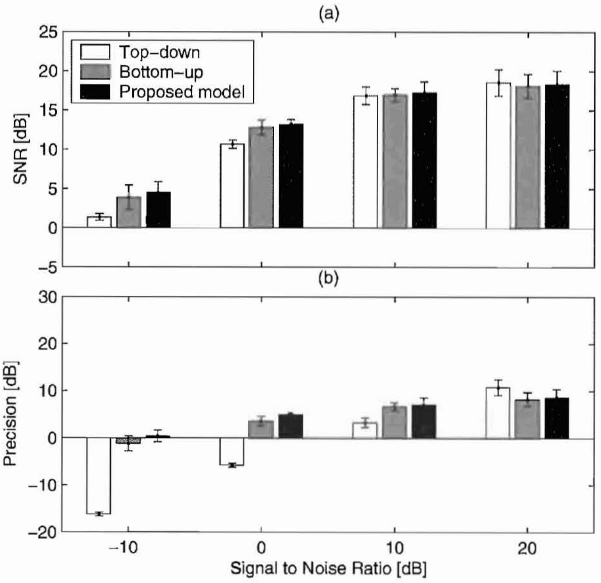

Fig. 5 Segregation accuracy when segregating a flute sound from a noisy sound: (a) SNR and (b) precision

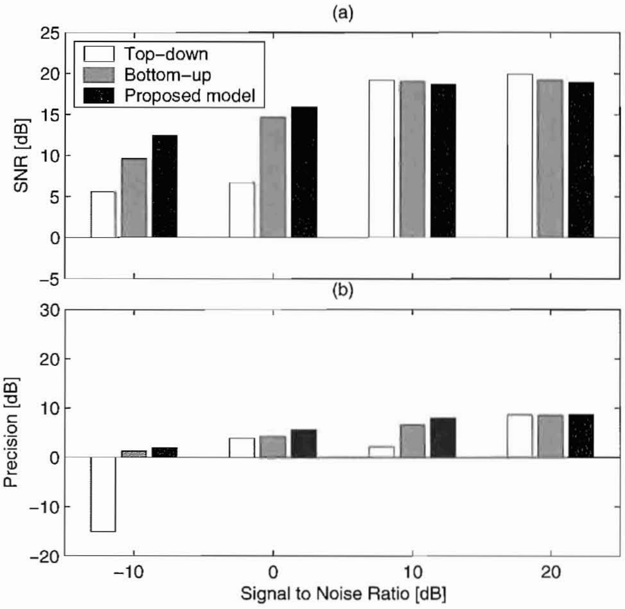

Fig. 6 Segregation accuracy when segregating a flute sound from a mixed sound: (a) SNR and (b) precision

horn (Eb2). For example, when the SNR of the mixed signal was 0 dB, it was possible to improve the SNR by about 16 dB from $f(t)$, and to improve the SNR by about 10 dB and the precision by about 2 dB in terms of segregation accuracy, compared with the top-down processing. This comparison shows the importance of separating each component from the overlapped components in each channel. These results show that the proposed model can selectively segregate the target, using the key of the target sound, with high accuracy.

The results of other second simulations for piano (G3) are shown in Fig. 7, where $f(t)$ was the target piano (G3) sound mixed with flute (A4), violin (C4), and horn (Eb2). For example, when the SNR of the mixed signal was 0 dB, it was possible to improve the SNR by about 12 dB from $f(t)$, and to improve the SNR by about 2 dB and the precision by about 5 dB with respect to segregation accuracy, compared with the top-down processing.

For the other target sounds (horn, violin), the results were similar to those shown in Figs. 6 and 7. When the SNR of the mixed signal was 0 dB, we could improve the SNR for the horn and violin sounds by about 7.3 dB, and 13.6 dB, respectively, from $f(t)$, and improve the SNR by about 3.6 dB, and 0.9 dB and the precision by about 9.3 dB, and 0.3 dB with respect to segregation accuracy compared with the top-down processing.

These comparisons again show the importance of separating each component from the overlapped components in each channel, and that the proposed model can selectively segregate the target, using the key of the target sound, with high accuracy.

4.3 Simulation 3

Next, to demonstrate that the proposed model can be applied to a realistic problem where the target performance sound must be segregated from mixed sound (which is a typical situation regarding the cocktail party effect), we carried out the following simulation. The original signals were as follows. Target $f_1(t)$ was a piano sound played "chu-rippu" (six notes: CDECDE), $f_2(t)$ was a flute sound played "kirakiraboshi" (seven notes: CCGGAAG), $f_3(t)$ was a violin sound played "choucho" (six notes: GEEFEE), and $f_4(t)$ was white noise. These were musical sounds taken from Japanese songs (except for $f_4(t)$). Inputs were the mixed signal $f(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t)$ and the keys of the symbol (piano) and notes (CDECDE, not including any time information) of the target. The task was to selectively segregate the target sound ("chu-rippu" of the piano sound) from mixture $f(t)$.

Figure 8 shows an example of the signal processing of the proposed model for this task. In this figure, panels A and B respectively show each original signal and the mixed signal $f(t)$ at an SNR of 0 dB. The instantaneous amplitudes $S_k(t)$s and phase $\phi_k(t)$s (panel C) are decomposed from $f(t)$ using the filterbank and then the candidates for the F-note (panel D) are extracted from $S_k(t)$s. The template of the target sound (panel E) is generated from the knowledge sources us-

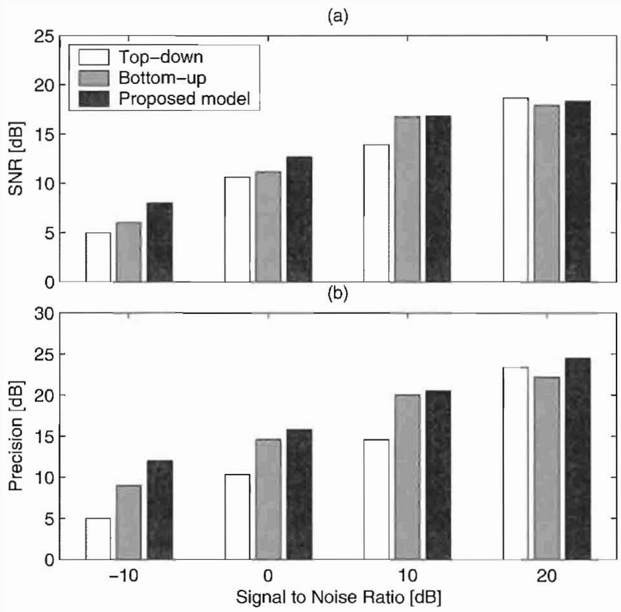

Fig. 7 Segregation accuracy when segregating a piano sound from a mixed sound: (a) SNR and (b) precision

ing keys. The segregated amplitude $A_k(t)$s (panel F) and phase $\theta_{1k}(t)$s are obtained from $S_k(t)$ and $\phi_k(t)$ using the constraints and template, and then the selectively segregated signal $\hat{f}_1(t)$ is reconstructed by the grouping block.

In this simulation, the proposed model improved the SNR by about 10.6 dB from $f(t)$. Moreover, the accuracy of the segregated target sound was improved by about 1.5 dB because of the better SNR and by about 2 dB because of the greater precision, compared with top-down processing. This suggests that this improvement reflects an advantage of the proposed model because top-down processing can precisely select the position of the target signal in the mixture and then bottom-up processing can separate the signal components of the target at the signal position from the mixture in the concurrent time-frequency region. In contrast, it was difficult to selectively segregate the target sound from the mixed sound using bottom-up processing without having some prior information because the target position could not be precisely selected. We have thus shown that our proposed model can be used to selectively segregate the sound of a target musical instrument from a mix of various sounds in a way similar to the cocktail party effect.

4.4 Consideration

The simulation results show that the proposed model based on our model-concept can selectively segregate a target instrument from a mixture of instruments. They also show that the proposed model performed best in the three simulations when using only top-down and only bottom-up processing. Although the sound separation using bottom-up processing that we previously proposed worked well in simulations 1 and 2 (signal extraction and concurrent separation), in simulation 3 it was too difficult to selectively segregate the target performance sound from the mixture of performance sounds without any useful prior information. Top-down processing worked in the three simulations, but this has an essential drawback in that it cannot segregate the target components from the mixture components in the concurrent time-frequency region. We confirmed that the proposed model can simultaneously solve the two issues that we addressed in Sec. 1 while the sound segregation model based on either bottom-up or top-down processing only cannot solve these simultaneously. This advantage was proven to be best achieved through the interaction of bottom-up and top-down processes.

5. Conclusion and Future Perspectives

In this paper, as the first step towards constructing a selective sound segregation model, we considered a simple basic problem of selective segregation for instrument sounds. We have proposed a novel model-concept of selective sound segregation that combines top-down and bottom-up processing and have implemented a model for selectively segregating instrument target sound from a sound mixture. We carried out three segregation simulations to evaluate the proposed model: (i) segregation of the target sound from a noise-added target sound (signal extraction), (ii) segregation of the target sound from a mix of four instrument sounds (concurrent separation), and (iii) segregation of a musical performance from the mixture of musical performance sounds (selective segregation).

Our results in the first two cases (signal extraction and concurrent separation) show that our model can selectively and highly accurately segregate a target sound not only from a noisy sound but also from a mix of various sounds. Our results also show that combining top-down and bottom-up processing is useful for selective sound segregation. The results of our third simulation (selective segregation) show that the proposed model can be applied to a more realistic sound segregation problem, such as the sort of situation where the cocktail party effect occurs. As the results, the proposed model was best constructed through the interaction of bottom-up and top-down processes so that the two issues in the problem can be solved and a reasonable prototype model of selective sound segregation can be achieved.

In our future work, we hope to establish a means of constructing a standard template for any instrument

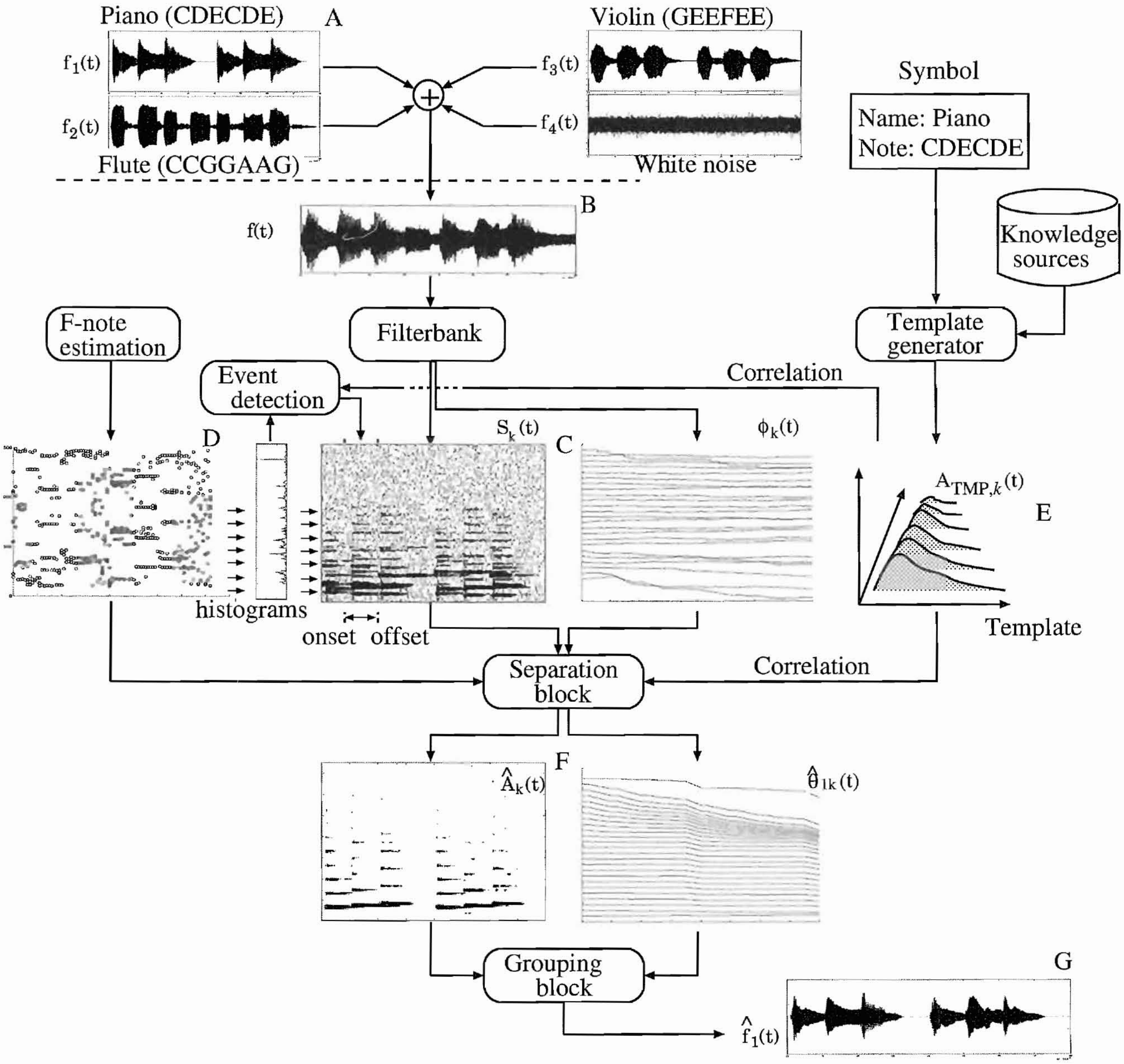

Fig. 8 Overview of signal processing for the proposed model: A: original signals (piano, flute, and violin) and background noise, B: mixed signal, C: instantaneous amplitudes $S_k(t)$ and phases $\phi_k(t)$, D: F-note candidates and their histograms, E: generated template, F: instantaneous amplitudes $A_k(t)$ and phases $\theta_{1k}(t)$, and G: segregated signal $\hat{f}_1(t)$

sound (e.g., optimization between the template and a real sound, and an HMM-based synthesis method) and a grouping rule for the interaction of top-down and bottom-up processing through other mathematical techniques. Moreover, we will adapt the model for various musical performance sounds and will also extend the model for speech segregation problems to develop this model concept as a model of the cocktail party effect, as shown in Fig. 1. We also hope to extend this model to a binaural processing model to deal with directional hearing.

We have already studied this applicability of the proposed model as a form of front-end processing for speech recognition systems in a preliminary study [29]. If successful for all perspectives, the developed general selective segregation model based on our model-concept may not only contribute to various types of signal processing for applications, but also play a role in modeling the mechanisms of a human's selective hearing system.

Acknowledgments

This work was partially supported by a Grant-in-Aid for Science Research from the Ministry of Education (No. 14780267, No. 15300061, No.17650047).

References

[1] E.G. Cherry: Some experiments on the recognition of speech with one and with two ears, J. Acoust. Soc. Am., Vol. 25, No. 5, pp. 975–979, Sept. 1953.

[2] S. Furui and M.M. Sondhi: Advances in Speech Signal Processing. Marcel Dekker, New York 1991.

[3] S.F. Boll: Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. on Acoustic, Speech and Signal Processing, Vol. ASSP-27, No. 2, pp. 113–120, April 1979.

[4] J.C. Junqua and J.P. Haton: Robustness in Automatic Speech Recognition - Fundamentals and Applications. Kluwer Academic Publishers, Boston, MA 1996.

[5] A. Papoulis: Signal Analysis. McGraw-Hill, New York, 1997.

[6] S. Shamsunder and G.B. Giannakis: Multichannel blind signal separation and recognition, IEEE Trans. on Speech and Audio Processing Vol. 5, No. 6, pp. 515–528, Nov. 1997.

[7] A. Papoulis: Probability, Random Variables, and Stochastic Process, 3rd Ed., McGraw-Hill, New York, 1991.

[8] H. Sawada, R. Mukai, S. Araki and S. Makino: Polar coordinate based nonlinear function for frequency-domain blind source separation, IEICE Trans. Fundamentals, Vol. E86-A, No. 3, pp. 590–596, April 2003.

[9] T. Nishikawa, H. Saruwatari and K. Shikano: Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA, IEICE Trans. Fundamentals, Vol. E86-A, No. 4, pp. 846–858, April 2003.

[10] A.S. Bregman: Auditory Scene Analysis: The Perceptual Organization of Sound, MIT Press, Cambridge, MA, 1990.

[11] A.S. Bregman: Auditory Scene Analysis: hearing in complex environments, in Thinking in Sounds, pp. 10–36, Oxford University Press, New York, 1993.

[12] M. Cooke and D.P.W. Ellis: The auditory organization of speech and other sources in listeners and computational models, Speech Communication, Vol. 35, No. 3, pp. 141–177, Oct. 2001.

[13] G. J. Brown: Computational auditory scene analysis: a representational approach, Ph.D. Thesis, University of Sheffield, 1992.

[14] A. de Cheveigné: Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model of auditory processing, J. Acoust Soc. Am., Vol. 93, No. 6, pp. 3271–3290, June 1993.

[15] M. P. Cooke: Modeling auditory processing and organization, Ph.D. Thesis, University of Sheffield. Cambridge University Press, Cambridge, 1993.

[16] K. Kashino and H. Tanaka: A computational model of auditory segregation of two frequency component - evaluation and integration of multiple cues, IEICE Trans. Vol. J77-A, No. 5, pp. 731–740, May 1994.

[17] M. Unoki and M. Akagi: A method of signal extraction from noise-added signal, Electronics and Communications in Japan, Part 3, Vol. 80, No. 11, pp. 1–11 (in English); translated from IEICE, Vol. J80-A, No. 3, pp. 444–453, March 1997 (in Japanese).

[18] M. Unoki and M. Akagi: Signal extraction from noisy signal based on auditory scene analysis, Speech Communication, Vol. 27, No. 3, pp. 261–279, April 1999.

[19] M. Unoki and M. Akagi: A method of extraction the harmonic tone from noisy signal based on auditory scene analysis, IEICE Trans., Vol. J82-A, No. 10, pp. 1497–1507, Oct. 1999.

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

[21] T. Nakatani and H.G. Okuno: A computational model of sound stream segregation with multi-agent paradigm, Proc. ICASSP-95, Vol. 4, pp. 2671–2674, May 1995.

[22] T. Kinoshita, S. Sakai and S. Tanaka: Musical source identification based on frequency component features, IEICE Trans., Vol. J83-D-II, No. 4, pp. 1073–1081, April 2000.

[23] K. Kashino, K. Nakadai, T. Kinoshita and H. Tanaka: Note recognition mechanisms in the OPTIMA processing architecture for music scene analysis, IEICE Trans. Vol. J79-D-II, No. 11, pp. 1751–1761, Nov. 1996.

[24] M. Goto: F0 estimation of melody and bass lines in musical audio signals, IEICE Trans. D-II Vol. J84-D-II, No. 1, pp. 12–22, Jan. 2001.

[25] K. Kashino and H. Murase: Sound source identification by adaptive template-mixture method — Application to ensemble music recognition —, IEICE Trans., Vol. J81-D-II, No. 7, pp. 1510–1517, July 1998.

[26] T. Kitahara, M. Goto and H. G. Okuno: Musical instrument identification considering pitch-dependent characteristics of timbre: a classifier based on F0-dependent

multivariate normal distribution, IPSJ Journal, Vol. 44, No. 10, pp. 2448–2458, Oct. 2003.

[27] T. Kitahara, M. Goto and H. G. Okuno: Acoustic-feature-based musical instrument hierarchy and its application to category-level recognition of unknown musical instruments, IPSJ Journal, Vol. 45, No. 3, pp. 680–689, March 2004.

[28] Tone generator MU-2000, YAMAHA.

[29] A. Haniu, M. Unoki and M. Akagi: A study on a speech recognition method based on the selective sound segregation in noisy environment, Proc. NCSP'05, pp. 403–406, March 2005.

Appendix: Bregman's regularities and constraints

As we know well, the human auditory system can easily segregate a desired signal in a noisy environment that simultaneously contains speech, noise, and reflections. Recently, this ability of the auditory system has been regarded as a function of an active scene analysis system. Auditory scene analysis (ASA) has become widely known as a result of Bregman's book [10]. Bregman claimed that to perform ASA, the human auditory system uses four psychoacoustically heuristic regularities related to an acoustic event [11], as shown in Table 1 (left column):

1. common onset and offset,
2. gradualness of change,
3. harmonicity, and
4. changes occurring in the acoustic event.

On the other hand, Unoki and Akagi have proposed that a CASA-based segregation method using these four regularities as constraints can solve a two-acoustic-source segregation problem as an ill-posed inverse problem. These constraints are listed in Table 1 (right column), and are briefly explained bellow:

(i) Common onset and offset. Suppose that T_S and T_E are the onset and offset of the fundamental component. If the signal component obtained by the kth channel is the signal component generated by the same acoustic source (that is, harmonic components), then $T_{k,\mathrm{on}}$ and offset $T_{k,\mathrm{off}}$ determined by the kth channel must coincide with T_S and T_E, respectively, as shown in Table 1(i).

(ii-a) Gradualness of change (polynomial approximation). Temporal differentiations of the instantaneous amplitude $A_k(t)$, the instantaneous phase $\theta_{1k}(t)$, and the fundamental frequency $F_0(t)$ must be represented by an Rth-oder differentiable piecewise polynomial as shown in Table 1(ii).

(ii-b) Gradualness of change (smoothness). Suppose that the instantaneous amplitude $A_k(t)$ and phase $\theta_{1k}(t)$ are defined in the closed-duration $[t_a, t_b]$ and satisfy constraint 1. If $A_k(t)$ and $\theta_{1k}(t)$ are as smooth as possible, the integrations shown in Table 1 (ii) must be minimized.

(iii) Harmonicity. $F_0(t)$ is the fundamental frequency and N_{F_0} is the number of harmonics of the highest order. The harmonic component must satisfy the multiple of $F_0(t)$ with N_{F_0} as shown in Table 1(iii).

(iv) Common AM. The normalized instantaneous amplitude of the output of the kth channel must approximate that of the ℓth channel as shown in Table 1(iv).

See [18, 19] for more details.

Masashi Unoki was born in Akita Pref., Japan, in 1969. He received the M.S. and Ph.D. degrees (Information Science) from Japan Advanced Institute of Science and Technology (JAIST), in 1996 and 1999, respectively. His main research interests are in auditory-motivated signal processing and the modeling of auditory systems. From 1998 to 2001, he was a JSPS research fellow. During 1999-2000, he was associated with the ATR Human Information Processing Laboratories as a visiting researcher, and in 2000-2001 he was a visiting research associate at CNBH, Dept. of Physiology, University of Cambridge. Since 2001 he has been on the faculty of the School of Information Science, JAIST and is now an associate professor. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). Dr. Unoki received the Sato Prize for Outstanding Paper from the ASJ in 1999 and the Yamashita Taro Prize for Young Researcher from the Yamashita Taro Research Foundation in 2005.

Masaaki Kubo was born in Ishikawa Pref., Japan, in 1978. He received the B.E. degree from the Faculty of Engineering Department of Electrical Engineering, Gunma University in 2001, and the M.S. degree from Japan Advanced Institute of Science and Technology in 2003. He has been with the Sony Corporation Home Network Company since 2003.

Atsushi Haniu was born in Saitama Pref., Japan, in 1973. He received the B.E. and M.E. degrees from the Faculty of Engineering Department of Metallurgical Engineering, Tokyo Institute of Technology in 1997 and 2001, respectively, and the M.S. degree from Japan Advanced Institute of Science and Technology in 2004. He has since been on the faculty of the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST) and is now a doctoral candidate. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Acoustical Society of Japan (ASJ). Mr. Haniu received the Student Paper Award at NCSP'05.

Masato Akagi received the B.E. degree from Nagoya Institute of Technology in 1979, and the M.E. and PhD. Eng. degrees from Tokyo Institute of Technology in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT), in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST) and is now a professor. His research interests include speech perception, the modeling of speech perception mechanisms of human beings, and signal processing of speech. During 1998, he was associated with the Research Laboratories of Electronics, MIT as a visiting researcher, and in 1993, he studied at the Institute of Phonetics Science, Univ. of Amsterdam. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of Japan (ASJ), the Institute of Electrical and Electronic Engineering (IEEE), the Acoustical Society of America (ASA), and the International Speech Communication Association (ISCA). Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, and the Sato Prize for Outstanding Paper from the ASJ in 1998 and 2005.

(Received June 15, 2006; revised August 7, 2006)