Segregation of vowel in background noise using the model of segregating two acoustic sources based on auditory scene analysis

Masashi Unoki and **Masato Akagi**
School of Information Science, JAIST
1-1 Asahidai, Tatsunokuchi,
Ishikawa, 923-1292
Japan

Abstract

This paper proposes an auditory sound segregation model based on auditory scene analysis. It solves the problem of segregating two acoustic sources by using constraints related to the heuristic regularities proposed by Bregman and by making an improvement to our previously proposed model. The improvement is to reconsider constraints on the continuity of instantaneous phases as well as constraints on the continuity of instantaneous amplitudes and fundamental frequencies in order to segregate the desired signal from a noisy signal precisely even in waveforms. Simulations performed to segregate a real vowel from a noisy vowel and to compare the results of using all or only some constraints showed that our improved model can segregate real speech precisely even in waveforms using all the constraints related to the four regularities, and that the absence of some constraints reduces the segregation accuracy.

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. It has been investigated by many researchers, who have proposed many methods. For example, in the investigation of robust speech recognition [Furui and Sondhi, 1991], there are noise reduction or suppression [Boll, 1979] and speech enhancement methods [Junqua and Haton, 1996]. In the investigation of signal processing, there is signal estimation using a linear system [Papoulis, 1977] and signal estimation based on a stochastic process for signal 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 in order 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. Therefore, it is difficult to solve this problem 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)", it 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 solve the sound segregation problem (ill-posed inverse problem) uniquely. In addition, it would 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; 1997], and a sound source separation system with the ability of automatic tone modeling [Kashino and Tanaka, 1993]. Typical top-down models include a segregation model based on psychoacoustic grouping rules [Ellis, 1994; 1996] and a computational

model of sound segregation agents [Nakatani *et al.*, 1994; 1995a; 1995b]. All 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 segregating two acoustic sources as a fundamental problem, and considering that it can be uniquely solved using not only amplitude but also phase information and using mathematical constraints related to the four psychoacoustically heuristic regularities [Unoki and Akagi, 1997a; 1999].

This fundamental problem is defined as follows [Unoki and Akagi, 1997a; 1999]. 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.

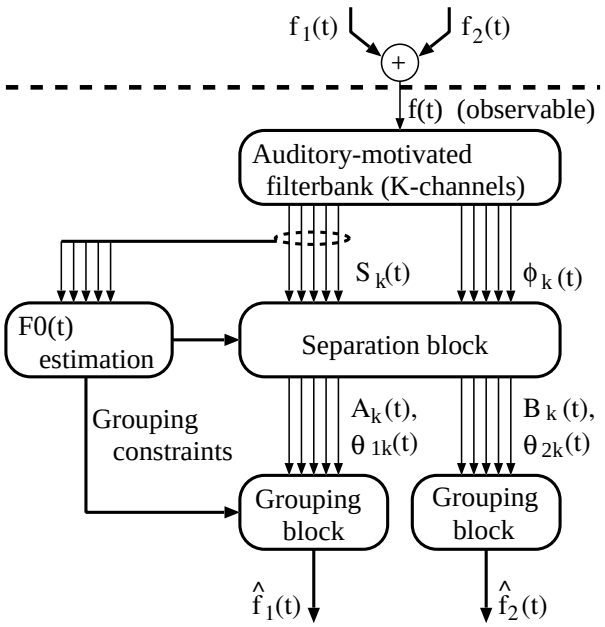

Figure 1: Auditory sound segregation model.

To solve this problem, we have proposed a basic method of solving it using constraints related to the four regularities [Unoki and Akagi, 1997b; 1997c] and the improved method [Unoki and Akagi, 1998; 1999]. However, the former cannot deal with the variation of the fundamental frequency, although it can segregate the synthesized signal from the noise-added signal. And the latter has difficulty completely determining the phases, although it can precisely segregate a vowel from a noisy vowel at certain amplitudes by constraining the continuity of the instantaneous amplitudes and fundamental frequencies.

This paper proposes a new sound segregation method to deal with real speech and noise precisely even in waveforms, by using constraints on the continuity of instantaneous phases as well as constraints on the continuity of instantaneous amplitudes and fundamental frequencies.

2 Auditory sound segregation model

In this paper, it is assumed that the desired signal $f_1(t)$ is 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.

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) **Slowness** (piecewise-differentiable polynomial approximation)	$dA_k(t)/dt = C_{k,R}(t)$ $d\theta_{1k}(t)/dt = D_{k,R}(t)$ $dF_0(t)/dt = E_{0,R}(t)$
(a) A single sound tends to change its properties smoothly and slowly		
(b) A sequence of sounds from the same source tends to change its properties slowly	(b) **Smoothness** (Spline interpolation)	$\sigma_A = \int_{t_a}^{t_b} [A_k^{(R+1)}(t)]^2 dt \Rightarrow \min$ $\sigma_\theta = \int_{t_a}^{t_b} [\theta_{1k}^{(R+1)}(t)]^2 dt \Rightarrow \min$
(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	$n \times F_0(t), \qquad n = 1, 2, \cdots, N_{F_0}$
(iv) Many changes that take place in an acoustic event will affect all the components of the resulting sound in the same way and at the same time	Correlation between the instantaneous amplitudes	$\frac{A_k(t)}{\lVert A_k(t)\rVert} \approx \frac{A_\ell(t)}{\lVert A_\ell(t)\rVert}, \qquad k \neq \ell$

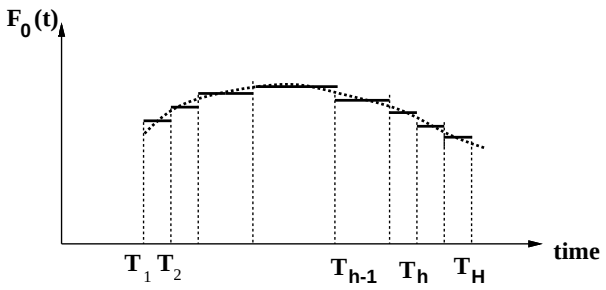

Figure 2: Temporal variation of the fundamental frequency.

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, constructed with $K = 128$, a bandwidth of 60–6000 Hz, and a sampling frequency of 20 kHz [Unoki and Akagi, 1997a; 1999]. $S_k(t)$ and $\phi_k(t)$ are determined by using the amplitude and phase spectra defined by the wavelet transform [Unoki and Akagi, 1997a; 1999].

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, 1998].

In this block, the Comb filter is defined by

$$\mathrm{Comb}(k,l) = \begin{cases} \frac{(\alpha+1)}{(2f_0\alpha^{K-k}(\alpha-1))}, & \omega_k = n\cdot\omega_l, \\ & 1 \le n \le N \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

where k and l are indices, ω_k and ω_l are the center frequencies in channels, and N is the number of harmonics of the highest order. Then, $\hat{l}$, which corresponds to the channel containing the fundamental wave, is determined by

$$\hat{l}(t; L_F) = \arg\max_{l \le L_F} \sum_{k=1}^{K} \mathrm{Comb}(k,l) S_k(t), \quad (11)$$

where L_F is the upper-limited search region of l. The estimated $F_0(t)$ is determined by

$$F_0(t) = \min_{L_F} \mathrm{std}(\omega_{\hat{l}}/2\pi). \quad (12)$$

In this paper, we let the parameters be $N = 10$ and $K/4 \le L_F \le K/2$.

Since the number of channels in $X_k(t)$ is finite, the estimated $F_0(t)$ takes a discrete value as shown in Fig. 2. 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 $F_0(t)$.

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, 1999]. $\hat{f}_1(t)$ 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 [Unoki and Akagi, 1999]. In this paper, onset $T_{k,\rm on}$ and offset $T_{k,\rm off}$ in $X_k(t)$ are determined as follows.

1. Onset $T_{k,\rm on}$ is determined by the nearest maximum point of $|d\phi_k(t)/dt|$ (within 25 ms) to the maximum point of $dS_k(t)/dt$.
2. Offset $T_{k,\rm off}$ is determined by the nearest maximum point of $|d\phi_k(t)/dt|$ (within 25 ms) to the minimum point of $dS_k(t)/dt$.

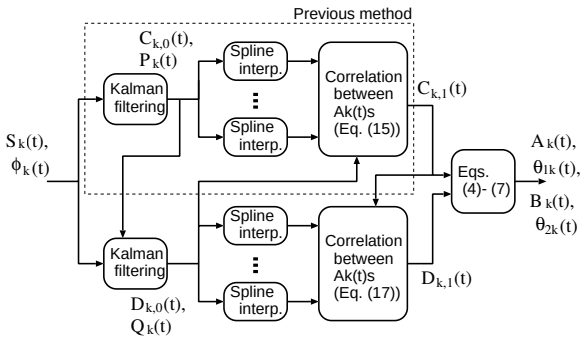

Figure 3: Signal processing of a separation block.

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

$$\ell = \frac{K}{2} - \left\lceil \frac{\log(n \cdot F_0(t)/f_0)}{\log \alpha} \right\rceil, \quad n = 1, 2, \cdots, N_{F_0}, \quad (13)$$

where α is the scale parameter and $\lceil \cdot \rceil$ 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 ($f_0 = 600$) [Unoki and Akagi, 1999].

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. In this paper, the improvement of the auditory sound segregation model is to reconsider the constraints on the continuity of $\theta_{1k}(t)$ as well as the constraints on the continuity of $A_k(t)$ and $F_0(t)$. 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, substituting $dA_k(t)/dt = C_{k,R}(t)$ into Eq. (4), we get 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

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{S_k(t)\cos(\phi_k(t) - \theta_{1k}(t)) + C_k(t)} \right), \quad (14)$$

where $C_k(t) = -\int C_{k,R}(t)dt - C_{k,0} = -A_k(t)$ [Unoki and Akagi, 1999].

The signal flow of the separation block is shown in Fig. 3. 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 (see Appendix B). Then, $\hat{C}_{k,1}(t)$ is determined by using

$$\hat{C}_{k,1} = \underset{\hat{C}_{k,0} - P_k \le C_{k,1} \le \hat{C}_{k,0} + P_k}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||}, \quad (15)$$

where $\hat{A}_k(t)$ is obtained by the spline interpolation and $\hat{\hat{A}}_k(t)$ is determined in across-channel that satisfies constraint (iii) as follows.

$$\hat{\hat{A}}_k(t) = \frac{1}{N_{F_0}} \sum_{\ell \in \mathbf{L}, \ell \ne k} \frac{\hat{A}_\ell(t)}{\|\hat{A}_\ell(t)\|}, \quad (16)$$

where $\mathbf{L}$ is a set of ℓ that satisfies Eq. (13). Finally, $\hat{D}_{k,1}(t)$ is determined by using

$$\hat{D}_{k,1} = \underset{\hat{D}_{k,0} - Q_k \le D_{k,1} \le \hat{D}_{k,0} + Q_k}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||}. \quad (17)$$

Since $\theta_{1k}(t)$ and $\theta_k(t)$ are determined from $\hat{D}_{k,1}(t)$ and $\hat{C}_{k,1}(t)$, 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.

2.5 Overview of the proposed model

An overview of signal processing of the proposed model is shown in Fig. 4. First, noisy vowel /a/ $f(t)$ shown in Fig. 4 A (the SNR of $f(t)$ is 10 dB) is decomposed into $S_k(t)$ and $\phi_k(t)$ as shown in Figs. 4 B and C, respectively. Next, $F_0(t)$ is estimated as shown in Fig. 4 D. The concurrent time-frequency region of the desired signal $f_1(t)$ is determined using constraints (i) and (iii) as shown in Figs. 4 E and F. Finally, the instantaneous amplitudes and the instantaneous phases of the two signals are determined from $S_k(t)$ and $\phi_k(t)$ using constraints (i-i) and (iv). The determined $A_k(t)$ and $\theta_{1k}(t)$ are shown in Figs. 4 H and I, respectively. The segregated signal $\hat{f}_1(t)$ is shown in Fig. 4 J. In this figure, note that the segregated $B_k(t)$, $\theta_{2k}(t)$, and $\hat{f}_2(t)$ are omitted.

3 Simulations

To show that the proposed model can segregate the desired signal $f_1(t)$ from a noisy signal $f(t)$ precisely even in waveforms, we evaluated the following two issues by using four simulations. One is to evaluate the advantage of constraints, and the other one is to evaluate whether or not the proposed model can precisely segregate the desired vowel from noisy vowel. In these simulations, we used the following signals:

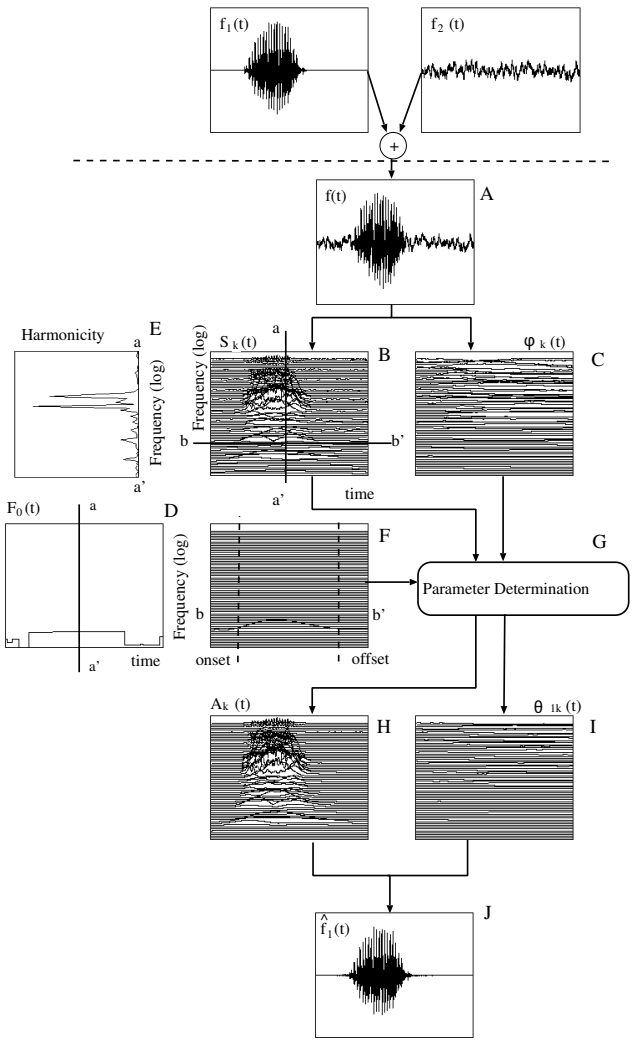

Figure 4: Overview of signal-flow in the proposed model.

(a) noisy synthesized AM–FM harmonic complex tone (LMA-synthesis vowel /a/) [Unoki and Akagi, 1998];

(b) noisy real vowel (/a/, /i/, /u/, /e/, /o/);

(c) noisy real continuous vowel (/aoi/); and

(d) concurrent double vowel (signal (b) + signal (c)),

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

We used an evaluation measure such as SNR to evaluate the segregation performance of the proposed method, as defined by

$$10\log_{10}\frac{\int_0^T f_1(t)^2 dt}{\int_0^T (f_1(t)-\hat{f}_1(t))^2 dt}, \tag{18}$$

where $f_1(t)$ is the original signal and $\hat{f}_1(t)$ is the segregated signal. This measure is called "segregation accuracy."

3.1 Evaluation of the constrains

To show the advantages of the constraints in Table 1, we compared the performances of our method under 11-conditions as shown in Table 2. In this simulation, we used two types of desired signal $f_1(t)$: simulation signal (a) and vowel /a/ of the male speaker (mau) in simulation signal (b); $f_2(t)$ was bandpassed pink-noise. Conditions 1 – 3 denote sound segregation using three of the constraints. Note that constraints (ii-b) and (iv) cannot be separately used because $C_{k,1}(t)$ and $D_{k,1}(t)$ are uniquely determined using these constraints. Conditions 4 – 6 denote sound segregation using the two of the constraints. Conditions 7 – 9 denote sound segregation using the only one of the constraints. Note that utilizing constraint (ii-a) corresponds to estimating $A_k(t)$ and $\theta_{1k}(t)$ using the Kalman filter for any channels. Condition 10 denotes sound segregation without all the constraints.

Segregation accuracy in this simulation for AM–FM complex tones is shown in Fig. 5. Segregation accuracy in this simulation for vowel /a/ is shown in Fig. 6. In these figures, segregation accuracy values above the dashed line show the improved accuracy, that is, noise reduction. The results show that segregation accuracy achieved by the proposed model was the best among the constraints. Moreover, comparisons between four groups (conditions 0–1–3–4–6–7, 0–1–2–3–4–5–9, 0–1–2–5–6–8, and 0–2–3) show that the absence of some constraints reduces the accuracy. These groups were selected by focusing on only one constraint in Table 2 and by omitting some of the constraints in turn. Hence, all the constraints related to the four regularities are useful for segregating the desired vowel from a noisy vowel.

An example of segregation in case of Fig. 5 (c) is shown in Fig. 7. An example of segregation in case of Fig. 6 (c) is shown in Fig. 8.

3.2 Evaluation of the proposed model

To show that the proposed method can segregate the desired vowel from a noisy vowel precisely even in waveforms, we performed three simulations using signals (b)–(d) under the following conditions:

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/) of the male (mau) or female (fkn) speaker and the other was /aoi/ of the female (fsu) or male (mht) speaker, and the dataset size was 40 (five vowels, two speakers, and four noise signals).

In addition, we compared the performances of the proposed model with those of other typical method (using constraints 1, 8 and, 10) for the above simulations. The

Table 2: Comparisons of constraints.

Constraints /condition No.	Proposed model	1	2	3	4	5	6	7	8	9	10
(i) synchronism of onset/offset	o	o	x	o	o	x	o	o	x	x	x
(ii-a) slowness	o	o	o	o	o	o	x	x	x	o	x
(iii) harmonicity	o	o	o	x	x	o	o	x	o	x	x
(ii-b) smoothness	o	x	o	o	x	x	x	x	x	x	x
(iv) correlation	o	x	o	o	x	x	x	x	x	x	x

"o": used constraint, "x": unused constraint

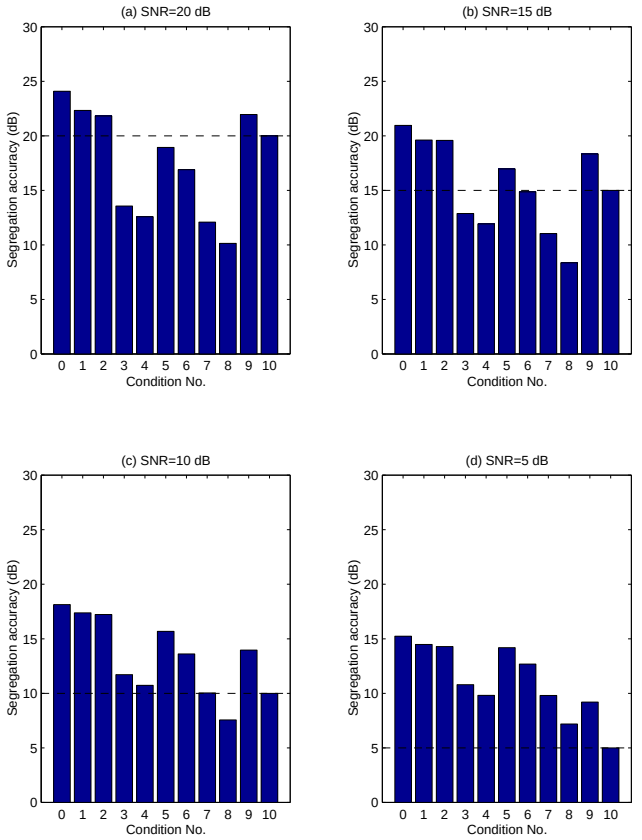

Figure 5: Segregation accuracies for evaluation using signal (a). Condition #0 denotes the use of the proposed method.

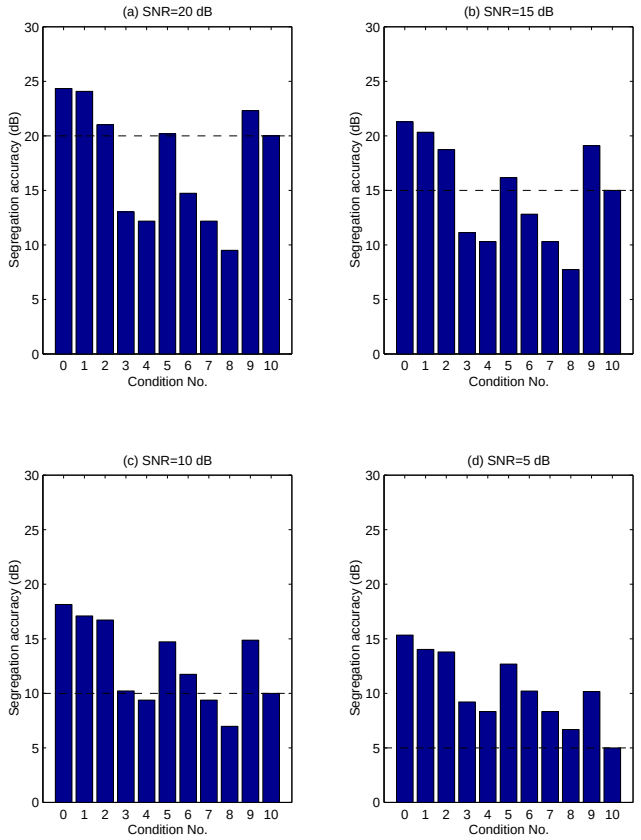

Figure 6: Segregation accuracies for evaluation using vowel /a/ of the male speaker in signal (b). Condition #0 denotes the use of the proposed method.

other methods using constraints 1, 8, and 10 correspond to:

(1) extracting the harmonics using the Comb filter and estimating $A_k(t)$ and $\theta_{1k}(t)$ using the Kalman filter,

(2) extracting the harmonics using the Comb filter, and

(3) doing nothing.

Comparison with condition (1) shows that the proposed method has the advantage of the smoothness of the instantaneous amplitude and phase, and comparison with condition (2) shows that it has the advantage of the instantaneous phase. Moreover, comparison with condition (3) shows that the proposed method's segregation accuracy was improved.

The segregation accuracies in the three simulations are shown in Figs. 9, 10 and 12. 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 obtained better segregation accuracy than the other three methods. They show that it can segregate the desired vowel from a noisy vowel precisely even in waveforms. In addition, the comparison between the proposed method and condition (2) shows that the simultaneous signals can be precisely segregated using the instantaneous amplitude and phase. The comparison with condition (3) shows that the improve-

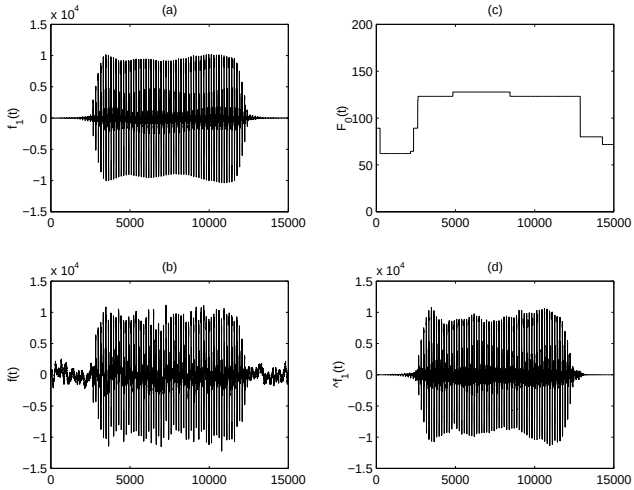

Figure 7: Example of segregation for noise-added AM-FM harmonic complex tone: (a) original /a/ $f_1(t)$, (b) mixed signal $f(t)$, (c) fundamental frequency $F_0(t)$, (d) segregated signal $\hat{f}_1(t)$.

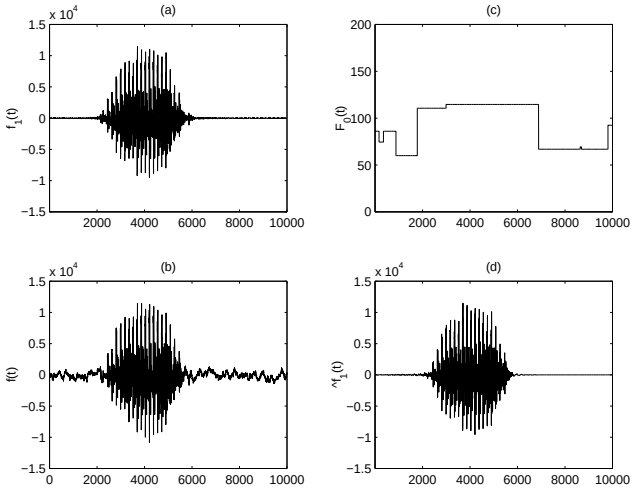

Figure 8: Example of segregation for noisy vowel /a/ of the male speaker (mau): (a) original /a/ $f_1(t)$, (b) mixed signal $f(t)$, (c) fundamental frequency $F_0(t)$, (d) segregated signal $\hat{f}_1(t)$.

ments in segregation accuracies at an SNR of 5 dB for simulations 1, 2 (in the case of pink-noise), and 3 were about 9, 7, and 4 dB, respectively.

3.3 Considerations

The above results show the advantage of the proposed method. However, when the SNR of $f(t)$ is 20 dB, the improvements in segregation accuracy using the proposed model, shown in Figs. 9, 10, and 12, were smaller. We suspect that the order of R for the polynomial approximation ($C_{k,R}(t)$, $D_{k,R}(t)$, and $E_{0,R}(t)$) affects the improvement in segregation accuracy obtainable using the proposed model.

The results of simulations 1 and 2 showed that the difference in segregation accuracy improvement depended on the two types of noise. We suspect that the construction of the constant Q filterbank used here affected the segregation accuracy depending on the noise type. Since it has a constant Q on any channel (the same filter shape within all channels), the power of the components for which pink noise passed through the channels would be distributed approximately equally, while the power of the components for which white noise passed through the channels would be concentrated at higher frequencies. On the other hand, the harmonic components of the desired signal should not be satisfied precisely at a higher frequency, while they should be satisfied precisely at a lower frequency. Therefore, we consider that the interplay between the above factors reduces the improvement in segregation accuracy obtained by using the proposed model, as shown in Figs. 9 (b) and 10 (b).

4 Conclusions

This paper proposed a new method of extracting the desired speech from noisy speech precisely even in waveforms by using constraints on the continuity of instantaneous phases as well as those on the continuity of instantaneous amplitudes and fundamental frequencies.

In order to show that the proposed model can extract real speech from noisy speech precisely even in waveforms, we demonstrated one evaluation and three simulations of segregating two acoustic sources. The result of evaluation showed that all constraints related to the four regularities are useful in order to segregate the desired vowel from a noisy vowel. The results of the three simulations showed that the proposed method can segregate the desired vowel from a noisy vowel precisely even in waveforms. It was also shown that the proposed method can precisely segregate the desired signal from the simultaneous signals using the instantaneous amplitude and phase.

Future work includes (1) reconsidering the order of the polynomial approximation of $C_{k,R}(t)$ and $D_{k,R}(t)$ for vowel segregation and (2) improving the proposed model so that it can be applied to consonants-vowel segregation.

5 Acknowledgments

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).

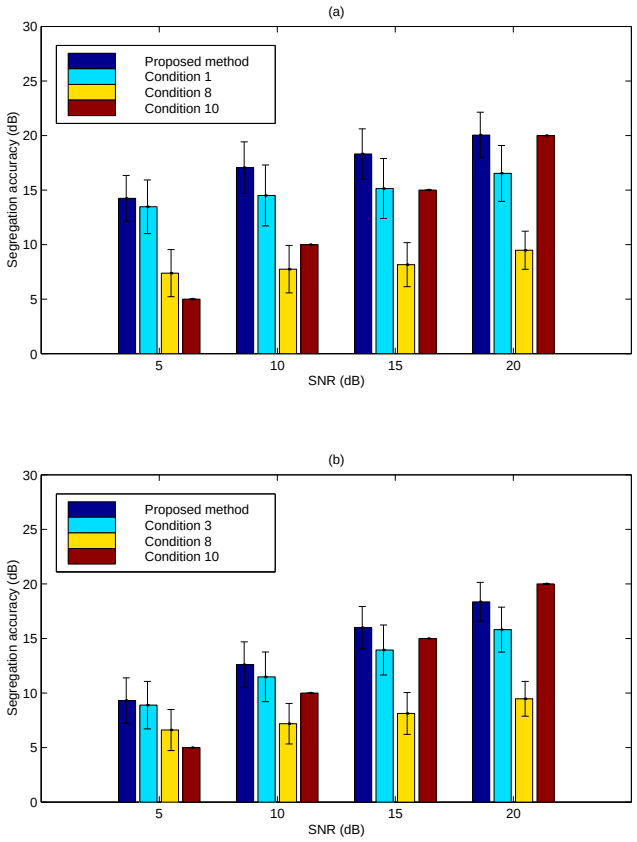

Figure 9: Segregation accuracy for simulation 1: (a) bandpassed pink noise, (b) bandpassed white noise.

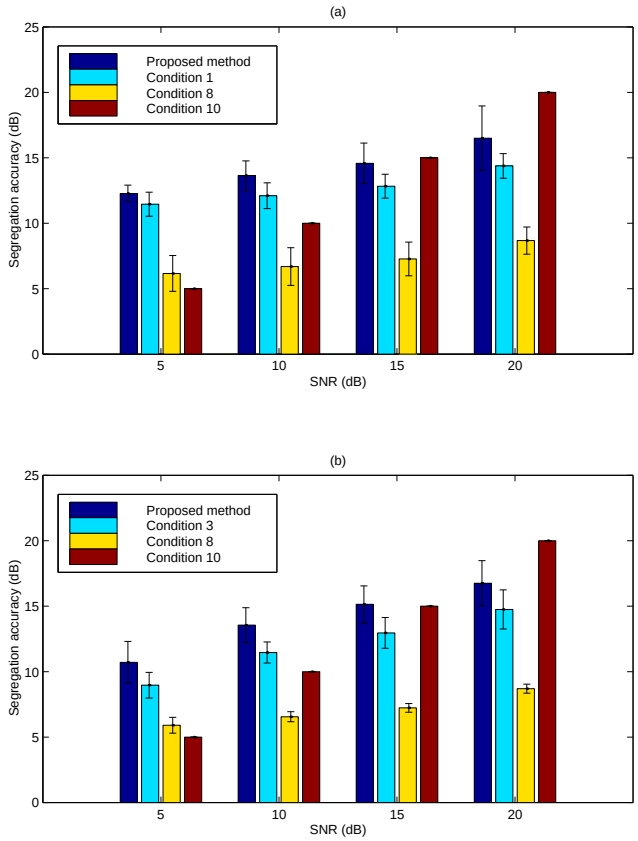

Figure 10: Segregation accuracy for simulation 2: (a) bandpassed pink noise, (b) bandpassed white noise.

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.

[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.

[de Cheveigné, 1997] de Cheveigné, A. "Concurrent vowel identification III: A neural model of harmonic interference cancellation," J. Acoust. Soc. Am. 101, 2857–2865, 1997.

[Ellis, 1994] Ellis, D. P. W. "A Computer Implementation of Psychoacoustic Grouping Rules," Proc. 12th Int. Conf. on Pattern Recognition, 1994.

[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.

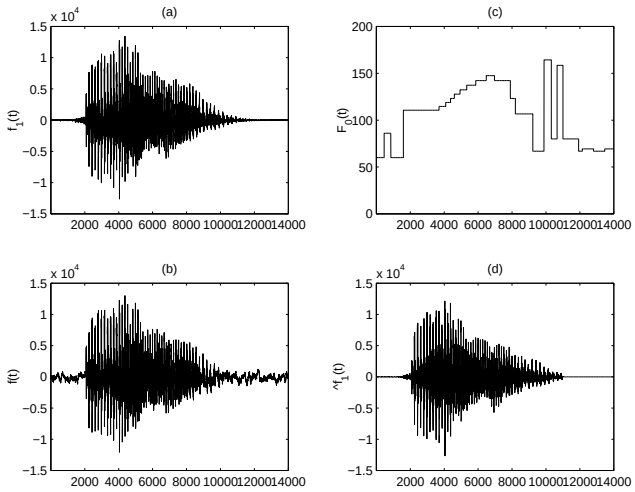

Figure 11: Example of simulation 2: (a) original /a/ $f_1(t)$, (b) mixed signal $f(t)$, (c) fundamental frequency $F_0(t)$, (d) segregated signal $\hat{f}_1(t)$.

[Kashino and Tanaka, 1994] Kashino, K. and Tanaka, H. "A Computational Model of Auditory Segregation of Two Frequency Components — Evaluation and Integration of Multiple Cues —," IEICE Vol. J77-A No. 5, pp. 731–740, May 1994.

[Nakatani *et al.*, 1994] Nakatani, T., Okuno, H. G., and Kawabata, T. "Unified Architecture for Auditory Scene Analysis and Spoken Language Processing," In Proc. of ICSLP '94, 24, 3, 1994.

[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.

[Unoki and Akagi, 1997a] 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, 1997b] Unoki, M. and Akagi, M. "A Method of Signal Extraction from Noisy Signal based on Auditory Scene Analysis," In Proc. IJCAI-97 Workshop on CASA'97, pp. 93-102, Nagoya, Japan, August 1997.

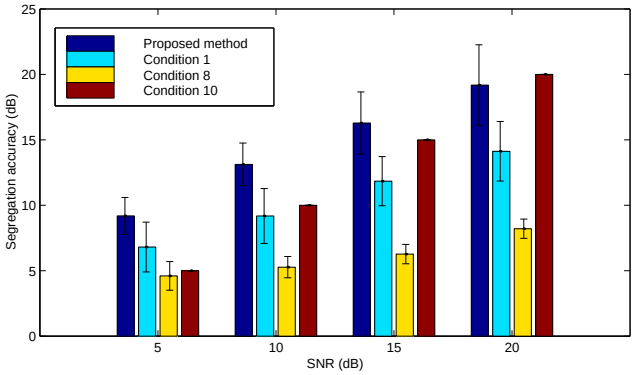

Figure 12: Segregation accuracies for simulation 3.

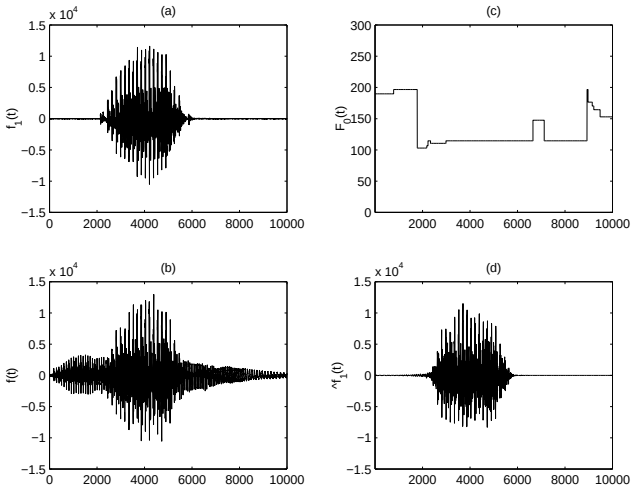

Figure 13: Example of simulation 3: (a) original /a/ $f_1(t)$, (b) mixed signal $f(t)$, (c) fundamental frequency $F_0(t)$, (d) segregated signal $\hat{f}_1(t)$.

[Unoki and Akagi, 1997c] Unoki, M. and Akagi, M. "A Method of Signal Extraction from Noisy Signal," In Proc. EuroSpeech'97, vol. 5, pp. 2583-2586, RHODOS-GREECE, Sept. 1997.

[Unoki and Akagi, 1998] Unoki, M. and Akagi, M. "Signal Extraction from Noisy Signal based on Auditory Scene Analysis," In Proc. ICSLP'98, vol. 4, pp. 1515–1518, Sydny, Australia, Dec. 1998.

[Unoki and Akagi, 1999] 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.

Appendix A: Determining the region estimated by the Kalman filtering

In this paper, we consider how to estimate $C_{k,0}$ and $D_{k,0}$ from the observed component $X_k(t)$ using the 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 $\Delta t = 1/f_s$

Table 3: Definitions of symbols for the Kalman filtering

Symbol	Estimation of $C_{k,0}(t)$	Estimation of $D_{k,0}(t)$
Observed signal $\mathbf{y}_m$	$X_k(t_m)$	$\exp(j\phi_k(t_m))$
State variable $\mathbf{x}_m$	$C_k(t_m)$	$\exp(jD_k(t_m))$
Observed noise $\mathbf{v}_m$	$X_{2,k}(t_m)$	$X_{2,k}(t_m)/S_k(t_m)$
System noise $\mathbf{w}_m$	w_m	w_m
State transition matrix $\mathbf{F}_m$	$\Delta C_k(t_m)$	$\Delta D_k(t_m)$
Observation matrix $\mathbf{H}_m$	$\exp(j\omega_k t_m)$	$\hat{C}_k(t_m)/S_k(t_m)$
Driving matrix $\mathbf{G}_m$	1	1
Initial value $\hat{\mathbf{x}}_{0\|-1}$	0	1
Initial value $\hat{\mathbf{\Sigma}}_{0\|-1}$	$S_k(t_0)$	$\exp(j\phi_k(t_0))$

and f_s is the sampling frequency. First, 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. Here, let the temporal variations of $C_k(t)$ and $D_k(t)$ at discrete time t_m be

$$C_k(t_{m+1}) = C_k(t_m)\Delta C_k(t_m) + w_m, \quad (19)$$

$$\Delta C_k(t_m) = 1 + \frac{C_k(t_m) - C_k(t_{m-1})}{C_k(t_m)}, \quad (20)$$

$$D_k(t_{m+1}) = D_k(t_m)\Delta D_k(t_m) + w_m, \quad (21)$$

$$\Delta D_k(t_m) = 1 + \frac{D_k(t_m) - D_k(t_{m-1})}{D_k(t_m)}, \quad (22)$$

where $t_0 = T_{h-1}$ and $t_M = T_h$. It is assumed that the variation error is represented by white noise with mean 0 and variance σ_m^2.

Next, for the system of the Kalman filtering problem:

$$\mathbf{x}_{m+1} = \mathbf{F}_m\mathbf{x}_m + \mathbf{G}_m\mathbf{w}_m \quad \text{(state)}, \quad (23)$$

$$\mathbf{y}_m = \mathbf{H}_m\mathbf{x}_m + \mathbf{v}_m \quad \text{(observation)}, \quad (24)$$

in order to estimate $C_{k,0}(t)$, we apply Eq. (19) to Eq. (23) and apply Eq. (1) to Eq. (24). Using the same steps as for the system of the Kalman filtering problem, in order to estimate $D_{k,0}(t)$, we apply Eq. (21) to Eq. (23) and apply the normalized Eq. (1) to Eq. (24). The parameters in Eqs. (23) and (24) are shown in Table 3.

In these systems, 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. The Kalman filtering problem is to determine the minimum variance requirement $\hat{\mathbf{x}}_{m|m}$ from the observed $\mathbf{y}_m$, $m = 0, 1, 2, \cdots, M$ as follows.

$$\hat{\mathbf{x}}_{m|m} = E(\mathbf{x}_m + \mathbf{y}_0, \cdots, \mathbf{y}_m) \quad (25)$$

It is calculated by sequentially solving the following equations:

1. **Filtering equation**

$$\hat{\mathbf{x}}_{m|m} = \hat{\mathbf{x}}_{m|m-1} + \mathbf{K}_m(\mathbf{y}_m - \mathbf{H}_m\hat{\mathbf{x}}_{m|m-1}) \quad (26)$$

$$\hat{\mathbf{x}}_{m+1|m} = \mathbf{F}_m\hat{\mathbf{x}}_{m|m} \quad (27)$$

2. **Kalman gain**

$$\mathbf{K}_m = \frac{\hat{\mathbf{\Sigma}}_{m|m-1}\mathbf{H}_m^{*T}}{\mathbf{H}_m\hat{\mathbf{\Sigma}}_{m|m-1}\mathbf{H}_m^{*T} + \mathbf{\Sigma}_{v_m}} \quad (28)$$

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

$$\hat{\mathbf{\Sigma}}_{m|m} = \hat{\mathbf{\Sigma}}_{m|m-1} - \mathbf{K}_m\mathbf{H}_m\hat{\mathbf{\Sigma}}_{m|m-1} \quad (29)$$

$$\hat{\mathbf{\Sigma}}_{m+1|m} = \hat{\mathbf{F}}_m\mathbf{\Sigma}_{m|m}\mathbf{F}_m^{*T} + \mathbf{G}_m\mathbf{\Sigma}_{w_m}\mathbf{G}_m^{*T} \quad (30)$$

4. **Initial state**

$$\hat{\mathbf{x}}_{0|-1} = \bar{\mathbf{x}}_0, \qquad \hat{\mathbf{\Sigma}}_{0|-1} = \mathbf{\Sigma}_{x_0}, \quad (31)$$

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

Finally, performing the Kalman filtering according to Eqs. (23) and (24), 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.

Appendix B: Candidate selection of $C_{k,1}(t)$ and $D_{k,1}(t)$ using the spline interpolation

In order to determine whether $A_k(t)$ and $\theta_{1k}(t)$ satisfy constraint (ii-b) as shown in Table 1, consider the selection of candidates for $C_{k,1}(t)$ and $D_{k,1}(t)$.

To estimate $C_{k,1}(t)$ and $D_{k,1}(t)$, where $R = 1$, that satisfy constraint (ii-b), we interpolate $A_k^{(R+1)}(t)$ and $\theta_{1k}^{(R+1)}(t)$, where $R = 1$, $A_k^{(R+1)}(t) = A_{k,i}$, and $\theta_{1k}^{(R+1)}(t) = \theta_{1k,i}$, $i = 1, 2, \cdots, I$ in $[t_a, t_b]$. According to constraint (ii-b), the smoothest interpolation function is the $(2R+1)$th-order spline function. This spline function is unique.

First, we determine candidates of $C_{k,1}(t)$ and $D_{k,1}(t)$ using the spline function within the estimated error region: $\hat{C}_{k,0}(t) - P_k(t) \le C_{k,1}(t) \le \hat{C}_{k,0}(t) + P_k(t)$ and $\hat{D}_{k,0}(t) - Q_k(t) \le D_{k,1}(t) \le \hat{D}_{k,0}(t) + Q_k(t)$. Then, selecting a correct solution from the candidates of $C_{k,1}(t)$ and $D_{k,1}(t)$, we can uniquely determine the smoothest $A_k(t)$ and $\theta_{1k}(t)$ from $C_{k,1}(t)$ and $D_{k,1}(t)$, respectively.

In this paper, we use the cubic spline function $(2R+1)$, where $R = 1$. The interpolation region is from $t_a = T_{h-1}$ to $t_b = T_h$. The interpolation interval is $\Delta\tau = 15 \times (2\pi/\omega_k)\Delta t$. Therefore, $I = \lceil (t_b - t_a)/\Delta\tau \rceil$.