Vowel segregation in background noise using the model of segregating two acoustic sources

Masashi Unoki[*†]
[*] ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun,
Kyoto 619-0288, JAPAN
unoki@hip.atr.co.jp

Masato Akagi[†]
[†] School of Information Science, JAIST
1-1 Asahidai, Tatsunokuchi, Nomi-gun,
Ishikawa, 923-1292, JAPAN
akagi@jaist.ac.jp

Abstract

This paper proposes an improved sound segregation model based on auditory scene analysis in order to overcome three disadvantages in our previously proposed model. The improved model solves the problem of segregating two acoustic sources by using constraints related to the heuristic regularities proposed by Bregman. In the improvements, we (1) reconsider the estimation of unknown parameters using Kalman filtering, (2) incorporate a constraint of channel envelopes with periodicity of the fundamental frequency into the grouping block, and (3) consider a constraint of smoothness of instantaneous amplitudes on channels. Simulations are performed to segregate a real vowel from a noisy vowel and to compare the results of using all or only some constraints. The proposed model can improve our previous model and precisely segregate real speech even in waveforms using all of the constraints related to Bregman's four regularities.

1 Introduction

The problem of segregating the desired signal from a noisy signal is an important issue not only in robust speech recognition systems but also in various types of signal processing. This problem has been investigated by many researchers, and many methods have been proposed. For example, an investigation of robust speech recognition [Furui and Sondhi, 1991], includes noise reduction or suppression [Boll, 1979] and speech enhancement methods [Junqua and Haton, 1996]. An investigation of signal processing includes signal estimation using a linear system [Papoulis, 1977] and signal estimation based on a stochastic process for signals and noise [Papoulis, 1991]. One recent proposal is Blind Separation [Shamsunder and Giannakis, 1997], which estimates the inverse-translation-operator (input-output translation function) by using the observed signal to estimate the original input.

However, in practice, it is difficult to segregate each original signal from a mixed signal, because this problem is an ill-posed inverse problem and the signals exist in a concurrent time-frequency region. Furthermore this problem is difficult to solve without using constraints on acoustic sources and the real environment.

On the other hand, the human auditory system can easily segregate the 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 called "Auditory Scene Analysis (ASA)". ASA has become widely known as a result of Bregman's book [Bregman, 1990]. Bregman has reported that the human auditory system uses four psychoacoustically heuristic regularities related to acoustic events to solve the problem of Auditory Scene Analysis. These regularities are (i) common onset and offset, (ii) gradualness of change, (iii) harmonicity, and (iv) changes occurring in the acoustic event [Bregman, 1993]. If an auditory sound segregation model were constructed using constraints related to these heuristic regularities, it should be possible to uniquely solve the sound segregation problem (ill-posed inverse problem). In addition, this model should be applicable not only to a preprocessor for robust speech recognition systems but also to various types of signal processing.

Some ASA-based investigations have shown that it is possible to solve the segregation problem by applying constraints to sounds and the environment. These approaches are called "Computational Auditory Scene Analysis (CASA)". Some CASA-based sound segregation models already exist. There are two main types of models, based on either bottom-up or top-down processes. Typical bottom-up models include an auditory sound segregation model based on acoustic events [Cooke, 1993, Brown, 1992], a concurrent harmonic sounds segregation model based on the fundamental frequency [de Cheveigné, 1993], and a sound source separation system with an automatic tone modeling ability [Kashino and Tanaka, 1993]. Typical top-down models include a segregation model based on psychoacoustic grouping rules [Ellis, 1996] and a computational model of sound segregation agents [Nakatani *et al.*, 1995a, Nakatani *et al.*, 1995b]. All of these models use some of the four regularities and the amplitude (or power) spectrum as the acoustic feature. Thus, they cannot completely extract the desired signal from a noisy signal when the signal and noise exist in the same frequency region.

In contrast, we have been tackling the problem of seg-

regating two acoustic sources as a fundamental problem. We believe that this problem can be uniquely solved by using amplitude, phase information, and mathematical constraints related to the four psychoacoustically heuristic regularities [Unoki and Akagi, 1997, Unoki and Akagi, 1999a].

This fundamental problem is defined as follows [Unoki and Akagi, 1997, Unoki and Akagi, 1999a]. First, only the mixed signal $f(t)$, where $f(t) = f_1(t) + f_2(t)$, can be observed. Next, $f(t)$ is decomposed into its frequency components by a filterbank (the number of channels is K). The output of the k-th channel $X_k(t)$ is represented by

$$X_k(t) = S_k(t)\exp(j\omega_k t + j\phi_k(t)). \quad (1)$$

Here, if the outputs of the k-th channel $X_{1,k}(t)$ and $X_{2,k}(t)$, which correspond to $f_1(t)$ and $f_2(t)$, are assumed to be

$$X_{1,k}(t) = A_k(t)\exp(j\omega_k t + j\theta_{1k}(t)), \quad (2)$$
$$X_{2,k}(t) = B_k(t)\exp(j\omega_k t + j\theta_{2k}(t)), \quad (3)$$

then the instantaneous amplitudes of the two signals $A_k(t)$ and $B_k(t)$ can be determined by

$$A_k(t) = \frac{S_k(t)\sin(\theta_{2k}(t) - \phi_k(t))}{\sin\theta_k(t)}, \quad (4)$$

$$B_k(t) = \frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{\sin\theta_k(t)}, \quad (5)$$

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. Instantaneous phases $\theta_{1k}(t)$ and $\theta_{2k}(t)$ can be determined by

$$\theta_{1k}(t) = -\arctan\left(\frac{Y_k(t)\cos\phi_k(t) - \sin\phi_k(t)}{Y_k(t)\sin\phi_k(t) + \cos\phi_k(t)}\right) + \arcsin\left(\frac{A_k(t)Y_k(t)}{S_k(t)\sqrt{Y_k(t)^2+1}}\right), \quad (6)$$

$$\theta_{2k}(t) = -\arctan\left(\frac{Y_k(t)\cos\phi_k(t) + \sin\phi_k(t)}{Y_k(t)\sin\phi_k(t) - \cos\phi_k(t)}\right) + \arcsin\left(-\frac{B_k(t)Y_k(t)}{S_k(t)\sqrt{Y_k(t)^2+1}}\right), \quad (7)$$

where

$$Y_k(t) = \sqrt{(2A_k(t)B_k(t))^2 - Z_k(t)^2}/Z_k(t), \quad (8)$$
$$Z_k(t) = S_k(t)^2 - A_k(t)^2 - B_k(t)^2. \quad (9)$$

Hence, $f_1(t)$ and $f_2(t)$ can be reconstructed by using the determined pair of [$A_k(t)$ and $\theta_{1k}(t)$] and the determined pair of [$B_k(t)$ and $\theta_{2k}(t)$] for all channels. However, $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ cannot be uniquely determined without some constraints, as is easily understood from the above equations. Therefore, this problem is an ill-inverse problem.

To overcome this problem, we have tried to construct a basic solution using constraints related to the four regularities [Unoki and Akagi, 1997, Unoki and Akagi, 1999a]. Thus, we have proposed a sound segregation model based on auditory scene analysis [Unoki and Akagi, 1999b]. This model solves the problem of segregating two acoustic sources by using constraints on the continuity of instantaneous phases as well as constraints on the continuity of instantaneous amplitudes and fundamental frequencies. In simulations, we showed that all constraints related to the four regularities are useful in segregating an AM-FM harmonic complex tone from a noisy AM-FM harmonic complex tone. We also showed that the proposed model can precisely segregate a real vowel from a noisy vowel even in waveforms.

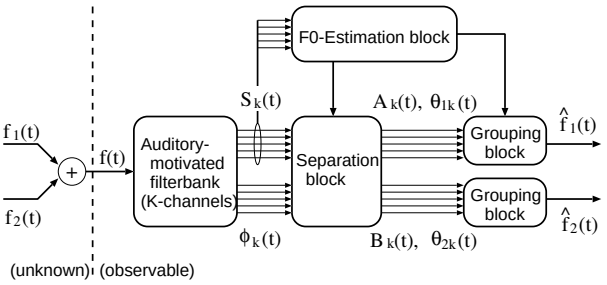

Figure 1: Auditory sound segregation model.

However, this model has the following disadvantages:

(1) the segregation accuracy improvement differs depending on the two types of noise (white or pink noise),

(2) it often fails to extract the components of envelopes with periodicity at the fundamental frequency by using the constraint of common onset / offset, and

(3) the segregated vowel feels as if it does not have good hearing quality.

This paper proposes an improved sound segregation model based on auditory scene analysis to overcome the above disadvantages.

2 Auditory sound segregation model

In this paper, the desired signal $f_1(t)$ is assumed to be a harmonic complex tone, where $F_0(t)$ is the fundamental frequency. The proposed model segregates the desired signal from the mixed signal by constraining the temporal differentiation of $A_k(t)$, $\theta_{1k}(t)$, and $F_0(t)$.

The proposed model is composed of four blocks: an auditory-motivated filterbank, an F_0 estimation block, a separation block, and a grouping block, as shown in Fig. 1. Constraints used in this model are shown in Table 1.

2.1 Auditory-motivated filterbank

The auditory-motivated filterbank decomposes the observed signal $f(t)$ into complex spectra $X_k(t)$. This filterbank is implemented as a constant Q gammatone filterbank that is constructed with $K = 128$, a bandwidth of 60–6000 Hz, and a sampling frequency of 20 kHz [Unoki and Akagi, 1997]. $S_k(t)$ and $\phi_k(t)$ are determined by using the amplitude and phase spectra defined by the wavelet transform [Unoki and Akagi, 1997].

2.2 F_0 estimation block

The F_0 estimation block determines the fundamental frequency of $f_1(t)$. This block is implemented as the Comb filtering on an amplitude spectrogram $S_k(t)$s [Unoki and Akagi, 1999b]. Since the number of channels in $X_k(t)$ is finite, the estimated $F_0(t)$ takes a discrete value. In addition, the fluctuation of $F_0(t)$ has a staircase shape and the temporal differentiation of $F_0(t)$ is zero at any segment. Therefore, this paper assumes that $E_{0,R}(t) = 0$ in constraint (ii) of Table 1 for a segment. Let the length of the above segment be $T_h - T_{h-1}$, where T_h is the continuous point of $F_0(t)$.

Table 1: Constraints corresponding to Bregman's psychoacoustical heuristic regularities.

Regularity (Bregman, 1993)	Constraint (Unoki and Akagi, 1999)	
(i) Unrelated sounds seldom start or stop at exactly the same time (**common onset/offset**)	Synchronism of 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		
(a) A single sound tends to smoothly and slowly change its properties	(a) **Slowness** (piecewise-differentiable polynomial approximation)	$dA_k(t)/dt = C_{k,R}(t)$ $d\theta_{1k}(t)/dt = D_{k,R}(t)$ $dF_0(t)/dt = E_{0,R}(t)$
(b) A sequence of sounds from the same source tends to slowly change its properties	(b) **Smoothness** (Spline interpolation)	$\sigma_A = \int_{t_a}^{t_b}[A_k^{(R+1)}(t)]^2 dt \Rightarrow \min$ $\sigma_\theta = \int_{t_a}^{t_b}[\theta_{1k}^{(R+1)}(t)]^2 dt \Rightarrow \min$ $\sigma_{A_k} = \sum_k[(\log A_k(t))^{(R+1)}]^2 \Rightarrow \min$ **(new)**
(iii) When a body vibrates with a repetitive period, these vibrations give rise to an acoustic pattern in which the frequency components are multiples of a common fundamental (**harmonicity**)	Multiples of the repetitive fundamental frequency	$n \times F_0(t), \quad n = 1, 2, \cdots, N_{F_0}$ $\ell = \frac{K}{2} - \left\lceil \frac{\log(n \cdot F_0(t)/f_0)}{\log \alpha} \right\rceil$
(iv) Many changes that take place in an acoustic event will affect all components of the resulting sound in the same way and at the same time	(a) **Slow modulation** Correlation between the instantaneous amplitudes	$\frac{A_k(t)}{\lVert A_k(t)\rVert} \approx \frac{A_\ell(t)}{\lVert A_\ell(t)\rVert}, \quad k \neq \ell$
	(b) **Fast modulation** Channel envelopes with periodicity at the $F_0(t)$	$\lvert F_0(t) - \hat{F}_0(t)\rvert \le \Delta F_0$ **(new)**

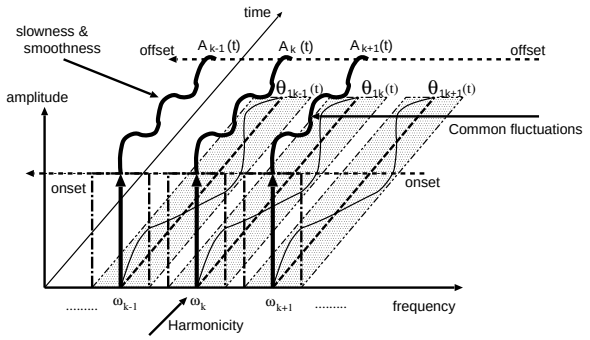

Figure 2: Concept of constraints related to Bregman's regularities.

2.3 Grouping block

The grouping block determines the concurrent time-frequency region of the desired signal using constraints (i) and (iii) in Table 1, and then reconstructs the segregated instantaneous amplitude and phase using the inverse wavelet transform [Unoki and Akagi, 1999a]. $\hat{f}_1(t)$ and $\hat{f}_2(t)$ are the reconstructed $f_1(t)$ and $f_2(t)$.

Constraint (i) is implemented by comparing the onset/offset $(T_{k,\rm on}, T_{k,\rm off})$ of $X_k(t)$ with the onset/offset $(T_{\rm S}, T_{\rm E})$ of $X_{\hat{\ell}}(t)$ corresponding to $F_0(t)$, where $\Delta T_{\rm S} =$ 25 ms and $\Delta T_{\rm E} =$ 50 ms. Onset $T_{k,\rm on}$ and offset $T_{k,\rm off}$ in $X_k(t)$ are determined by using the nearest maximum or minimum point of $|d\phi_k(t)/dt|$ and $dS_k(t)/dt$.

Constraint (iii) is implemented by determining the channel number corresponding to the integer multiples of $F_0(t)$. The channel number ℓ of $X_\ell(t)$, in which the harmonic components exist in the output of the ℓ-th channel, is determined by using (iii) in Table 1. $\lceil\cdot\rceil$ is the ceil symbol, meaning the approximation of the closest integer value toward positive infinity. In addition, K is an even number and f_0 is the center frequency of the analyzing wavelet in the constant Q gammatone filterbank.

2.4 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, as shown in Fig. 2 [Unoki and Akagi, 1999b].

Constraint (ii) is implemented such that $C_{k,R}(t)$ and $D_{k,R}(t)$ are linear $(R = 1)$ piecewise-differentiable 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, constrain the second-order polynomials ($A_k(t) = \int C_{k,1}(t)dt + C'_{k,0}$ and $\theta_{1k}(t) = \int D_{k,1}(t) + D'_{k,0}$).

In the segment $T_h - T_{h-1}$ that can be determined by $E_{0,R}(t) = 0$, the terms $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ are determined by the following steps. First, the estimation 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 (See Appendix A). Next, the candidates of $C_{k,1}(t)$ at any $D_{k,1}(t)$ are selected by using the spline interpolation in the estimated error region [Unoki and Akagi, 1999a]. Then, $\hat{C}_{k,1}(t)$ is determined by using

$$\hat{C}_{k,1} = \underset{\hat{C}_{k,0} - P_k \le C_{k,1} \le \hat{C}_{k,0} + P_k}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||}, \qquad (10)$$

where $\hat{A}_k(t)$ is obtained by the spline interpolation and $\hat{\hat{A}}_k(t)$ is determined in the across-channel that satisfies constraint (iii). Finally, $\hat{D}_{k,1}(t)$ is determined by using

$$\hat{D}_{k,1} = \underset{\hat{D}_{k,0} - Q_k \le D_{k,1} \le \hat{D}_{k,0} + Q_k}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||}. \qquad (11)$$

Since, $\theta_k(t)$ is determined by

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{S_k(t)\cos(\phi_k(t) - \theta_{1k}(t)) + C_k(t)}\right), \qquad (12)$$

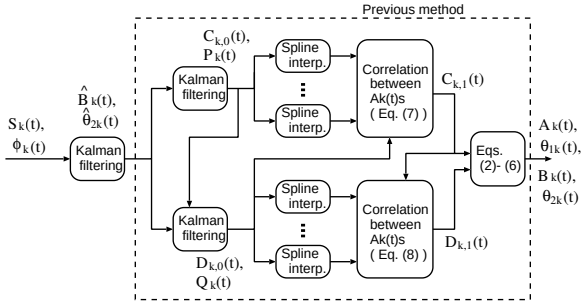

Figure 3: Signal processing of a separation block.

where $C_k(t) = -\int C_{k,R}(t)dt - C_{k,0} = -A_k(t)$ [Unoki and Akagi, 1999a], and $\theta_{1k}(t)$ is determined from $\hat{D}_{k,1}(t)$, the terms $A_k(t)$, $B_k(t)$, and $\theta_{2k}(t)$ can be determined from Eq. (4), Eq. (5), and $\theta_{2k}(t) = \theta_k(t) + \theta_{1k}(t)$, respectively.

3 Improvements to previous model

To overcome the three disadvantages in the previous model, we improved the following three respects.

3.1 Estimation of $C_{k,0}(t)$ and $D_{k,0}(t)$

We previously set the statistical parameters (mean and variance) for $A_k(t)$ and $\theta_{1k}(t)$ with ad-hoc values without considering the distribution of the power of noise. As a result, the estimation of $C_{k,0}(t)$ and $D_{k,0}(t)$ was influenced by the power of the noise components that passed through the channels.

Since we need to know the statistical parameters of $A_k(t)$ and $\theta_{1k}(t)$, we improved the estimation of $C_{k,0}(t)$ and $D_{k,0}(t)$ by using Kalman filtering as follows (See Appendix A about details):

1. estimate $B_k(t)$ and $\theta_{2k}(t)$ with Kalman filtering,
2. estimate $A_k(t)$ and $\theta_{1k}(t)$ from Eqs. (4) and (6),
3. calculate mean and deviation of $A_k(t)$ and $\theta_{1k}(t)$, and
4. estimate $C_{k,0}(t)$ and $D_{k,0}(t)$ with the previous method [Unoki and Akagi, 1999b].

3.2 Constraint of envelopes with periodicity

Fig. 4 shows an original signal /a/ and its instantaneous amplitudes, $A_k(t)$s. Channel envelopes with periodicity at the fundamental frequency $F_0(t)$ exist at higher frequencies. This is caused by using constant Q filterbank.

In our previous model, the constraint of common onset and offset often failed to extract channel envelopes with the original added white noise, but the constraint could extract channel envelopes with the original added pink noise. This is because the power of noise components at a higher frequency disturbed the detection of the onset and offset of the desired signal using constraint (i).

We reconsider constraint (iv) to solve this problem. We divide constraint (iv) into two temporal modulations: slow modulation and fast modulation. Then, we regard slow temporal modulation as common fluctuations of $A_k(t)$ and regard the fast temporal modulation as channel envelopes with periodicity at the fundamental frequency.

In order to detect channel envelopes with periodicity at $F_0(t)$, we implement a detection of the difference

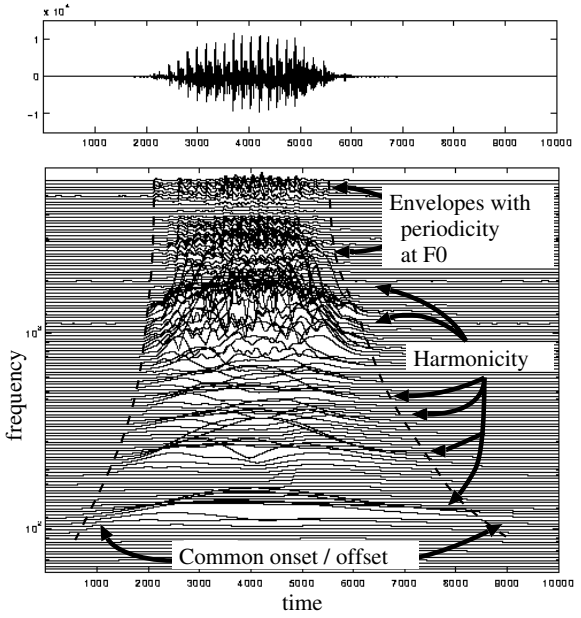

Figure 4: Original signal /a/ (top) and its instantaneous amplitudes, $A_k(t)$s (bottom).

between $F_0(t)$ and $\hat{F}_0(t)$ as the fast modulation, where $F_0(t)$ is determined by the F_0 estimation block and $\hat{F}_0(t)$ is estimated by using the autocorrelation between $S_k(t)$ for any t as follows.

$$A_{\mathrm{corr,k}}(t,\tau) = \frac{1}{T}\int_t^{t+T} S_k(x)S_k(x+\tau)dx, \quad (13)$$

$$\hat{T}_0(t) = \underset{\tau_{\min}\le\tau\le\tau_{\max}}{\arg\max}\ A_{\mathrm{corr,k}}(t,\tau), \quad (14)$$

$$\hat{F}_0(t) = 1/\hat{T}_0(t), \quad (15)$$

where $\tau_{\min} = 1/400$, $\tau_{\max} = 1/60$, $T = 1/60$, and τ is lag length. The above common value, 60, means the lowest frequency of the filterbank. In this paper, ΔF_0 of constraint (iv-b) in Table 1 is 10 Hz.

3.3 Smoothness of $A_k(t)$ on channels

In our previous model, we set $A_k(t)$s on the non-desired signal region, which is not constraint by (i) and (iii), have been to zero. However, zero-setting might cause a decrease in the quality of the segregated signal form our experience and some reports [Boll, 1979].

To consider the above disadvantage, we consider the constraint of smoothness on the frequency axis instead of zero-setting. This constraint is shown in Table 1 (ii-b). For any t, we set values, which are determined by using spline interpolation, into $A_k(t)$s when these are not satisfied in grouping constraints.

3.4 Overview of the proposed model

Fig. 5 shows an overview of signal processing of the proposed model. First, the noisy vowel /a/ $f(t)$ shown in Fig. 5 A (the SNR of $f(t)$ is 10 dB) is decomposed into $S_k(t)$ and $\phi_k(t)$ as shown in Figs. 5 B and C, respectively. Next, $F_0(t)$ is estimated as shown in Fig. 5 D. The concurrent time-frequency region of the desired signal $f_1(t)$ is determined using constraints (i), (iii), and (iv-b) as shown in Figs. 5 E, F, and G. Finally, the instantaneous amplitudes and the instantaneous phases

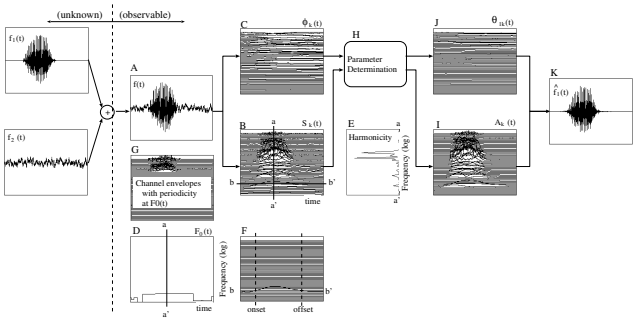

Figure 5: Overview of signal-flow in proposed model.

of the two signals are determined from $S_k(t)$ and $\phi_k(t)$ using constraints (ii) and (iv-a). The determined $A_k(t)$ and $\theta_{1k}(t)$ are shown in Figs. 5 I and J, respectively. The segregated signal $\hat{f}_1(t)$ is shown in Fig. 5 K. In this figure, the segregated $B_k(t)$, $\theta_{2k}(t)$, and $\hat{f}_2(t)$ are omitted.

4 Simulations

To show that the proposed method can segregate the desired vowel from noisy vowel even in waveforms, we performed the following three simulations:

1. vowel segregation (/a/, /i/, /u/, /e/, /o/) from a noisy vowel: the dataset size was 160 (five vowels, four speakers, four noise signals, and two types of noise);
2. vowel segregation (/aoi/) from a noisy vowel: the dataset size was 32 (one vowel, four speakers, four noise signals, and two types of noise); and
3. vowel segregation from another vowel (double vowel condition): one vowel was (/a/, /i/, /u/, /e/, /o/) from the male (mau) or female (fkn) speaker and the other was /aoi/ from the female (fsu) or male (mht) speaker, and the dataset size was 40 (five vowels, two speakers, and four noise signals).

The speech signals were the Japanese vowels of four speakers (two males and two females) in the ATR-database [ATR Tech. Rep., 1988]. The noise was pink or white noise and the SNRs of noisy signals ranged from 5 to 20 dB in 5-dB steps.

4.1 Evaluation measures

We used two measures to evaluate the segregation performance of the proposed method.

One was the ratio of the original $f_1(t)$ (signal) and the difference between the original and the segregated signal $\hat{f}_1(t)$ (noise), as defined by

$$10\log_{10}\frac{\int_0^T f_1(t)^2 dt}{\int_0^T (f_1(t)-\hat{f}_1(t))^2 dt}. \qquad \text{(dB)} \qquad (16)$$

The aim of using this measure was to evaluate whether a segregation model can segregate a desired signal from a noisy signal precisely even in waveforms. This measure is called "segregation accuracy."

The other measure was an objective distortion estimator for hearing aids such as the spectrum distortion reconsidered with the simultaneous and temporal masking effects. This measure is defined by

$$\sqrt{\frac{1}{W}\sum_{\omega}^{W}\left(20\log_{10}\frac{\tilde{X}_1(\omega)}{\tilde{\hat{X}}_1(\omega)}\right)^2}, \qquad \text{(dB)} \qquad (17)$$

where $\tilde{X}_1(\omega)$ and $\tilde{\hat{X}}_1(\omega)$ are the amplitude spectra of $f_1(t)$ and $\hat{f}_1(t)$, respectively. This is called "auditory-oriented spectral distortion (ASD)" [Mizumachi and Akagi, 1999]. In the above equation, the frame length is 21.3 ms, the frame shift is a quarter of the frame length, W is the analyzable bandwidth of the filterbank (about 6 kHz), the sampling frequency is 48 kHz, and the window function is Hamming [Mizumachi and Akagi, 1999]. Since the sampling frequency of our model is 20 kHz, we have to do an up-sampling of 20 kHz to 48 kHz to use the ASD.

4.2 Comparison with the other model

In addition, we compared the proposed model's performance with the performances of other typical methods for the above simulations. The other methods correspond to:

(1) Previous model [Unoki and Akagi, 1999b],

(2) Segregation model using constraints (ii-a) and (iii) (labeled by Cond. 1),
   - extracting the harmonics using the Comb filter
   - determining $A_k(t)$ and $\theta_{1k}(t)$ from $C_{k,0}(t)$ and $D_{k,0}(t)$

(3) Segregation model using constraints (iii) (labeled by Cond. 2),
   - extracting the harmonics using the Comb filter
   - $A_k(t) = S_k(t)$ and $\theta_{1k}(t) = \phi_k(t)$

(4) Spectral subtraction [Boll, 1979] on the gammatone filterbank (labeled SS), and
   - bias (mean of noise) subtraction
   - half-wave rectification
   - reduction of noise residual

(5) Segregation model using no constraints (labeled by Cond. 3).
   - all-pass filtering
   - $A_k(t) = S_k(t)$ and $\theta_{1k}(t) = \phi_k(t)$ for all k

We compared with the above conditions as follows:

(1) can the proposed method improve our previous model,

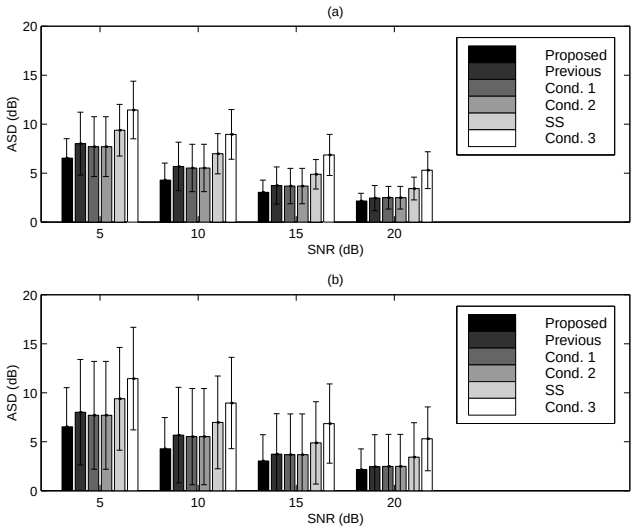

Figure 6: ASD for simulation 1: (a) bandpassed pink noise, (b) bandpassed white noise.

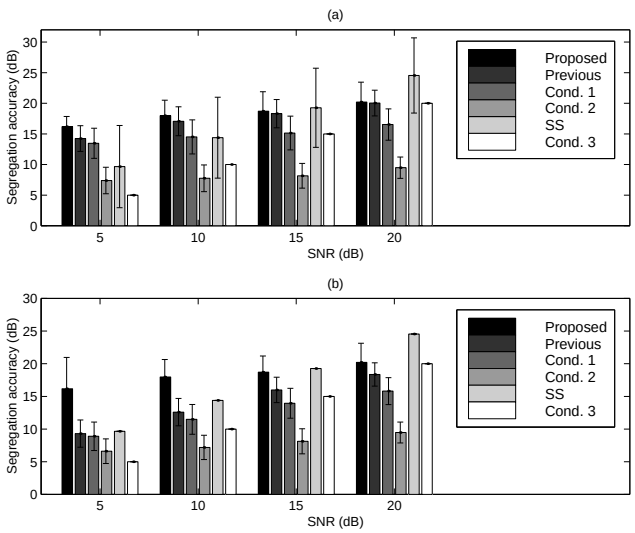

Figure 7: Segregation accuracy for simulation 1: (a) bandpassed pink noise, (b) bandpassed white noise.

(2) does it have the advantage of the smoothness of the instantaneous amplitude and phase,

(3) does it have the advantage of the instantaneous phase,

(4) is it more useful than basic noise suppression, and

(5) is the proposed method's segregation accuracy improved as noise separation.

4.3 Results and considerations

Figs. 6, 8 and 10 show the auditory-oriented spectral distortion (ASDs) in the three simulations. In these figures, the bar height shows the mean of the ASD and the error bar shows its standard deviation. Figs. 7, 9 and 11 show the segregation accuracies in the three simulations. In these figures, the bar height shows the mean of the segregation accuracy and the error bar shows its standard deviation.

The results show that the proposed method is better than our previous model, and that it obtained better segregation accuracy than the other five methods. The

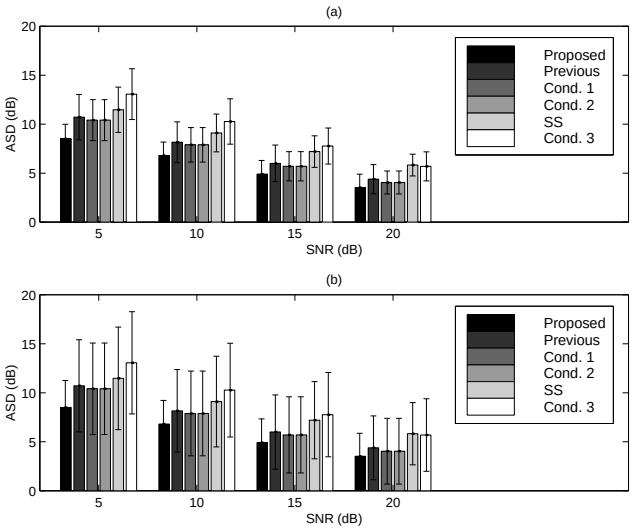

Figure 8: ASD for simulation 2: (a) bandpassed pink noise, (b) bandpassed white noise.

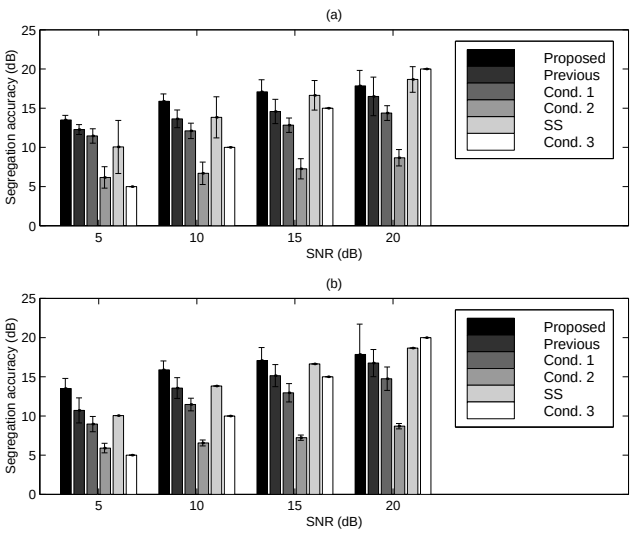

Figure 9: Segregation accuracy for simulation 2: (a) bandpassed pink noise, (b) bandpassed white noise.

proposed model can precisely segregate a desired vowel from a noisy vowel even in waveforms, and it can reduce the ASD for sound segregation. However, we cannot conclude that it can precisely segregate a desired vowel form a noisy vowel in hearing, by using only the ASD.] We need to do hearing tests by using a subjective measure. The comparisons with conditions (2) and (3) show that the simultaneous signals can be precisely segregated using the instantaneous amplitude and phase. The comparison with condition (4) shows that the proposed model is more useful than spectral subtraction in the measure of segregation accuracy and the ASD. The comparison with condition (5) shows that the improvements in segregation accuracies at an SNR of 5 dB for simulations 1, 2 (in the case of pink-noise), and 3 were about 12, 8, and 5 dB, respectively.

5 Conclusions

This paper proposes an improved sound segregation model based on auditory scene analysis in order to over-

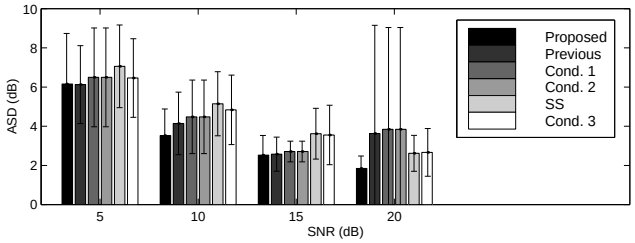

Figure 10: ASD for simulation 3.

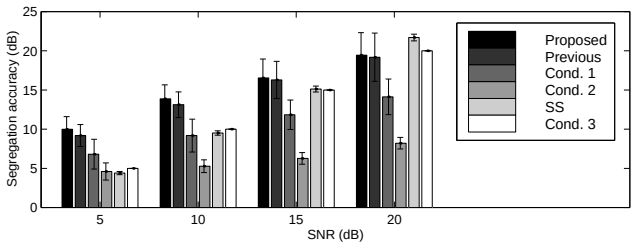

Figure 11: Segregation accuracies for simulation 3.

come three disadvantages in our previous model. This improved model solves the problem of segregating two acoustic sources by using constraints related to the heuristic regularities proposed by Bregman. We first reconsider the estimation method of $C_{k,0}(t)$ and $D_{k,0}(t)$, then incorporate the constraint of channel envelopes with periodicity of the fundamental frequency into the grouping block, and finally consider the constraint of smoothness of $A_k(t)$ on channels.

We demonstrated that the proposed model can improve the previous model and that it can precisely segregate real speech from noisy speech in three simulations of segregating two acoustic sources. The evaluations showed that the proposed model can improve the previous model, and that all constraints related to the four regularities are useful in order to segregate the desired vowel from a noisy vowel. Furthermore, the proposed method can precisely segregate the desired signal from noisy signal, compared with basic spectral subtraction.

In the future work, we will (1) do hearing tests for vowel segregation by using the proposed model and some other model, and (2) improve the proposed model so that it can be applied to consonants-vowel segregation.

6 Acknowledgments

The authors wish to thank H. Kawahara of Wakayama University and T. Irino of ATR-HIP for helpful suggestions about this work, and Y. Ichinose and S. Katagiri of ATR-HIP for the arrangement. This work was supported by a Grant-in-Aid for science research from the Ministry of Education (Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists and No. 10680374) and by CREST (Core Research for Evolutional Science and Technology) of Japan Science and Technology Corporation (JST).

References

[ATR Tech. Rep., 1988] Takeda, K., Sagisaka, Y., Katagiri, S., Abe, M., and Kuwabara, H. Speech Database User's Manual, ATR Technical Report TR-I-0028, 1988.

[Boll, 1979] Boll, S. F. "Suppression of Acoustic Noise in Speech using Spectral Subtraction," IEEE Trans. on Acoustic, Speech, and Signal Processing, Vol. ASSP-27, April, 1979.

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

[Bregman, 1993] Bregman, A.S. "Auditory Scene Analysis: hearing in complex environments," in Thinking in Sounds, (Eds. S. McAdams and E. Bigand), pp. 10–36, Oxford University Press, New York, 1993.

[Brown, 1992] Brown, G.J. "Computational Auditory Scene Analysis: A Representational Approach," Ph.D. Thesis, University of Sheffield, 1992.

[Brown and Hwang, 1992] Brown, R.G., Hwang, P.Y.C., "Introduction to Random Signals and Applied Kalman Filtering," 2nd ed. Wiley, New York, Chapters 5-6, pp. 210-288, 1992.

[Cooke, 1993] Cooke, M. P. "Modeling Auditory Processing and Organization," Ph.D. Thesis, University of Sheffield, 1991 (Cambridge University Press, Cambridge, 1993).

[de Cheveigné, 1993] de Cheveigné, A. "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. 93, 3271–3290, 1993.

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

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

[Hansen and Nandkumar, 1995] Hansen, J. H. L. and Nandkumar, S. "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process," J. Acoust. Soc. Am. 97(6), June 1995.

[Junqua and Haton, 1996] Junqua, J. C. and Haton, J. P. ROBUSTNESS IN AUTOMATIC SPEECH RECOGNITION, – fundamentals and applications –, Kluwer Academic Publishers, Boston, 1996

[Kashino and Tanaka, 1993] Kashino, K. and Tanaka, T., "A sound source separation system with the ability of automatic tone modeling," Proc. of Int. Computer Music Conference, pp. 248–255, 1993.

[Mizumachi and Akagi, 1999] Mizumachi, M. and Akagi, M. "An objective distortion estimator for hearing aids and its application to noise reduction," In Proc. Eurospeech'99, vol. 6, pp. 2619-2622, Budapest, Hungary, Sept. 1999.

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

[Nakatani *et al.*, 1995b] Nakatani, T., Okuno, H. G., and Kawabata, T., "Residue-driven Architecture for Computational Auditory Scene Analysis," In Proc. of IJCAI-95, pp. 165–172, August 1995.

[Papoulis, 1977] Papoulis, A. Signal Analysis. McGraw-Hill, New York, 1977.

[Papoulis, 1991] Papoulis, A. Probability, Random Variables, and Stochastic Process. Third Edition, McGraw-Hill, New York, 1991.

[Shamsunder and Giannakis, 1997] Shamsunder, S. and Giannakis, G. B. "Multichannel Blind Signal Separation and Recognition," IEEE Trans. on Speech and Audio Processing, vol. 5, No. 6, Nov. 1997.

Table 2: Definitions of symbols for Kalman filtering

Symbol	Est. of $B_k(t)$	Est. of $\theta_{2k}(t)$	Est. of $C_{k,0}(t)$	Est. of $D_{k,0}(t)$
Observed signal $\mathbf{y}_m$	$X_k(t_m)$	$\exp(j\phi_k(t_m)$	$X_k(t_m)$	$\exp(j\phi_k(t_m))$
State variable $\mathbf{x}_m$	$B_k(t_m)$	$\exp(j\theta_{2k}(t))$	$C_k(t_m)$	$\exp(jD_k(t_m))$
Observed noise $\mathbf{v}_m$	$X_{1,k}(t_m)$	$X_{1,k}(t_m)/S_k(t_m)$	$X_{2,k}(t_m)$	$X_{2,k}(t_m)/S_k(t_m)$
System noise $\mathbf{w}_m$	w_m	w_m	w_m	w_m
State transition matrix $\mathbf{F}_m$	$1+\frac{B_k(t_m)-B_k(t_{m-1})}{B_k(t_m)}$	$1+\frac{\theta_{2k}(t_m)-\theta_{2k}(t_{m-1})}{\theta_{2k}(t_m)}$	$1+\frac{C_k(t_m)-C_k(t_{m-1})}{C_k(t_m)}$	$1+\frac{D_k(t_m)-D_k(t_{m-1})}{D_k(t_m)}$
Observation matrix $\mathbf{H}_m$	$\exp(j\omega_k t_m)$	$B_k(t_m)/S_k(t_m)$	$\exp(j\omega_k t_m)$	$\hat{C}_k(t_m)/S_k(t_m)$
Driving matrix $\mathbf{G}_m$	1	1	1	1
Initial value $\hat{\mathbf{x}}_{0\|-1}$	$\mathrm{std}(S_k(t))^2$	$\mathrm{cov}(\exp(j\phi_k(t)))$	$\mathrm{std}(\hat{A}_k(t))^2$	$\mathrm{cov}(\exp(j\hat{\theta}_{1k}(t)))$
Initial value $\hat{\mathbf{\Sigma}}_{0\|-1}$	$S_k(t_0)$	$\exp(j\phi_k(t_m))$	$\hat{A}_k(t_0)$	$\exp(j\hat{\theta}_{1k}(t_0))$

[Unoki and Akagi, 1997] Unoki, M. and Akagi, M. "A Method of Signal Extraction from Noise-Added Signal," Electronics and Communications in Japan, Part 3, Vol. 80, No. 11, pp. 1-11, 1997 (in English), Translated from IEICE, vol. J80-A, no. 3, March 1997 (in Japanese).

[Unoki and Akagi, 1999a] Unoki, M. and Akagi, M. "A method of signal extraction from noisy signal based on Auditory Scene Analysis," Speech Communication 27, pp. 261-279, April 1999.

[Unoki and Akagi, 1999b] Unoki, M. and Akagi, M. "Segregation of vowel in background noise using the model of segregating two acoustic sources based on auditory scene analysis," In Proc. EuroSpeech'99, vol. 6, pp. 2575-2578, Budapest, Hungary, Sept. 1999.

Appendix A: Reconsidered estimation method of $C_{k,0}(t)$ and $D_{k,0}(t)$

The system of the Kalman filtering is defined by

$$\mathbf{x}_{m+1} = \mathbf{F}_m\mathbf{x}_m + \mathbf{G}_m\mathbf{w}_m \quad \text{(state)}, \tag{18}$$

$$\mathbf{y}_m = \mathbf{H}_m\mathbf{x}_m + \mathbf{v}_m \quad \text{(observation)}, \tag{19}$$

where the mean and variance of the terms, $\mathbf{x}_0$, $\mathbf{w}_m$, and $\mathbf{v}_m$, are known, and $\mathbf{F}_m$, $\mathbf{G}_m$, $\mathbf{H}_m$, and $\mathbf{v}_m$ are known matrices [Brown and Hwang, 1992]. The Kalman filtering problem determines the minimum variance requirement $\hat{\mathbf{x}}_{m|m}$ from the observed $\mathbf{y}_m$, $m = 0, 1, 2, \cdots, M$ as follows.

$$\hat{\mathbf{x}}_{m|m} = E(\mathbf{x}_m + \mathbf{y}_0, \cdots, \mathbf{y}_m) \tag{20}$$

The minimum variance is sequentially calculated.

1. **Filtering equation**

$$\hat{\mathbf{x}}_{m|m} = \hat{\mathbf{x}}_{m|m-1} + \mathbf{K}_m(\mathbf{y}_m - \mathbf{H}_m\hat{\mathbf{x}}_{m|m-1}) \tag{21}$$

$$\hat{\mathbf{x}}_{m+1|m} = \mathbf{F}_m\hat{\mathbf{x}}_{m|m} \tag{22}$$

2. **Kalman gain**

$$\mathbf{K}_m = \frac{\hat{\mathbf{\Sigma}}_{m|m-1}\mathbf{H}_m^{*T}}{\mathbf{H}_m\hat{\mathbf{\Sigma}}_{m|m-1}\mathbf{H}_m^{*T} + \mathbf{\Sigma}_{v_m}} \tag{23}$$

3. **Covariance equation for the estimated-error**

$$\hat{\mathbf{\Sigma}}_{m|m} = \hat{\mathbf{\Sigma}}_{m|m-1} - \mathbf{K}_m\mathbf{H}_m\hat{\mathbf{\Sigma}}_{m|m-1} \tag{24}$$

$$\hat{\mathbf{\Sigma}}_{m+1|m} = \hat{\mathbf{F}}_m\mathbf{\Sigma}_{m|m}\mathbf{F}_m^{*T} + \mathbf{G}_m\mathbf{\Sigma}_{w_m}\mathbf{G}_m^{*T} \tag{25}$$

4. **Initial state**

$$\hat{\mathbf{x}}_{0|-1} = \bar{\mathbf{x}}_0, \qquad \hat{\mathbf{\Sigma}}_{0|-1} = \mathbf{\Sigma}_{x_0}, \tag{26}$$

The symbols $\bar{}$ and $\mathbf{\Sigma}$ are the mean and variance of a random variable, respectively.

In this paper, we reconsider how to estimate $C_{k,0}$ and $D_{k,0}$ from the observed component $X_k(t)$ using Kalman filter. The estimation duration is $[T_{h-1} - T_h]$. It is then decomposed into discrete time $t_m = m \cdot \Delta t$, $m = 0, 1, 2, \cdots, M$, where the sampling period is $\Delta t = 1/f_s$ and f_s is the sampling frequency.

First, $B_k(t)$ and $\theta_{2k}(t)$ are estimated the Kalman filtering with the parameters in Eqs. (18) and (19) as shown in Table 2. By performing the Kalman filtering according to Eqs. (18) and (19), we obtain the minimal-variance estimated value $\hat{\mathbf{x}}(t_m) = \hat{\mathbf{x}}_{m|m}$ and the covariance matrix $\hat{\mathbf{e}}(t_m) = \hat{\mathbf{\Sigma}}_{m|m}$ at discrete time t_m. Therefore, we obtain the estimated $\hat{B}_k(t) = |\hat{\mathbf{x}}(t)|$ and $\hat{\theta}_{2k}(t) = |\hat{\mathbf{x}}(t)|$.

Next, we obtain the estimated $\hat{A}_k(t)$ and $\hat{\theta}_{1k}(t)$ from Eqs. (4) and (6) from the above parameters.

Finally, $C_{k,0}(t)$ and $D_{k,0}(t)$ are estimated with our previous Kalman filtering [Unoki and Akagi, 1999b] with the parameters in Eqs. (18) and (19) as shown in Table 2. Note that let $C_k(t)$ and $D_k(t)$ be $C_k(t) = \int C_{k,0}(t)dt$ and $D_k(t) = \int D_{k,0}(t)dt$, respectively. By Performing the Kalman filtering according to Eqs. (18) and (19), we obtain the minimal-variance estimated value $\hat{\mathbf{x}}(t_m) = \hat{\mathbf{x}}_{m|m}$ and the covariance matrix $\hat{\mathbf{e}}(t_m) = \hat{\mathbf{\Sigma}}_{m|m}$ at discrete time t_m. As a result, the estimated $\hat{C}_{k,0}(t)$ and $\hat{D}_{k,0}(t)$, and the estimated errors $P_k(t)$ and $Q_k(t)$ are determined by $\hat{C}_{k,0}(t) = |d\hat{\mathbf{x}}(t)/dt|$ and $P_k(t) = |d\hat{\mathbf{e}}(t)/dt|$, $\hat{D}_{k,0}(t) = \arg(d\hat{\mathbf{x}}(t)/dt)$, and $Q_k(t) = \arg(d\hat{\mathbf{e}}(t)/dt)$, respectively.