A model for selective segregation of a target instrument sound from the mixed sound of various instruments

Masashi Unoki, Masaaki Kubo, and Masato Akagi

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

Abstract

This paper proposes a selective sound segregation model for separating target musical instrument sound from the mixed sound of various musical instruments. The model consists of two blocks: a model of segregating two acoustic sources based on auditory scene analysis as bottom-up processing, and a selective processing based on knowledge sources as top-down processing. Two simulations were carried out to evaluate the proposed model. Results showed that the model could selectively segregate not only the target instrument sound, but also the target performance sound, from the mixed sound of various instruments. This model, therefore, can also be adapted to computationally model the mechanisms of a human's selective hearing system.

1. Introduction

Let us consider the problem of selective sound segregation (Fig. 1). Here, the sound of three musical performances, independently played by flute, piano, and violin, are mixed together. When we try to listen separately to the target sound (e.g., the piano sound) from among the mixed sound, we can easily selectively segregate the target sound if we know what the target is and have previously listened to it. In general, this type of situation arises from what is called the "cocktail party effect" [3] and is an important issue not only with regard to automatic music description systems, but also regarding various types of signal processing such as that of hearing aid systems and robust speech-recognition systems. In practice, though, it is difficult to construct a computational model that can process signals in this way, because the signals exist in a concurrent time-frequency region and this problem is an ill-inverse problem. Therefore, we need to use reasonable constraints to solve the problem.

Recently, sound segregation models based on "computational auditory scene analysis (CASA)" have been proposed to solve the above problem by using Bregman's regularities [2]. In particular, in the case of musical sound, CASA are called "music scene analysis" [6], 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 [4, 5, 6]. The underlying concept of these models is to computationally model the ability of the auditory system as a function of active scene analysis [2]. There are two main types of segregation models, based on either bottom-up (e.g., [7]) or top-down (e.g., [4, 6]) processes.

To realize a selective sound segregation model as shown in Fig. 1, we have to resolve two issues: (1) how to precisely select the target sound within a real environment, and (2) how to completely separate the target sound from the mixed sound in which overlapped components exist in a concurrent time-frequency region. However, since bottom-up and top-down processes focus only on either issue (1) or (2), respectively, each alone cannot be used to realize a selective sound segregation model. This paper proposes a model concept for selectively segregating a target instrument sound from a mixture of various sounds by combining top-down and bottom-up processing.

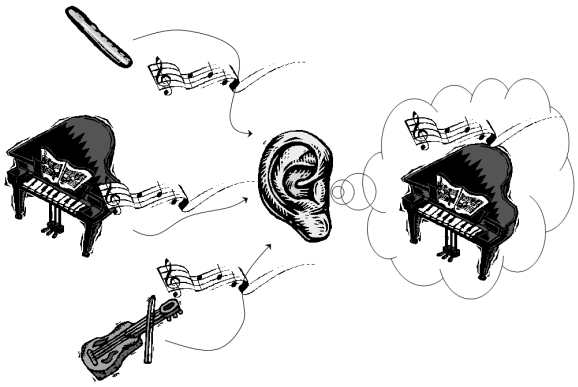

Figure 1: The selective sound segregation problem.

2. Selective sound segregation model

The proposed selective segregation model is shown in Fig. 2. This model is based on the two types of processing: top-down processing to select the position of the target sound in the mixed sound (to resolve (1), as shown by the dashed-line in Fig.2), and bottom-up processing to separate the target from the other sounds in the concurrent time-frequency region (to resolve (2), the dotted line in Fig. 2). The bottom-up processing is the same method proposed [7], but it has been modified so that it can be combined with top-down processing.

2.1. Model concept and definition

In this model, the original signals ($f_1(t)$, $f_2(t)$, $f_3(t)$, and so on) are not known, nor it is known how many different sounds there are. The only model inputs are the observed mixed signal $f(t)$ and a knowledge key such as the symbol for the target instrument name (here this is $f_1(t)$). To deal with top-down information, we assume that the exact target sound can exist

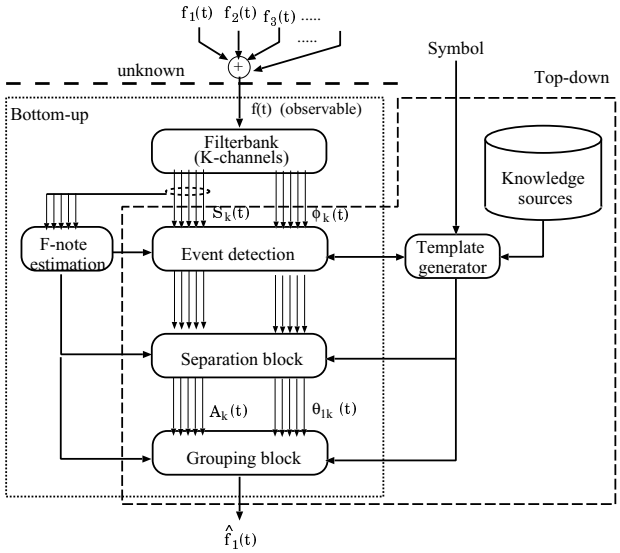

Figure 2: Selective sound segregation model.

Table 1: Constraints corresponding to Bregman's regularities.

Regularity (Bregman, 1993)	Constraint (Unoki, 1999)
(i) common onset/offset	$\lvert T_{\rm S} - T_{k,{\rm on}}\rvert \le \Delta T_{\rm S}$, $\lvert T_{\rm E} - T_{k,{\rm off}}\rvert \le \Delta T_{\rm E}$
(ii) gradualness of change (slowness)	$dA_k(t)/dt = C_{k,R}(t)$ $d\theta_{1k}(t)dt = D_{k,R}(t)$ $dF_0(t)/dt = E_{0,R}(t)$
(smoothness)	$\int_{t_a}^{t_b}[A_k^{(R+1)}(t)]^2dt \Rightarrow \min$ $\int_{t_a}^{t_b}[\theta_{1k}^{(R+1)}(t)]^2dt \Rightarrow \min$
(iii) harmonicity	$n \times F_0(t), n = 1, 2, \cdots N_{F_0}$
(iv) common AM	$\frac{A_k(t)}{\lVert A_k(t)\rVert} \approx \frac{A_\ell(t)}{\lVert A_\ell(t)\rVert}, \quad k \ne \ell$

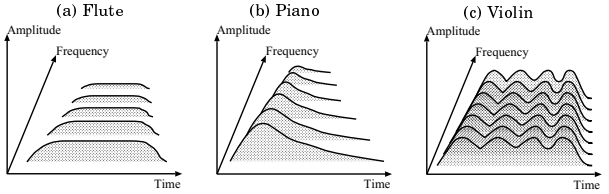

Figure 3: Typical template for the target instrument.

in anywhere in the mixed sound, 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 [7].

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

$$X_k(t) = S_k(t)\exp(j\omega_k t + j\phi_k(t)), \tag{1}$$

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 $A_k(t)\exp(j\omega_k t + j\theta_{1k}(t))$, and $B_k(t)\exp(j\omega_k t + j\theta_{2k}(t))$, then the instantaneous amplitudes $A_k(t)$ and $B_k(t)$ can be determined as

$$A_k(t) = S_k(t)\sin(\theta_{2k}(t) - \phi_k(t))/\sin\theta_k(t), \tag{2}$$
$$B_k(t) = S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))/\sin\theta_k(t), \tag{3}$$

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

However, $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ cannot be uniquely determined without some constraints. This is easily understood by considering the above equations. The problem, therefore, is the ill-inverse problem. To solve this problem, we previously proposed a basic model that uses constrains related to the four Bregman's regularities [7], as shown in Table 1.

2.2. Model implementation

The basic problem given above is for 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., $f_2(t) + f_3(t) + \cdots + f_N(t)$). The problem is then solved using the solution based on the Auditory Scene Analysis [7].

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

This filterbank decomposes the observed signal $f(t)$ into complex spectra $X_k(t)$. It is designed as a constant narrow-bandwidth filterbank with $K = 500$, a 20-Hz bandwidth, a FIR-type bandpass filter, and a 20-kHz sampling frequency. $S_k(t)$ and $\phi_k(t)$ are determined by using the Hilbert transform of $X_k(t)$ [7].

The F-note estimation block determines the candidates for the fundamental frequency of the musical instrument sound obtaining some peaks in the auto-correlation function in terms of the frequency region at each time of $S_k(t)$s. The histograms for each candidate are then calculated according to the time axes in the time-frequency region. Some of the candidates with higher histogram values are passed to the event-detection block and the final estimated F-note, $F_0(t)$, 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.

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, duration, and general acoustical feature of the musical instrument sound. The shapes of the standard template for flute, piano, and violin are shown in Fig. 3. In this paper, templates were obtained from the averaged instantaneous amplitude of the target under various conditions (normalized duration and normalized harmonicity etc.). This can be extended by analyzing all of the sounds as was done in [6].

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 is selected from the candidates of F-note $F_0(t)$ while this block searches whether the extracted amplitude based on the harmonicity of each candidate of F-note matches the template based on the correlations. This corresponds to constraint (iii). The estimated event of the target can then be obtained from a 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 harmonicity of the selected fundamental frequency. This corresponds to constraint (i).

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) is implemented such that $C_{k,R}(t)$ and $D_{k,R}(t)$ are linear ($R = 1$) polynomials, in order to reduce the computational cost of estimating $C_{k,R}(t)$ and $D_{k,R}(t)$. In 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. (2), 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

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{S_k(t)\cos(\phi_k(t) - \theta_{1k}(t)) + C_k(t)}\right), \tag{4}$$

where $C_k(t) = -\int C_{k,R}(t)dt - C_{k,0} = -A_k(t)$ [7].

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. Then, $\hat{C}_{k,1}(t)$ is determined by using

$$\hat{C}_{k,1} = \underset{\hat{C}_{k,0} - P_k \le C_{k,1} \le \hat{C}_{k,0} + P_k}{\arg\max} \frac{< \hat{A}_k, A_{\mathrm{TMP,k}} >}{||\hat{A}_k|| \cdot ||A_{\mathrm{TMP,k}}||}, \tag{5}$$

where $\hat{A}_k(t)$ is obtained through spline interpolation and $A_{\mathrm{TMP,k}}(t)$ is template such as one shown in Fig. 3. Finally, $\hat{D}_{k,1}(t)$ is determined by using

$$\hat{D}_{k,1} = \underset{\hat{D}_{k,0} - Q_k \le D_{k,1} \le \hat{D}_{k,0} + Q_k}{\arg\max} \frac{< \hat{A}_k, A_{\mathrm{TMP,k}} >}{||\hat{A}_k|| \cdot ||A_{\mathrm{TMP,k}}||}. \tag{6}$$

The difference between our proposed model and the previous model is that we use a template of $A_{\mathrm{TMP,k}}(t)$ instead of the averaged $\hat{A}_k(t)$ [7]. 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. (2), Eq. (3), and $\theta_{2k}(t) = \theta_k(t) + \theta_{1k}(t)$, respectively.

The grouping block merges the instantaneous amplitudes $A_k(t)$s and phases $\theta_{1k}(t)$ in the concurrent time-frequency region of the target using constraints (i) and (iii) from Table 1, and then reconstructs them into the segregated signal $\hat{f}_1(t)$.

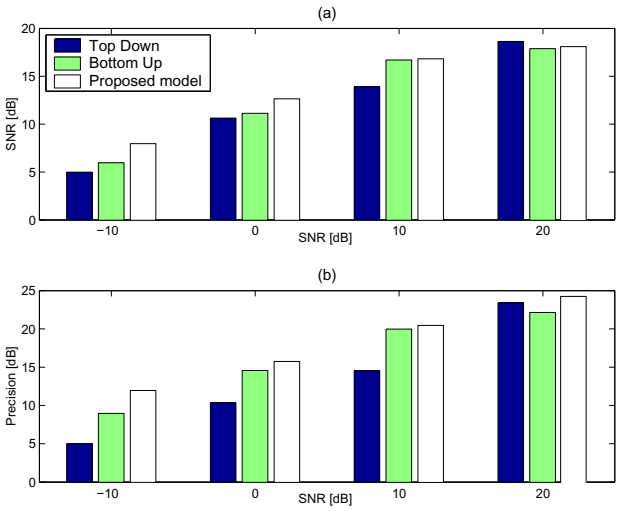

Figure 4: Segregation accuracy when segregating a piano sound from a mixed sound: (a) SNR and (b) precision.

3. Simulations

First, 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 simulations of the segregation of each sound from four mixed sounds (piano, flute, horn, and violin). Five types of mixed signal $f(t)$ were used as simulation stimuli in each simulation, where the SNRs 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).

To evaluate the segregation performance of our proposed method, we used the following two measures. Both measures show improvement if they become positive higher values.

$$\mathrm{SNR} = 10\log_{10}\frac{\int_0^T f_1(t)^2 dt}{\int_0^T (f_1(t) - \hat{f}_1(t))^2 dt} \tag{7}$$

$$\mathrm{Precision} = \frac{1}{T}\int_0^T \left(10\log_{10}\frac{\sum_{k=1}^K \tilde{A}_k(t)^2 dt}{\sum_{k=1}^K (\tilde{A}_k(t) - A_k(t))^2}\right) dt. \tag{8}$$

Moreover, to show the advantages of the proposed model, we compared the performance of the model when (a) using only top-down processing (only extracting the harmonic component of the target sound, not segregating it in each channel) and (b) using bottom-up processing (i.e., using a previous model [7]).

The results of the first simulations for piano (G3) are shown in Fig. 4, 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 as segregation accuracy, compared with the top-down processing. This comparison shows the importance of separating each component from the overlapped components in each channel. Our results show that the proposed model can selectively segregate the target, using the key of the target sound, with high accuracy. For the other target sounds (flute, horn, violin), the results were similar to those shown in Fig. 4. When the SNR of the mixed signal was 0 dB, we could improve the SNR for the flute,

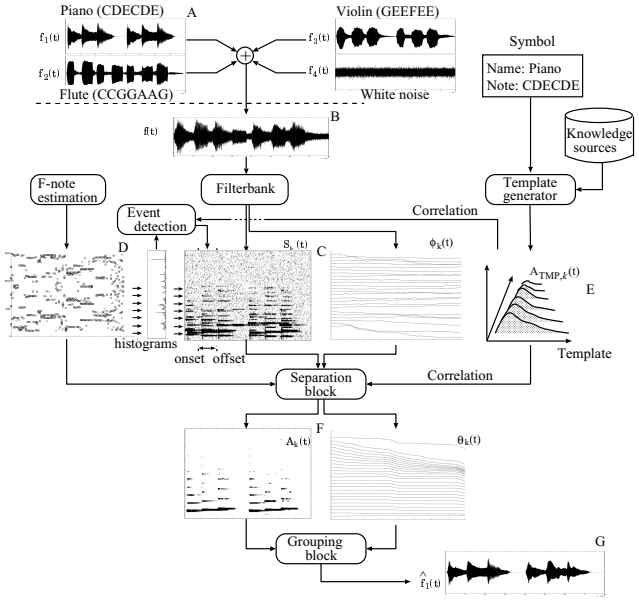

Figure 5: Overview of signal processing for the proposed model.

horn, and violin sounds, by about 16.7 dB, 7.3 dB, and 13.6 dB , respectively, from $f(t)$, and improve the SNR by about 2.0 dB, 3.6 dB, and 0.9 dB and the precision by about 9.9 dB, 9.3 dB, and 0.3 dB as segregation accuracy compared with the top-down processing.

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 as shown in Fig. 1 (which is a typical situation resulting from 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 $f(t)$.

Figure 5 shows an example of the signal processing of the proposed model for this task. In this figure, panels A and B show each original signal and the mixed signal $f(t)$ at an SNR of 0 dB, respectively. 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 of 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 using 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 selective-segregated signal $\hat{f}_1(t)$ is reconstructed by the grouping block.

In this simulation, the proposed model improved the SNR 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. In this simulation, it was difficult to selectively segregate the target sound from the mixed sound using bottom-up processing without having some prior information. We have thus shown that our proposed model can be used to selectively segregate the sound of a target musical instrument performance sound from a mix of various sounds such as one resulting from the cocktail party effect.

4. Conclusions

In this paper, we have proposed a selective sound segregation model that combines top-down and bottom-up processing. We carried out two segregation simulations to evaluate the proposed model - one in which a target sound was segregated from a mix of four instrument sounds, and one in which a musical performance sound was segregated from a mixed musical performance. Our results in the first case showed that our model can selectively and highly accurately segregate a target instrument sound from a mix of various sounds. Our results also showed that combining top-down and bottom-up processing is useful for selective sound segregation. The results of our second simulation showed that the proposed model can be applied to a more realistic sound segregation problem, such as the sort of situation that results from the cocktail party effect. The advantages of our proposed model make it applicable to preprocessing for a musical scene analysis system and for replaying of a target sound. This model, therefore, can also be adapted to computationally model the mechanisms of a human's selective hearing system.

In our future work, we hope to establish a means of constructing a standard template for any instrument sound (e.g., optimization between the template and a real sound), and then we will adapt the model for various musical performance sounds.

5. Acknowledgment

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

6. References

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

[2] Cooke, M. and Ellis, D.P.W., "The auditory organization of speech and other sources in listeners and computational models," Speech Communication, vol. 35, no. 3, pp. 141-177, Oct. 2001.

[3] Cherry, E.G., "Some experiments on the recognition of speech with one and with two ears," J. Acoust. Soc. Am., 25, pp. 975-979, 1953.

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

[5] Goto, M., "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.

[6] Kinoshita, T., Sakai, S., and Tanaka, S., "Musical source identification based on frequency component features," IEICE Trans. vol. J83-D-II, no. 4, pp. 1073-1081, April 2000.

[7] Unoki, M. and Akagi, M., "Signal Extraction from Noisy Signal based on Auditory Scene Analysis," Speech Communication, vol. 27, no. 3 , pp. 261–279, April 1999.