TITLE

A Method of Signal Extraction from Noisy Signal based on Auditory Scene Analysis

AUTHORS

† Masashi UNOKI

† Masato AKAGI

AFFILIATION

† School of Information Science,
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa-ken, 923-1292 Japan

Abstract

This paper proposes a method of extracting the desired signal from a noisy signal, addressing the problem of segregating two acoustic sources as a model of acoustic source segregation base on Auditory Scene Analysis. Since the problem of segregating two acoustic sources is an ill-inverse problem, constraints are needed to determine an unique solution. The proposed method uses the four heuristic regularities proposed by Bregman as constraints and uses the instantaneous amplitude and phase of noisy signal components that have passed through a wavelet filterbank as features of acoustic sources. Then the model can extracts the instantaneous amplitude and phase of the desired signal. Simulations were performed to segregate the harmonic complex tone from a noise-added harmonic complex tone and to compare the results of using all or only some constraints. The results show that the method can segregate the harmonic complex tone precisely using all the constraints related to the four regularities and that the absence some constraints reduces the accuracy.

Résumé

Cet article propose une méthode pour extraire le signal voulu à partir d'un signal bruité addressant ainsi le problème de ségrégation de deux sources acoustiques comme modèle de ségrégation de sources acoustiques basé sur l'analyse de la scène auditive. Comme le problème de ségrégation de deux sources acoustiques est un problème mal-inversé il est nécessaire d'utiliser des contraintes afin de déterminer une solution unique. La méthode proposée adopte les quatres régles proposées par Bregman comme contraintes physiques et utilise comme propriétés des sources acoustiques, l'amplitude instantannée et la phase des composants du signal bruité après son passage par une banque de filtres. Le modèle proposé peut alors extraire l'amplitude instantannée et la phase du signal voulu. Des simulations pour la ségrégation du ton harmonique complexe à partir d'un ton harmonique bruité et la comparaison des résultats obtenus lors de l'utilisation de tous ou d'une partie des contraintes ont été effectuées. Les resultats montrent que la méthode proposée peut effectuer une ségrégation précise du ton harmonic complexe lors de l'utilisation de toutes les contraintes en relation avec les quatres régles de Bregman et une diminution de la précision lors de l'utilisation de seulement une partie de ces contraintes.

Zusammenfassung

In der vorliegenden Abhandlung wird ein Verfahren zur Extraktion des gewünschten Signals aus dem verrauschten Signal vorgeschlagen, um so das Problem der Trennung von zwei akustischen Signalquellen in Form eines Modells für Akustiksignalquellen-Trennung auf der Basis von "Auditory Scene Analysis" zu lösen. Da das Problem der Trennung von zwei akustischen Signalquellen ein inverses ILL-Problem darstellt, sind zur Ermittlung einer eindeutigen Lösung Einschränkungen erforderlich. Das vorgeschlagene Verfahren beruht auf den vier heuristischen Regelmäβigkeiten von Bergman als Einschränkungen und macht sich als Merkmale von akustischen Signalquellen die Momentanamplitude und Phase von Rauschsignalkomponenten zunutze, die eine "Wavelet"-Filterbank passiert haben. Anschlieβend kann das Modell die Momentanamplitude und Phase des gewünschten Signals extrahieren. Es werden Simulationen zur Trennung einer komplexen Oberwelle von einer verrauschten komplexen Oberwelle sowie der Vergleich der Resultate unter Anwendung aller bzw. einiger Einschränkungen durchgefuhrt. Die Ergebnisse zeigen, daβ mit Hilfe des vorgeschlagenen Verfahrens eine präzise Trennung von komplexen Oberwellen mit allen Einschrankungen in bezug auf die vier Regelmäβigkeiten möglich ist und daβ unter dem Fehlen bestimmter Einschränkungen die Genauigkeit leidet.

Keyword

auditory scene analysis, Bregman's regularities, the problem of segregating two acoustic sources, constant Q gammatone filterbank

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 , Shamsunder and Giannakis, 1997] and signal estimation based on a stochastic process for signal and noise [Papoulis, 1991].

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

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 reported that to perform the problem of ASA the human auditory system uses four psychoacoustically heuristic regularities related to acoustic events:

(i) common onset and offset,

(ii) gradualness of change,

(iii) harmonicity, and

(iv) changes occurring in the acoustic event [Bregman, 1993].

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 segregation models already exist. There are two main types of models of auditory segregation, based on either bottom-up or top-down processes.

Typical bottom-up models include an auditory segregation model based on acoustic events [Brown, 1992 , Cooke, 1993], a concurrent harmonic sounds segregation model based on the fundamental frequency [de Cheveigné, 1993 ,de Cheveigné, 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 , Ellis, 1996] and a computational model of sound segregation agents [Nakatani *et al.*, 1994 , Nakatani *et al.*, 1995a , Nakatani *et al.*, 1995b]. All these segregation models use regularities (i) and (iii), and the amplitude (or power) spectrum as the acoustic feature, so, they cannot completely extract the desired signal from a noisy signal if the signal and noise exist in the same frequency region.

We think that, using the same approach as in CASA, it should be possible to solve the signal segregation problem (ill-problem) uniquely, using constraints related to the four regularities. In addition, we have discussed the need to use not only the amplitude spectrum but also the phase spectrum in order to completely extract the desired signal from a noisy signal in which the signal and noise exist in the same frequency region [Unoki and Akagi, 1997a]. There have been two investigations based on this idea.

As the first step, the problem of segregating a sinusoidal signal from a noise-added sinusoidal signal can be solved using constraints related to two of the four regularities, (ii) and (iv) [Unoki and Akagi, 1997a]. Then, the problem of segregating an amplitude modulated (AM) complex tone from noise-added or concurrent AM complex tones can be solved using the four regularities [Unoki and Akagi, 1997b].

This paper introduces the general problem of segregating two acoustic source as summary of the above results. Then it proposes a method of extracting the desired signal (harmonic tone) from a noisy signal (noisy harmonic tone) based on auditory scene analysis.

2 Auditory segregation model

In this paper, we define the problem of segregating two acoustic sources as "segregating a mixed signal into the original signal components, where the mixed signal is composed of two signals generated by any two acoustic sources". The essential idea of the proposed model is that (a) the observed signal is decomposed into its frequency components by an auditory filterbank (frequency decomposition), (b) components of each signal are segregated from the decomposed components, and (c) components of each signal are grouped by a grouping process into each signal. The auditory segregation model is shown in Fig. 1. This process is formulated as follows.

Fig. 1

2.1 Formulation of the problem of segregating two acoustic sources

First, only the mixed signal $f(t)$, where $f(t) = f_1(t) + f_2(t)$, can be observed in the proposed model. Here, $f_1(t)$ is the desired signal and $f_2(t)$ is a noise or the other signal. The observed signal $f(t)$ is decomposed into its frequency components by an auditory-motivated filterbank (the number of channels is K). The output of the k-th channel $X_k(t)$ is represented by

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

where $X_{1,k}(t)$ and $X_{2,k}(t)$ are components of $f_1(t)$ and $f_2(t)$ that have passed through the filterbank, respectively.

Second, the outputs of the k-th channel, which correspond to $f_1(t)$ and $f_2(t)$, are assumed to be

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

and

$$X_{2,k}(t) = B_k(t)\exp(j\omega_k t + j\theta_{2k}(t)). \tag{4}$$

Here, ω_k is the center frequency of the k-th channel (the auditory filter) and $\theta_{1k}(t)$ and $\theta_{2k}(t)$ are the instantaneous input phases of $f_1(t)$ and $f_2(t)$, respectively. Using this assumption, the instantaneous amplitude $S_k(t)$ and the instantaneous output phase $\phi_k(t)$ are represented by

$$S_k(t) = \sqrt{A_k^2(t) + 2A_k(t)B_k(t)\cos\theta_k(t) + B_k^2(t)} \tag{5}$$

and

$$\phi_k(t) = \arctan\left(\frac{A_k(t)\sin\theta_{1k}(t) + B_k(t)\sin\theta_{2k}(t)}{A_k(t)\cos\theta_{1k}(t) + B_k(t)\cos\theta_{2k}(t)}\right). \tag{6}$$

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

$$A_k(t) = \frac{S_k(t)\sin(\theta_{2k}(t) - \phi_k(t))}{\sin\theta_k(t)} \tag{7}$$

and

$$B_k(t) = \frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{\sin\theta_k(t)}, \tag{8}$$

where $\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$ and $\theta_k(t) \neq n\pi, n \in \mathbf{Z}$. Focusing on the output value of the k-th channel at time t, the relationships between every instantaneous amplitude and every instantaneous phase are shown in Fig. 2.

Hence, since the instantaneous amplitude $S_k(t)$ and the instantaneous output phase $\phi_k(t)$ are observable (see Sec. 3.1.1), and if the instantaneous input phases $\theta_{1k}(t)$ and $\theta_{2k}(t)$ are determined, then $A_k(t)$ and $B_k(t)$ can be determined by the above equations.

Finally, $f_1(t)$ and $f_2(t)$ can be reconstructed by using the grouping of the instantaneous amplitude and the instantaneous phase for all channels. Thus, $\hat{f}_1(t)$ and $\hat{f}_2(t)$ are the reconstructed $f_1(t)$ and $f_2(t)$, respectively.

However, in the above formulation, it is difficult to uniquely and simultaneously determine the instantaneous amplitudes ($A_k(t)$ and $B_k(t)$) and the instantaneous phases ($\theta_{1k}(t)$ and $\theta_{2k}(t)$) using $S_k(t)$ and $\phi_k(t)$, because there are currently no equations for determining two such instantaneous phases and the segregation of two acoustic sources is an ill-inverse problem. Therefore, in this paper, we try solving the problem of segregating two acoustic sources by constraining the desired signal using the four regularities.

Fig. 2

2.2 Assumption and constraints of the proposed model

In this paper, it is assumed that the desired signal $f_1(t)$ is a harmonic complex tone, consisting of the fundamental frequency $F_0(t)$ and the harmonic components, which are multiplies of $F_0(t)$. The proposed model segregates the desired signal from the mixed signal by constraining the temporal differentiation of the instantaneous amplitude, the instantaneous phase, and the fundamental frequency. Here, the relationship between the four regularities [Bregman, 1993] and the constraints concerned is shown in Table 2. These constraints are defined as follows.

Tables 1 and 2

[Constraint1] (Gradualness of change (polynomial approximation)) Temporal differentiations of the instantaneous amplitude $A_k(t)$, the instantaneous phase $\theta_{1k}(t)$, and the fundamental frequency $F_0(t)$ must be represented by an R-th-order differentiable piecewise polynomial as follows:

$$\frac{dA_k(t)}{dt} = C_{k,R}(t), \tag{9}$$

$$\frac{d\theta_{1k}(t)}{dt} = D_{k,R}(t), \tag{10}$$

and

$$\frac{dF_0(t)}{dt} = E_{0,R}(t), \tag{11}$$

where $C_{k,R}(t)$, $D_{k,R}(t)$, and $E_{0,R}(t)$ are R-th-order differentiable piecewise polynomials. Here, $A_k(t)$, $\theta_{1k}(t)$, and $F_0(t)$ are represented by $A_k(t) = \int C_{k,R}(t)dt + C_{k,0}$, $\theta_{1k}(t) = \int D_{k,R}(t)dt + D_{k,0}$, and $F_0(t) = \int E_{0,R}(t)dt + E_{0,0}$, respectively. □

[Constraint2] (Harmonicity) $F_0(t)$ is the fundamental frequency, and N_{F_0} is the number of harmonics of the highest order. The harmonic component must satisfy

$$n \cdot F_0(t), \qquad n = 1, 2, \cdots, N_{F_0}. \tag{12}$$

□

[Constraint3] (Common onset and offset) Suppose that T_{S} and T_{E} are the onset and offset of the fundamental component. If the signal component obtained by the k-th channel is the signal component generated by the same acoustic source (that is, harmonic components), then onset $T_{k,\mathrm{on}}$ and offset $T_{k,\mathrm{off}}$ determined by the k-th channel must coincide with T_{S} and T_{E} respectively. That is, the differences in onset and offset must satisfy

$$|T_{\mathrm{S}} - T_{k,\mathrm{on}}| \leq \Delta T_{\mathrm{S}} \tag{13}$$

and

$$|T_{\mathrm{E}} - T_{k,\mathrm{off}}| \leq \Delta T_{\mathrm{E}}, \tag{14}$$

respectively. □

[Constraint4] (Gradualness of change (smoothness)) Suppose that the amplitude envelope $A_k(t)$ is defined in the closed-duration $[t_a, t_b]$ and satisfies constraint 1. If $A_k(t)$ is as smooth as possible, then the following integral must be minimized:

$$\sigma = \int_{t_a}^{t_b} [A_k^{(R+1)}(t)]^2 dt, \tag{15}$$

where $A_k^{(R+1)}(t)$ is determined by $C_{k,R}(t)$ in constraint 1. □

[Constraint5] (Correlation between the instantaneous amplitudes $A_k(t)$) The normalized amplitude envelope of the output of the k-th channel must approximate that of the ℓ-th channel as follows:

$$\frac{A_k(t)}{\|A_k(t)\|} \approx \frac{A_\ell(t)}{\|A_\ell(t)\|}, \qquad k \neq \ell, \tag{16}$$

where $\|\cdot\|$ is the norm symbol. The norm of $A_k(t)$, determined by $\|A_k(t)\|$, is determined as $\|A_k(t)\| = \sqrt{\int_0^t |A_k(\tau)|^2 d\tau}$. □

Substituting constraint (9) in Eq. (7), we get the linear differential equation of the instantaneous input phase difference $\theta_k(t)$. By solving this linear differential equation, we can determine $\theta_k(t)$ as follows.

[Lemma1] From constraint 1, a general solution of the input phase $\theta_k(t)$ is determined by

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{S_k(t)\cos(\phi_k(t) - \theta_{1k}(t)) + C_k(t)}\right), \tag{17}$$

where $C_k(t) = -\int C_{k,R}(t)dt - C_{k,0} = -A_k(t)$. The $C_k(t)$ is called the "undetermined function".

(Proof) See appendix A. □

From Lemma 1, if $C_k(t)$ is determined, then $\theta_k(t)$ is uniquely determined by the above equation. Moreover, if $D_{k,R}(t)$ is determined, then the two instantaneous input phases can be determined using $\theta_k(t)$ and $D_{k,R}(t)$. Therefore, if the two R-th-order polynomials $C_{k,R}(t)$ and $D_{k,R}(t)$ are determined as some kind of optimization problem, the two instantaneous amplitudes and the two instantaneous phases can be estimated. Although it is possible to estimate the coefficients $C_{k,r}(t)$ and $D_{k,r}(t)$, $r = 0, 2, = \cdots, R$, there is a problem that the computational cost of estimating two polynomials increases greatly.

In this paper, in order to reduce the computational cost, we assumed that $C_{k,R}(t)$ is a linear ($R = 1$) polynomial ($dA_k(t)/dt = C_{k,1}(t)$) and $D_{k,R}(t)$ is zero ($d\theta_{1k}(t)/dt = D_{k,0} = 0$) in constraint 1. In this assumption, the instantaneous amplitude $A_k(t)$ which can be allowed to undergo a temporal change in region, constrains the second order polynomial ($A_k(t) = \int C_{k,1}(t)dt + C_{k,0}$). Moreover, the instantaneous phase $\theta_{1k}(t)$, which is constrained (i.e. $\theta_{1k}(t) = D_{k,0}$), cannot be allowed to temporarily change. Here, if the number of channels K is very large, each frequency of the signal component that passed through the channel approximately coincides with the center frequency of each channel. Even if the above condition is false, its frequency difference can be represented by $D_{k,0}$.

This paper solves the problem of segregating the desired signal $f_1(t)$ from the mixed signal, in which noise $f_2(t)$ is added to the localized $f_1(t)$, under the above assumption.

2.3 Overview of the proposed model

An overview of signal flow in the proposed model is shown in Fig. 3.

First, this model can observe the mixed signal $f(t)$ (Fig. 3. A). The observed signal is decomposed into an instantaneous amplitude $S_k(t)$ and an instantaneous phase $\phi_k(t)$ using

an auditory-motivated filterbank (Fig. 3. B, C). Next, the fundamental frequency $F_0(t)$ of the desired signal is determined using an amplitude spectrogram (or an other method) (Fig. 3. D). Next, the concurrent time-frequency region of the segregation target is determined using grouping constraints. The frequency region, in which harmonic exist, is determined using the fundamental frequency $F_0(t)$ and the constraint of harmonicity (Fig. 3. E, and focus on a-a') . The time region in which harmonics exist is determined using the constraint of common onset and offset (Fig. 3 F, and focus on b-b').

Next, in the determined concurrent time-frequency region, the instantaneous amplitude $A_k(t)$ and $\theta_{1k}(t)$ of the desired signal $f_1(t)$ are determined from the instantaneous amplitude $S_k(t)$ and the instantaneous output phase $\phi_k(t)$. These are determined by estimating $C_{k,1}(t)$ and $D_{k,0}$, which are constraining $A_k(t)$ and $\theta_{1k}(t)$, respectively (Fig. 3. H, I). In particular, $C_{k,1}(t)$ is estimated using the constraints of smoothness and correlation between the instantaneous amplitudes $A_k(t)$. And $D_{k,0}$ is estimated using the estimated $C_{k,1}(t)$ and constraint of correlation between the instantaneous amplitudes $A_k(t)$.

Finally, the segregated signal is reconstructed the grouping the instantaneous amplitude $A_k(t)$ and the instantaneous input phase $\theta_{1k}(t)$ (Fig. 3. J).

Fig. 3

3 Algorithm implementation

3.1 An auditory-motivated filterbank

In this investigation, a filterbank is implemented considering two points: (1) consideration of the properties of the auditory system and, (2) detection of a discontinuous point dealing with the complex spectrum. In order to construct a constant Q filterbank, this paper uses the gammatone filter as an analyzing wavelet.

The gammatone filter, which is an auditory filter designed by Patterson [Patterson *et al.*, 1995], simulates the response of the basilar membrane. Its impulse response is given by

$$gt(t) = At^{N-1}\exp(-2\pi b_f \mathrm{ERB}(f_0)t)\cos(2\pi f_0 t), \qquad t \geq 0, \tag{18}$$

where A, b_f, and N are parameters, and $At^{N-1}\exp(-2\pi b_f \mathrm{ERB}(f_0)t)$ is the amplitude term represented by the Gamma distribution, f_0 is the center frequency, and $\mathrm{ERB}(f_0)$ is an equivalent rectangular bandwidth in $f_0(t)$. In addition, amplitude characteristics of the

gammatone filter are represented approximately by

$$GT(f) \approx \left[1 + \frac{j(f - f_0)}{b_f \mathrm{ERB}(f_0)}\right]^{-N}, \qquad 0 < f < \infty, \tag{19}$$

where $GT(f)$ is the Fourier transform of $gt(t)$.

In order to determine phase information, let the analyzing wavelet be the extend gammatone filter in Eq. (18) using the Hilbert transform. This analyzing wavelet is represented by

$$\psi(t) = At^{N-1} \exp(j2\pi f_0 t - 2\pi b_f \mathrm{ERB}(f_0)t), \tag{20}$$

where $f_0 = 600$ Hz, $N = 4$, and $b_f = 0.25$. Here, the bandwidth of the $\psi(t)$ becomes a quarter of the bandwidth of the auditory filter (about 1/4 ERB). The characteristics of the analyzing wavelet $\psi(t)$ of Eq. (20) are shown in Fig. 4 .

Next, the wavelet filterbank is designed using the wavelet transform (see appendix B). Here, let the wavelet transform of $f(t)$ be

$$\tilde{f}(a, b) = |\tilde{f}(a, b)| \exp(j \arg(\tilde{f}(a, b))), \tag{21}$$

where $|\tilde{f}(a, b)|$ is the amplitude spectrum and $\arg(\tilde{f}(a, b))$ is the phase spectrum; a is the scale parameter and b is the shift parameter.

Finally, an auditory-motivated filterbank is designed with the conditions shown in Table. 3. The frequency characteristics of the wavelet filterbank are shown in Fig. 5.

Table 3

Figs. 4 and 5

3.1.1 Calculation of instantaneous amplitude $S_k(t)$ and instantaneous output phase $\phi_k(t)$

The instantaneous amplitude $S_k(t)$ of Eq. (5) and the instantaneous output phase $\phi_k(t)$ of Eq. (6) can be calculated using the following lemma.

[Lemma2] The instantaneous amplitude $S_k(t)$ is calculated by

$$S_k(t) = |\tilde{f}(\alpha^{k-\frac{K}{2}}, t)|, \qquad a = \alpha^{k-\frac{K}{2}}, b = t, \tag{22}$$

where $|\tilde{f}(a, b)|$ is the amplitude spectrum defined by the wavelet transform. The instantaneous output phase $\phi_k(t)$ is calculated by

$$\phi_k(t) = \int_0^t \left(\frac{d}{d\tau} \arg\left(\tilde{f}(\alpha^{k-\frac{K}{2}}, \tau)\right) - \omega_k\right) d\tau, a = \alpha^{k-\frac{K}{2}}, b = t, \tag{23}$$

where $\arg(\tilde{f}(a, b))$ is the phase spectrum defined by the complex wavelet transform.

Proof. See appendix C. □

3.2 Grouping block

This section describes the method of estimating the fundamental frequency, constraint 2 of harmonicity, and constraint 3 of common onset and offset, in order to constraint the time-frequency region in which the desired signal exists.

3.2.1 Determination of the fundamental frequency

In this paper, the fundamental frequency of the complex tones is estimated using TEMPO (a method of time-domain excitation extraction based on a minimum perturbation operator) [Kawahara, 1997].

Next, consider the constraint of gradualness of change in Eq. (11) for the estimated fundamental frequency $F_0(t)$. The estimated $F_0(t)$ can take continuous values. However, since the number of channels in the auditory-motivated filterbank is finite, the center frequencies of the auditory-motivated filterbank cannot take continuous values. Therefore, it is difficult to deal with continuous temporal variation of $F_0(t)$. In this paper, we assume that $E_{0,R}(t) = 0$ in Eq.(11) for a small segment, means that $F_0(t)$ is constant in a small segment. Here, the above small segment is determined using the following equation, as the duration for which the temporal variation of $F_0(t)$ has the same variance as $F_0(t)$.

$$\frac{1}{T_h - T_{h-1}} \int_{T_{h-1}}^{T_h} \left| F_0(t) - \overline{F_0(t)} \right|^2 dt \leq (\Delta F_0)^2, \tag{24}$$

where the length of the small segment is $T_h - T_{h-1}$ and $\Delta(F_0)^2$ is the variance of $F_0(t)$. In this paper, $\Delta F_0 = 1$ Hz.

The relationship between $F_0(t)$ and the small segments using constraint 1 is shown in Fig. 6. For $F_0(t)$, as shown by the dotted line in Fig. 6, segregated duration ($F_0(t)$ duration) is applied to small segments from Eq. (24). The number of split segments is $H - 1$.

Fig. 6

3.2.2 Grouping constraints

For the fundamental frequency $F_0(t)$ in each small segment, the constraint of harmonicity, and the constraint of common onset and offset are implemented as follows.

First, from constraint 2, 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

$$\ell = \frac{K}{2} - \left\lceil \frac{\log(n \cdot F_0(t)/f_0)}{\log \alpha} \right\rceil, \quad n = 1, 2, \cdots, N_{F_0}, \tag{25}$$

where α is the scale parameter shown in Table 2 and $\lceil \cdot \rceil$ is the ceil symbol, meaning the approximation of the closest integer value toward positive infinity.

Next, from constraint 3, let the onset and offset of the fundamental component, $T_{\rm S}$ and $T_{\rm E}$, be $T_{\rm S} = T_{h-1}$ and $T_{\rm E} = T_h$, respectively. Moreover, we assume that $\Delta T_{\rm S} = 50$ msec and $\Delta T_{\rm E} = 100$ msec. Here $\Delta T_{\rm S}$ is taken from the result of a psychoacoustical experiment on the synchronism of onset [Kashino and Tanaka, 1994].

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 $|\frac{d\phi_k(t)}{dt}|$ (within 25 ms) to the maximum point of $|\frac{dS_k(t)}{dt}|$.

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

Here, constraint 2 acts on the signal component of $f_1(t)$ in the log-frequency domain, and constraint 3 acts the signal component of $f_1(t)$ in the time domain.

3.3 Separation block

In the separation block, the instantaneous amplitude $A_k(t)$ and the instantaneous input phase $\theta_{1k}(t)$ are estimated from $S_k(t)$ and $\phi_k(t)$ for the concurrent time-frequency region. This is done by the following steps, which optimize $C_{k,1}(t)$ and $D_{k,0}$.

Step. 1 Let $D_{k,0}$ in Eq. (10) be any value within $-\pi/2 \le D_{k,0} \le \pi/2$.

Step. 2 Using the Kalman filter, determined the estimated region, which is $\hat{C}_{k,0}(t) - P_k(t) \le C_{k,1}(t) \le \hat{C}_{k,0}(t) + P_k(t)$, where $\hat{C}_{k,0}(t)$ is the estimated value and $P_k(t)$ is the estimated error.

Step. 3 Select candidates of $C_{k,1}(t)$ using the spline interpolation in the estimated error region $\hat{C}_{k,0} - P_k(t) \le C_{k,1}(t) \le \hat{C}_{k,0}(t) + P_k(t)$.

Step. 4 Determine $C_{k,1}(t)$ using the correlation between the instantaneous amplitudes $A_k(t)$.

Step. 5 Repeating Steps.1 to 4, determine $D_{k,0}$ using the correlation between the instantaneous amplitudes $A_k(t)$

Step. 6 Determine $\theta_k(t)$ from $\hat{C}_{k,1}(t)$ and determine $\theta_{1k}(t)$ from $\hat{D}_{k,0}$. Then, determine $\theta_{2k}(t) = \theta_k(t) + \theta_{1k}(t)$.

Step. 7 Determine $A_k(t)$ and $B_k(t)$ from Eqs. (7) and (8), respectively.

3.3.1 Determination for the estimated region using the Kalman filter

In this section, we consider how to estimate $C_{k,0}$ 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 $\Delta t = 1/f_s$, where f_s is the sampling frequency. Here, let the temporal variation of $C_{k,0}(t)$ at discrete time t_m be

$$C_{k,0}(t_{m+1}) = C_{k,0}(t_m)\Delta C_k(t_m) + w_m, \tag{26}$$

$$\Delta C_k(t_m) = 1 + \frac{C_{k,0}(t_m) - C_{k,0}(t_{m-1})}{C_{k,0}(t_m)}, \tag{27}$$

where $t_0 = T_{h-1}$ and $t_M = t_h$. It is assumed that $C_{k,0}(t_m)$ times $\Delta C_k(t_m)$, and that the variation error is represented by white noise with mean 0 and variance σ_m.

Next, for the system of the Kalman filtering problem:

$$\mathbf{x}_{m+1} = \mathbf{F}_m\mathbf{x}_m + \mathbf{G}_m\mathbf{w}_m \qquad (\textbf{state}) \tag{28}$$

$$\mathbf{y}_m = \mathbf{H}_m\mathbf{x}_m + \mathbf{v}_m \quad (\textbf{observation}), \tag{29}$$

applying Eq. (26) to Eq. (28) and applying Eq. (2) to Eq. (29). The parameters in Eqs. (28) and (29) are shown in Table 4.

Next, performing the Kalman filtering (see Appendix D) according to Eqs. (28) and (29), we obtain the minimal-variance estimated value $|\hat{\mathbf{x}}_{m|m}|$ and covariance matrix $|\hat{\mathbf{\Sigma}}_{m|m}|$ at discrete time t_m. As a result, the estimated $\hat{C}_{k,0}(t)$ and the estimated error $P_k(t)$ are determined by $\hat{C}_k(t) = -|\hat{\mathbf{x}}_{m|m}|$ and $P_k(t) = |\hat{\mathbf{\Sigma}}_{m|m}|$, respectively. Therefore, the estimated error region for $C_{k,1}(t)$ is

$$\hat{C}_{k,0}(t) - P_k(t) \leq C_{k,1}(t) \leq \hat{C}_{k,0}(t) + P_k(t). \tag{30}$$

Table 4

3.3.2 Candidate selection of $C_{k,1}(t)$ using the spline interpolation

In order to determined whether $A_k(t)$ satisfied Eq. (15), consider the selection of candidates for $A_k(t)$. Suppose that $A_k^{(R+1)}(t)$ is the instantaneous amplitude of $f_1(t)$ given by any $C_{k,R}(t)$, and $\tau_1, \tau_2, \cdots, \tau_i$ are within the opened-duration (t_a, t_b), where $t_a < \tau_1 < \cdots < \tau_i < t_b$. In addition, suppose that $A_{k,i} := A_k^{(R+1)}(t_i)$ is the value of the instantaneous amplitude at time τ_i.

To estimate $C_{k,1}(t)$, where $R = 1$, that satisfies Eq. (15) we interpolate $A_k^{(R+1)}(t)$, where $R = 1$ and, $A_k^{(R+1)}(t) = A_{k,i}$, $i = 1, 2, \cdots, I$ in $[t_a, t_b]$. According to constraint 4, the smoothest interpolation function is the $(2R+1)$th-order spline function. This spline function is unique [de Boor, 1978].

As shown in Fig. 7, first we determine candidates of $C_{k,1}(t)$ using the spline function within the estimated error region: $\hat{C}_{k,0}(t) - P_k(t) \leq C_{k,1}(t) \leq \hat{C}_{k,0}(t) + P_k(t)$. Then, select a correct solution from the candidates of $C_{k,1}(t)$, we can uniquely determine the smoothest $A_k(t)$ from $C_{k,1}(t)$.

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

Fig. 7

3.3.3 Determination of $C_{k,1}(t)$ using correlation between the instantaneous amplitudes

Select an optimal solution from the candidates of $C_{k,1}(t)$ using Eq. (16) in constraint 5. This process is done by selecting $C_{k,1}(t)$ when the correlation between the instantaneous amplitude ($A_k(t)$ and $A_\ell(t)$) is maximum at any $C_{k,1}(t)$ within the estimated error region.

$$\hat{C}_{k,1} = \underset{\hat{C}_{k,0} - P_k \leq C_{k,1} \leq \hat{C}_{k,0} + P_k}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||}, \tag{31}$$

where $< \cdot >$ is an operation of the inner product, $\hat{A}_k(t)$ is the instantaneous amplitude obtained by interpolating $C_k(t)$, and $\hat{\hat{A}}_k(t)$ is the instantaneous amplitude determined as

$$\hat{\hat{A}}_k(t) = \frac{1}{N_{F_0}} \sum_{\ell \in \mathbf{L}} \frac{\hat{A}_\ell(t)}{\|\hat{A}_\ell(t)\|}, \tag{32}$$

where, L is the set symbol of ℓ that satisfies Eq. (12).

3.3.4 Determination of $D_{k,0}(t)$ using correlation between the instantaneous amplitudes

Setting $D_{k,0}$ to any value within $[-\pi/2, \pi/2]$, we can determine $\hat{C}_{k,1}(t)$ at any $D_{k,0}$ using the above method. An optimal solution $\hat{D}_{k,0}$ is determined by

$$\hat{D}_{k,0} = \underset{-\pi/2 \le D_{k,0} \le \pi/2}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||} \tag{33}$$

where $\hat{A}_k(t)$ is the instantaneous amplitude determined by $\hat{C}_{k,1}$, and $\hat{\hat{A}}_k(t)$ is determined by Eq. (32).

4 Simulations

We carried out three simulations on segregating two acoustic sources using noise-added signal $f(t)$, to show that the proposed method can extract the desired signal $f_1(t)$ from the mixed signal $f(t)$. These simulations were composed as follows:

1. Extracting an AM complex tone from a noise-added AM complex tone.
2. Extracting one AM complex tone from mixed AM complex tones.
3. Extracting a speech signal (vowel) from a noisy speech.

In simulations 1 and 2 the fundamental frequency did not vary temporally, while in simulation 3 it did. The purpose of simulation 1 was to examine the assumptions of the problem of segregating two acoustic sources; the purpose of simulation 2 was to examine the case in which the concurrent signal component exists in the same frequency region; and the purpose of simulation 3 was to examine whether the proposed method can be applied the problem of segregating a vowel from a noisy vowel.

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

One was the temporal average of the segregated error in terms of the instantaneous amplitude $A_k(t)$. The aim of using this measure was to evaluate the segregation in terms of the amplitude envelope where signal and noise exist in the same frequency region. This measure is called "Precision", and is defined by

$$\frac{1}{T} \int_0^T \left(10 \log_{10} \frac{\sum_{k=1}^K A_k(t)^2}{\sum_{k=1}^K (A_k(t) - \hat{A}_k(t))^2} \right) dt \tag{34}$$

where $A_k(t)$ is the amplitude envelope of the original signal $f_1(t)$, and $\hat{A}_k(t)$ is the amplitude envelope of the segregated signal $\hat{f}_1(t)$.

The other measure was the spectrum distortion. The aim of using this measure was to evaluate the extraction of the desired signal $\hat{f}_1(t)$ from noise-added signal $f(t)$. This measure is defined by

$$\sqrt{\frac{1}{W}\sum_{\omega}^{W}\left(20\log_{10}\frac{\tilde{F}_1(\omega)}{\tilde{\hat{F}}_1(\omega)}\right)^2}, \tag{35}$$

where $\tilde{F}_1(\omega)$ and $\tilde{\hat{F}}_1(\omega)$ are the amplitude spectra of $f_1(t)$ and $\hat{f}_1(t)$, respectively. In the above equation, the frame length is 51.2 ms, the frame shift is 25.6 ms, W is the analyzable bandwidth of the filterbank (about 6 kHz), and the window function is Hamming.

Here, in the measure of precision, a higher value means high accuracy of segregation. Conversely, in the measure of spectrum distortion, a lower value means high accuracy of segregation.

Next, in order to show the advantages of the constraints shown in Table 2, we compare the following conditions for the three simulations:

1. Condition 1: Extract the harmonics using the Comb filter and predict $A_k(t)$ using the Kalman filtering.

2. Condition 2: Extract the harmonics using the Comb filter.

3. Condition 3: Do nothing.

Here, condition 1 corresponds to constraint of gradualness of change (smoothness) being omitted; condition 2 corresponds to constraints of gradualness of change (smoothness) and harmonicity being omitted; and condition 3 corresponds to no constraints being applied at all.

The evaluated value of the noise reduction is the improvement in the two measures between using the proposed method and using condition 3.

4.1 Simulation 1

This simulation assumed that $f_1(t)$ was an AM complex tone as shown in Fig. 8, where $F_0(t) = 200$ Hz, $N_{F_0} = 10$, and the tone's instantaneous amplitude was sinusoidal (10 Hz), and that $f_2(t)$ was bandpassed pink noise, having a bandwidth of about 6 kHz. Five types of $f(t)$ were used as simulation stimuli, where the SNRs of $f(t)$ ranged from 0 to 20 dB in 5-dB steps. The SNR shows the ratio of signal to noise in the concurrent time region.

For example, when the SNR of $f(t)$ was 10 dB, as shown in Fig. 9, the proposed method could segregate $A_k(t)$ with high accuracy and could extract $\hat{f}_1(t)$, shown in Fig. 10, from it. In this case, the precision for $A_k(t)$ is shown in Fig. 11 (top panel). In addition, the average SDs of $\hat{f}_1(t)$ and $f(t)$ for five simulations are shown in Fig. 11 (bottom panel). It was possible to improve the precision by about 7.3 dB and the spectrum distortion by about 17.6 dB as noise reduction, comparing the proposed method with condition 3. Hence, the proposed model could extract with high precision the amplitude information of signal $f_1(t)$ from a noise-added signal $f(t)$ in which the signal and noise were in the same frequency region.

Figs. 8–11

4.2 Simulation 2

This simulation assumed that $f_1(t)$ was an AM complex tone the same as Fig. 8 and that $f_2(t)$ was another AM complex tone, where $F_0(t) = 300$ Hz, $N_{F_0} = 10$, and the tone's instantaneous amplitude was sinusoidal (15 Hz). Therefore, harmonics of $f_1(t)$ and $f_2(t)$ in multiples of 600 Hz (for example, the third harmonic of $f_1(t)$ and second harmonic of $f_2(t)$) exist in the same frequency region. Five types of $f(t)$ were used as simulation stimuli, where the SNRs of $f(t)$ ranged from 0 to 20 dB in 5-dB steps.

For example, when the SNR of $f(t)$ was 10 dB, as shown in Fig. 12, the proposed method could segregate $A_k(t)$ with high accuracy and could extract $\hat{f}_1(t)$, shown in Fig. 13, from the $f(t)$, even when two components of the signals existed in the same frequency region. In this case, the precision for $A_k(t)$ is shown in Fig. 14 (top panel). In addition, the average SDs of $\hat{f}_1(t)$ and $f(t)$ for five simulations are shown in Fig. 14 (bottom panel). It was possible to improve the precision by about 3.1 dB and the spectrum distortion by about 7.2 dB as noise reduction, comparing the proposed method with condition 3.

Hence, just like the results of the previous simulations, the proposed model could also extract with high precision the amplitude information of signal $f_1(t)$ from a noise-added signal $f(t)$ in which two AM complex tones existed in the same frequency region. Here, comparing the results of the proposed method with condition 2, we see that the segregated accuracy using the proposed method shoed more improvement with simulation 2.

Figs. 12–14

4.3 Simulation 3

This simulation assumed that $f_1(t)$ was a vowel /a/ synthesized by the log magnitude approximation (LMA) [Imai *et al.*, 1977 ,Imai, 1978] as shown in Fig. 15, where averaged $\overline{F_0(t)} = 125$ Hz, jitter was 5 Hz (from 123 to 128 Hz), and $f_2(t)$ was bandpassed pink noise, where the bandwidth was about 6 kHz. In this simulation, $N_{F_0} = 40$. Five types of $f(t)$ were used as simulation stimuli, where the SNRs of $f(t)$ ranged from 0 to 20 dB in 5-dB steps.

For example, when the SNR of $f(t)$ was 10 dB as shown in Fig. 16, the proposed method could segregate $A_k(t)$ with high accuracy and could extract $\hat{f}_1(t)$, shown in Fig. 17, from $f(t)$. In this case, the precision for $A_k(t)$ is shown in Fig. 18 (top panel). In addition, the average SDs of $\hat{f}_1(t)$ and $f(t)$ for five simulations are shown in Fig. 18 (bottom panel). It was possible to improve the precision by about 4.8 dB and the spectrum distortion by about 7.3 dB as noise reduction, comparing the proposed method with condition 3. Here, comparing the amplitude spectrum of original signal $f_1(t)$ with that of $\hat{f}_1(t)$ or $f(t)$, the proposed method could clearly reduce the noise-component from the observed amplitude spectrum, as shown in Fig. 19. In this figure, the amplitude spectra are shown in one frame in middle point of signal duration. Hence, the proposed model could also extract with high accuracy the amplitude information of speech $f_1(t)$ from a noisy speech $f(t)$ in which speech and noise existed in the same frequency region. Hence, this method can be applied in cases where a speech signal is to be extracted from noisy speech.

Figs. 15–18

4.4 Comparison the proposed model with the other method

We compared the proposed method and the other method (under three conditions) for the above three simulations. As shown by the results of Figs. 11, 14, and 18, the segregation accuracy using the proposed method was better than the other three conditions. Comparing the proposed method and condition 1 shows the advantage of the constraint of gradualness of change (smoothness). Comparing the proposed method and condition 2 shows the advantage of constraining the gradualness of change (smoothness) and harmonicity and also shows the segregation accuracy in the same frequency region in which concurrent harmonic components exist. Comparing the proposed method and condition 3 shows the improved accuracy of the proposed method. The result for these three simulations and three conditions show that the proposed method can segregate the desired signal from a

mixed signal, with about 4 dB better precision and about 10 dB improvement in spectrum distortion.

As the above results, the problem of segregating the desired vowel from noisy vowel, using constraints related to Bregman's regularities is examined in this section.

5 Conclusions

This paper proposed a method of extracting the desired signal from a noisy signal, addressing the problem of segregating two acoustic sources as a model of acoustic source segregation based on Auditory Scene Analysis. The problem of segregating two acoustic sources, discussed here, is an ill inverse problem in which the instantaneous amplitude and the instantaneous phase of the desired signal must be determined using the instantaneous amplitude and the instantaneous phase of the observed signal. This method uses the instantaneous amplitude and the instantaneous phase of signal component passed through a wavelet filterbank. It can solve this problem using constraints, which are related to the four heuristic regularities proposed by Bregman.

As an example of segregation using the proposed method, we demonstrated three simulations of segregating two acoustic sources. These simulations were:

1. Extracting an AM complex tone from a noise-added AM complex tone.

2. Extracting one AM complex tone from mixed AM complex tones.

3. Extracting a speech signal from a noisy speech.

In these simulations, two measures were used to evaluate the proposed method. One was precision, which is the temporal average of the segregated error for $A_k(t)$ and the other was the spectrum distortion for the extracted signal. Moreover, the segregation accuracy between using the proposed model and using the other method, where it uses the omitted constraints in the proposed model, were evaluated by computer simulations.

The results of simulations 1 and 2 showed that the proposed method could extract with high precision the AM complex tone not only from a noise-added AM complex tone but also from mixed AM complex tones, in which signal and noise existed in the same frequency region. In particular, it was possible to reduce the SD by about 20 dB as noise reduction, using the proposed method. Moreover, the results of simulation 3 showed that the proposed method could also extract the speech signal from noisy speech.

Comparing the proposed method with the three conditions, we found that using it with constraints related to the four regularities is better than using the other method (under three conditions). In particular, it could segregate the desired signal from the mixed signal, with about 4 dB better precision and about 10 dB improvement in spectrum distortion.

Acknowledgments

This work was supported by Grant-in-Aid for Scientific research from the Ministry of Education (Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists). It was partially supported by Grant-in-Aid for Scientific Research from the Ministry of Education (No. 07308026 and No. 10680374), and by CREST (Core Research for Evolutional Science and Technology) of the Japan Science and Technology Corporation (JST).

References

[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. and Hwang, P. Y. C. Introduction to Random Signals and Appled Kalman Filtering, second edition, Chapter 5–6, pp. 210–288, John Wiley and Sons, Inc., New York, 1992.

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

[Chui, 1992] Chui, C.K. An Introduction to Wavelets, Academic Press, Boston, MA, 1992.

[de Boor, 1978] de Boor, C. A Practical Guide to Spline. Springer-Verlag, New York, 1978.

[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.

[Imai, 1978] Imai, S. "Low bit rate cepstral vocoder using the log magnitude approximation filter," Proc. 1978 international conference on acoustics, speech, and signal process., pp. 441-444, 1978.

[Imai *et al.*, 1977] Imai, S., Kitamura, T., and Takeya, H. "A direct approximation technique of log magnitude response for digital filters," IEEE trans. on IEEE Trans. on Acoustics, speech and signal processing, ASSP-25, [2] pp. 127-133, 1977.

[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.

[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.

[Kawahara, 1997] Kawahara, H. "STRAIGHT - TEMPO: A Universal Tool to Manipulate Linguistic and Para-Linguistic Speech Information," In Proc. SMC-97, Oct. 12-15, Orlando, Florida, USA.

[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.

[Patterson *et al.*, 1995] Patterson, R. D., Allerhand, M., and Giguére, C. "Time-domain modelling of peripheral auditory processing: a modular architecture and a software platform," J. Acoust. Soc. Am. 98, 1890–1894.

[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, pp. 444-453, March 1997 (in Japanese).

[Unoki and Akagi, 1997b] 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.

Appendix A: Proof of Lemma 1

Let $\Psi_k(t) = \phi_k(t) - \theta_{1k}(t)$. Rearranging Eq. (7), we obtain

$$A_k(t) = S_k(t)\cos\Psi_k(t) - S_k(t)\cot\theta_k(t)\sin\Psi_k(t). \tag{36}$$

Differentiating both terms in t in the above equation, we obtain

$$y'(t) + \frac{P'(t)}{P(t)}y(t) = \frac{Q'(t)}{P(t)} - \frac{C_{k,R}(t)}{P(t)}, \tag{37}$$

where $y(t) = \cot\theta_k(t)$, $P(t) = S_k(t)\sin\Psi_k(t)$, and $Q(t) = S_k(t)\cos\Psi_k(t)$. Since the above equation is a linear differential equation, a general solution $y(t)$ is determined by

$$y(t) = \frac{1}{P(t)}\left(Q(t) - \int C_{k,R}(t)dt + C\right), \tag{38}$$

where C is an undermined coefficient. Hence, from

$$\cot\theta_k(t) = \frac{S_k(t)\cos\Psi_k(t) - \int C_{k,R}(t)dt + C}{S_k(t)\sin\Psi_k(t)}, \tag{39}$$

we obtain

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{S_k(t)\cos(\phi_k(t) - \theta_{1k}(t) + C_k(t)}\right), \tag{40}$$

where $C_k(t) = -\int C_{k,R}(t)dt + C$. On the other hand, applying Eq. (36) into Eq. (40), we obtain

$$\begin{aligned} A_k(t) &= S_k(t)\left(\cos\Psi_k(t) - \cos\Psi_k(t) - \frac{C_k(t)}{S_k(t)}\right) \\ &= -C_k(t). \end{aligned} \tag{41}$$

Considering that $C = -C_{k,0}$, that is the same result as integrating Eq. (9).

Appendix B: Wavelet transform

First, we summarize the wavelet transform and the inverse wavelet transform [Chui, 1992] for designing a constant Q filterbank.

The integral wavelet transform for $f(t)$ is defined by

$$\tilde{f}(a,b) = \frac{1}{\sqrt{|a|}}\int_{-\infty}^{\infty} f(t)\overline{\psi\left(\frac{t-b}{a}\right)}dt, \tag{42}$$

where a is the "scale parameter," b is the "shift parameter," $a, b \in \mathbf{R}$ with $a \neq 0$, and $\overline{\psi}$ is the conjugate of ψ. The integral basis function is $\psi(t)$ scale-transformed by parameter a and shifted by parameter b. The selection of $\psi(t)$ allows much mathematical freedom; however, in general, $\psi(t)$ is determined to be an integrable function that satisfies the following "admissibility condition":

$$G_\psi := \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|}d\omega < \infty, \tag{43}$$

where $\hat{\psi}$ is the Fourier transform of ψ, It follows that $\hat{\psi}$ is a continuous function, so that the finiteness of G_ψ in Eq. (43) implies that $\hat{\psi}(0) = 0$, or equivalently, $\int_{-\infty}^{\infty}\psi(t)dt = 0$. If the above equation is satisfied, ψ is called a "basic wavelet," and a unique inverse transform exists as follows:

$$f(t) = \frac{1}{G_\psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\tilde{f}(a,b)\psi\left(\frac{t-b}{a}\right)\frac{dadb}{a^2} \tag{44}$$

Since the analyzing wavelet $\psi(t)$ in Eq. (20) approximately satisfies the admissibility condition because $|\psi(0)| \approx 0$, it can be considered that this analyzing wavelet is the basic wavelet.

Equations. (42)–(44) are a continuous wavelet transform. The discrete wavelet transform corresponding to these equations is represented by

$$\psi_{p,q}(t) := \alpha^{-p/2}\psi\left(\frac{t - q \cdot b}{\alpha^p}\right), \tag{45}$$

$$\tilde{f}_{p,q} := \tilde{f}(\alpha^p, q/f_s) = \int_{-\infty}^{\infty} f(t)\overline{\psi}_{p,q}(t)dt, \tag{46}$$

and

$$f(t) = \frac{1}{G_{p,q}} \sum_p \sum_q \tilde{f}_{p,q}\psi_{p,q}(t), \tag{47}$$

where p and q are integer parameters.

Appendix C: Proof of Lemma 2

The wavelet transform in Eq. (21) is a complex representation of the output of the analyzing filter in Eq. (2):

$$\begin{aligned} X_k(t) &= S_k(t)e^{j(\omega_k t + \phi_k(t))} \\ &:= \tilde{f}(a, b), \qquad a = \alpha^{k-\frac{K}{2}}, b = t \end{aligned} \tag{48}$$

Taking the absolute value for both terms, we obtain

$$|X_k(t)| = S_k(t) = |\tilde{f}(\alpha^{k-\frac{K}{2}}, t)|. \tag{49}$$

Similarly, comparing phase terms between Eqs. (48) and (21), we obtain

$$\omega_k t + \phi_k(t) = \arg(\tilde{f}(a, b)). \tag{50}$$

Since the phase spectrum $\arg(\tilde{f}(a, b))$ is represented by

$$\arg(\tilde{f}(a, b)) = \arctan \frac{Im\{\tilde{f}(a, b)\}}{Re\{\tilde{f}(a, b)\}}, \tag{51}$$

it becomes a periodical ramp function within

$$-\pi \leq \arg(\tilde{f}(a, b)) \leq \pi. \tag{52}$$

Differentiating both terms in Eq. (50), we get

$$\omega_k + \frac{d\phi_k(t)}{dt} = \frac{\partial}{\partial t} \arg(\tilde{f}(\alpha^{k-\frac{K}{2}}, t)). \tag{53}$$

After clearing, we obtain

$$\frac{d\phi_k(t)}{dt} = \frac{\partial}{\partial t} \arg(\tilde{f}(\alpha^{k-\frac{K}{2}}, t)) - \omega_k. \tag{54}$$

Hence, the instantaneous output phase $\phi_k(t)$ is represented by

$$\phi_k(t) = \int \left(\frac{d}{dt} \arg\left(\tilde{f}(\alpha^{k-\frac{K}{2}}, t)\right) - \omega_k\right) dt. \tag{55}$$

Appendix D: Kalman filtering

The system of considering the Kalman filtering problem is a linear stochastic state-observation description as follows:

$$\begin{aligned} \mathbf{x}_{m+1} &= \mathbf{F}_m \mathbf{x}_m + \mathbf{G}_m \mathbf{w}_m \qquad (\textbf{state }) \\ \mathbf{y}_m &= \mathbf{H}_m \mathbf{x}_m + \mathbf{v}_m \quad (\textbf{observation}), \end{aligned}$$

where $\mathbf{x}_m$ and $\mathbf{y}_m$ are random variables, and $\mathbf{F}_m$, $\mathbf{G}_m$, and $\mathbf{H}_m$, are state transition matrix, observation matrix, and driving matrix, respectively.

In this system, mean and variance with x_0, w_m, and 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 $\hat{\mathrm{x}}_{m|m}$ from the observed y_m, $m = 0, 1, 2, \cdots, M$ as follows.

$$\hat{\mathrm{x}}_{m|m} = E(\mathrm{x}_m + \mathbf{y}_0, \cdots, \mathrm{y}_m) \tag{56}$$

The Kalman filter is called an algorithm that obtains a solution to the above problem [Brown and Hwang, 1992].

It is calculated by sequentially performing the following: [Brown and Hwang, 1992].

1. **Filtering equation**

$$\hat{\mathbf{x}}_{m|m} = \hat{\mathbf{x}}_{m|m-1} + \mathbf{K}_m(\mathbf{y}_m - \mathbf{H}_m \hat{\mathbf{x}}_{m|m-1}) \tag{57}$$

$$\hat{\mathbf{x}}_{m+1|m} = \mathbf{F}_m \hat{\mathbf{x}}_{m|m} \tag{58}$$

2. **Kalman gain**

$$\mathbf{K}_m = \frac{\hat{\mathbf{\Sigma}}_{m|m-1} \mathbf{H}_m^{*T}}{\mathbf{H}_m \hat{\mathbf{\Sigma}}_{m|m-1} \mathbf{H}_m^{*T} + \mathbf{\Sigma}_{v_m}} \tag{59}$$

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

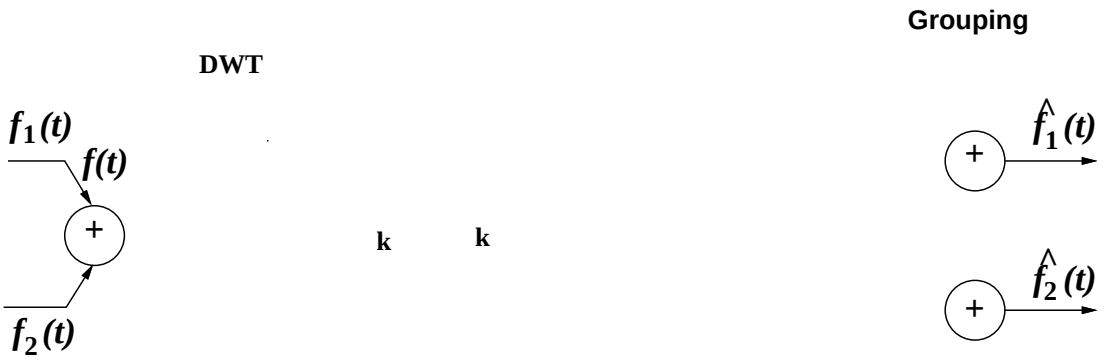
$$\hat{\mathbf{\Sigma}}_{m|m} = \hat{\mathbf{\Sigma}}_{m|m-1} - \mathbf{K}_m \mathbf{H}_m \hat{\mathbf{\Sigma}}_{m|m-1} \tag{60}$$

$$\hat{\mathbf{\Sigma}}_{m+1|m} = \hat{\mathbf{F}}_m \mathbf{\Sigma}_{m|m} \mathbf{F}_m^{*T} + \mathbf{G}_m \mathbf{\Sigma}_{w_m} \mathbf{G}_m^{*T} \tag{61}$$

4. **Initial state**

$$\hat{\mathbf{x}}_{0|-1} = \bar{\mathbf{x}}_0, \qquad \hat{\mathbf{\Sigma}}_{0|-1} = \mathbf{\Sigma}_{x_0}, \tag{62}$$

We remark that symbols $\bar{}$ and $\mathbf{\Sigma}$ are mean and variance of a random variable, respectively.



DWT : Discrete Wavelet Transform
IDWT : Inverse Discrete Wavelet Transform

Figure 1: Auditory segregation model.

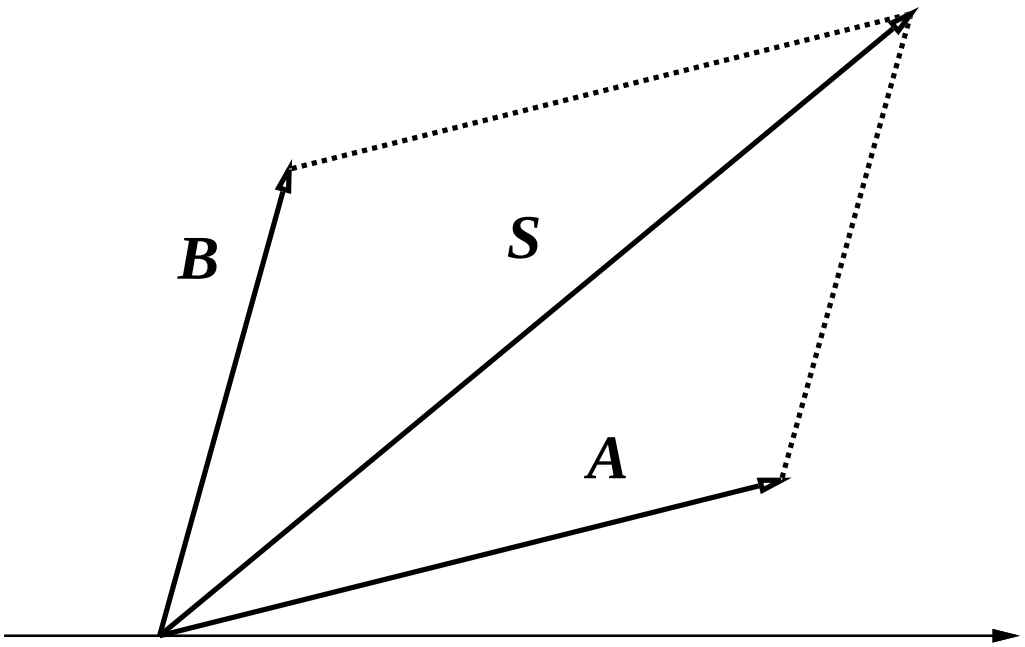

B
S
A

Figure 2: Vector-plane of complex spectrum.

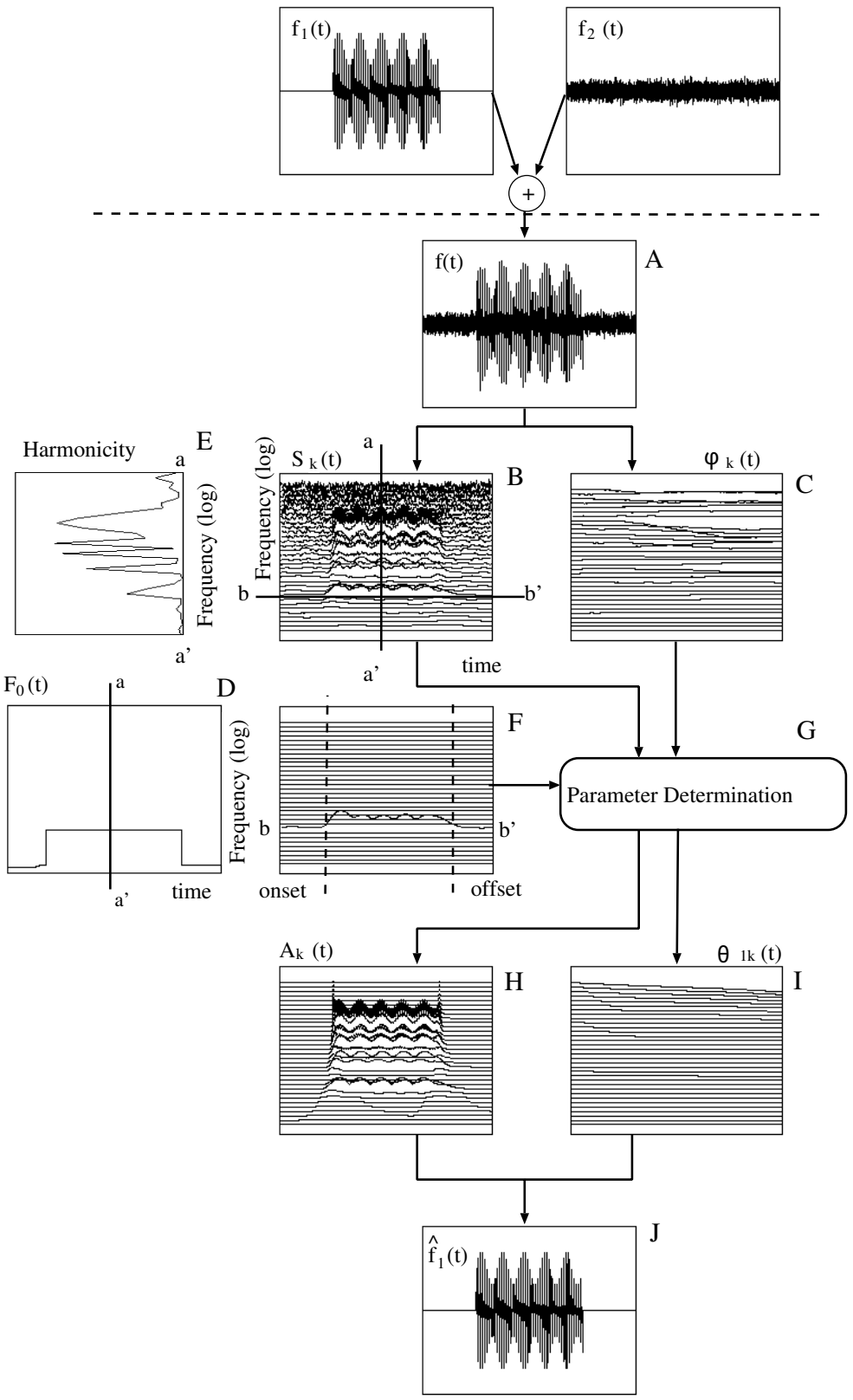


Figure 3: Overview of signal-flow in the proposed model.

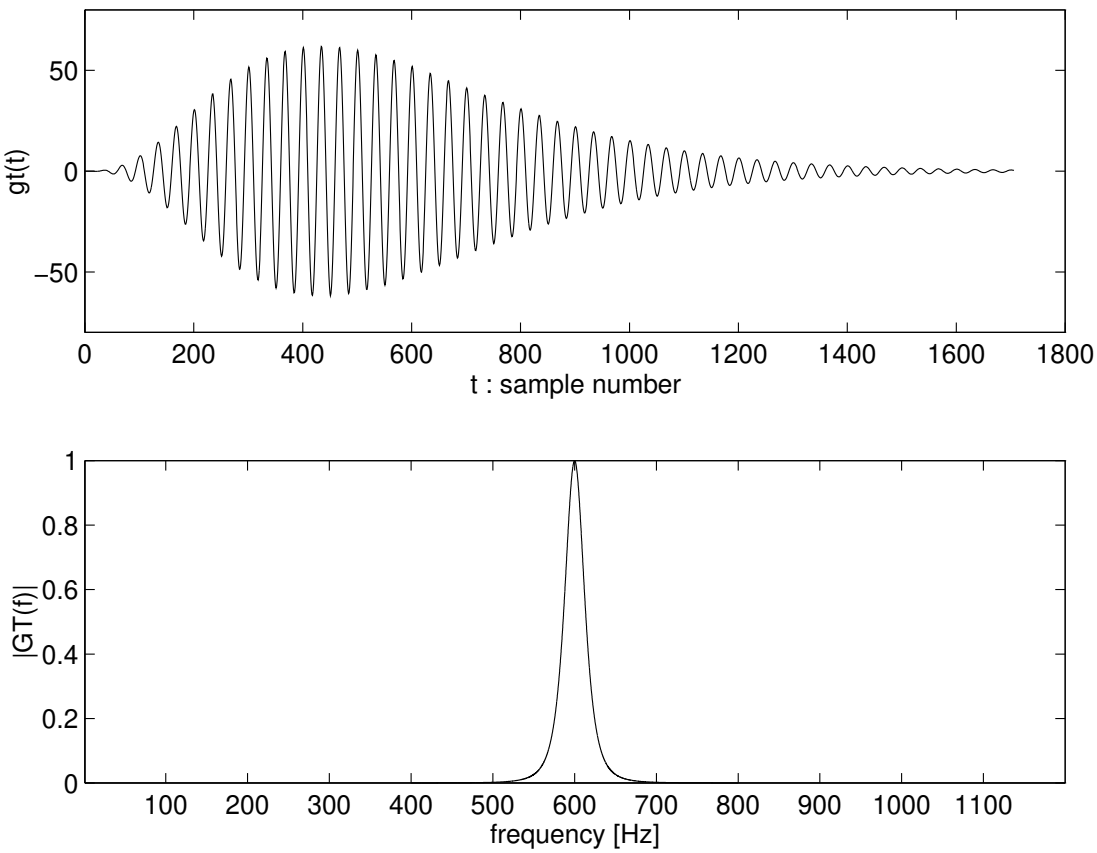


Figure 4: Characteristics of basic wavelet $\psi(t)$: (top panel) $\mathrm{Re}\{\psi(t)\}$, (bottom panel) $\hat{\psi}(f)$ (gammatone filter at $f_0 = 600$ Hz, $N = 4$, $b_f = 0.25$).

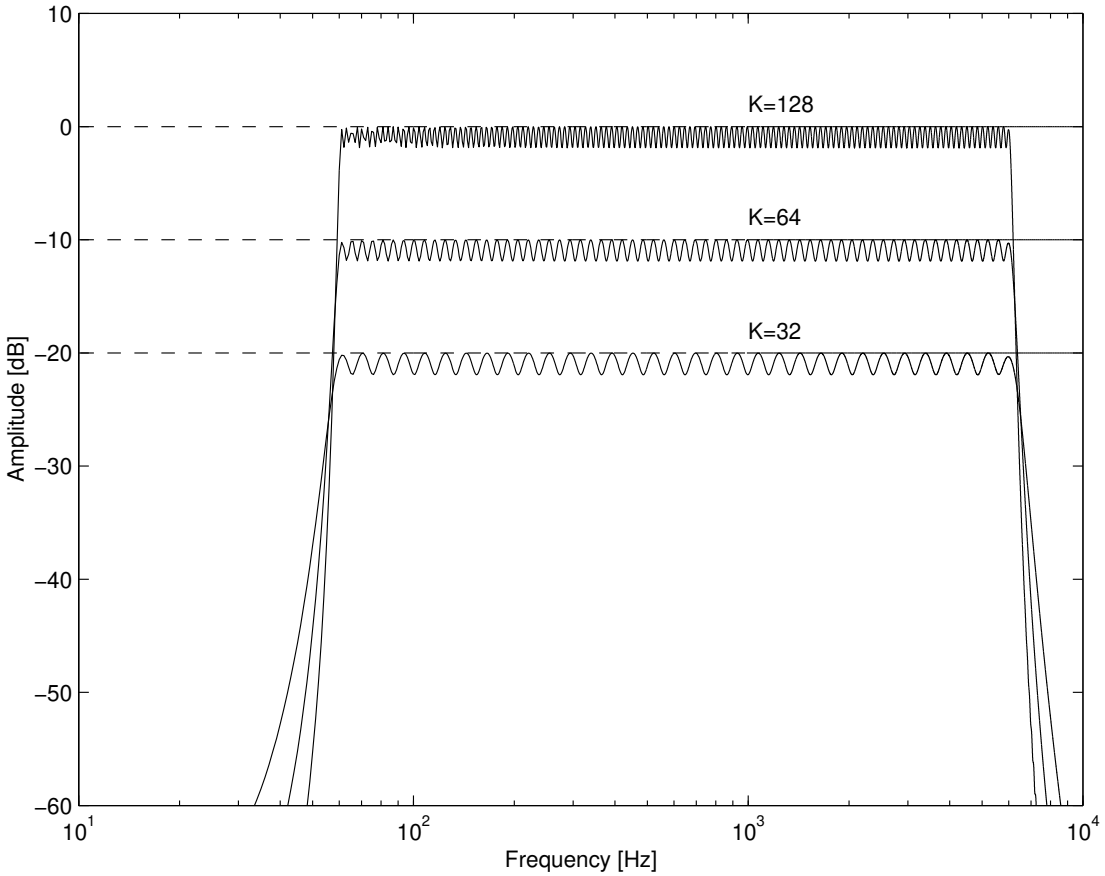


Figure 5: Frequency characteristics of the wavelet filterbank (the criterion levels are 0 dB for $K = 128$; -10 dB for $K = 64$; and -20 dB for $K = 32$.).

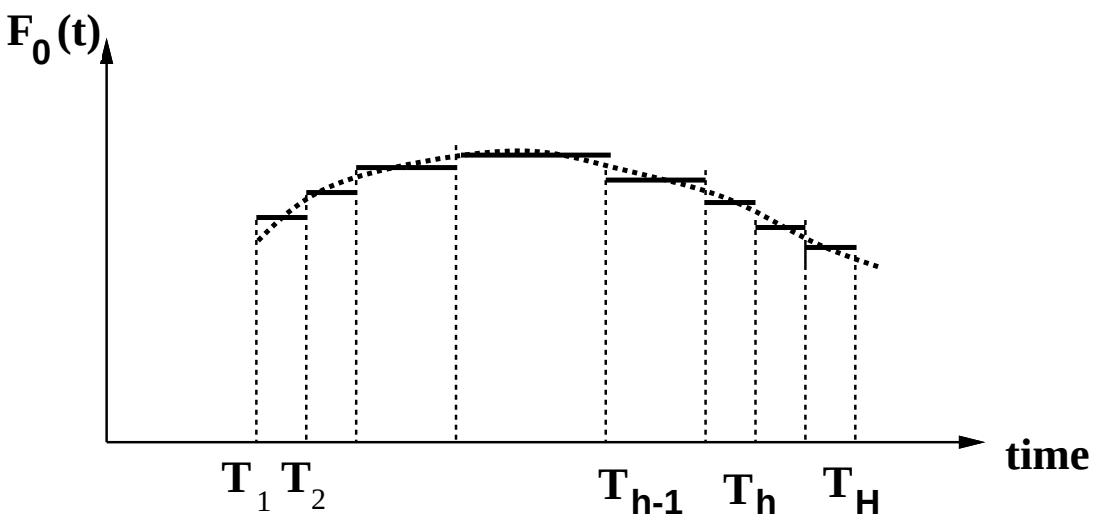


Figure 6: Temporal variation of the fundamental frequency.

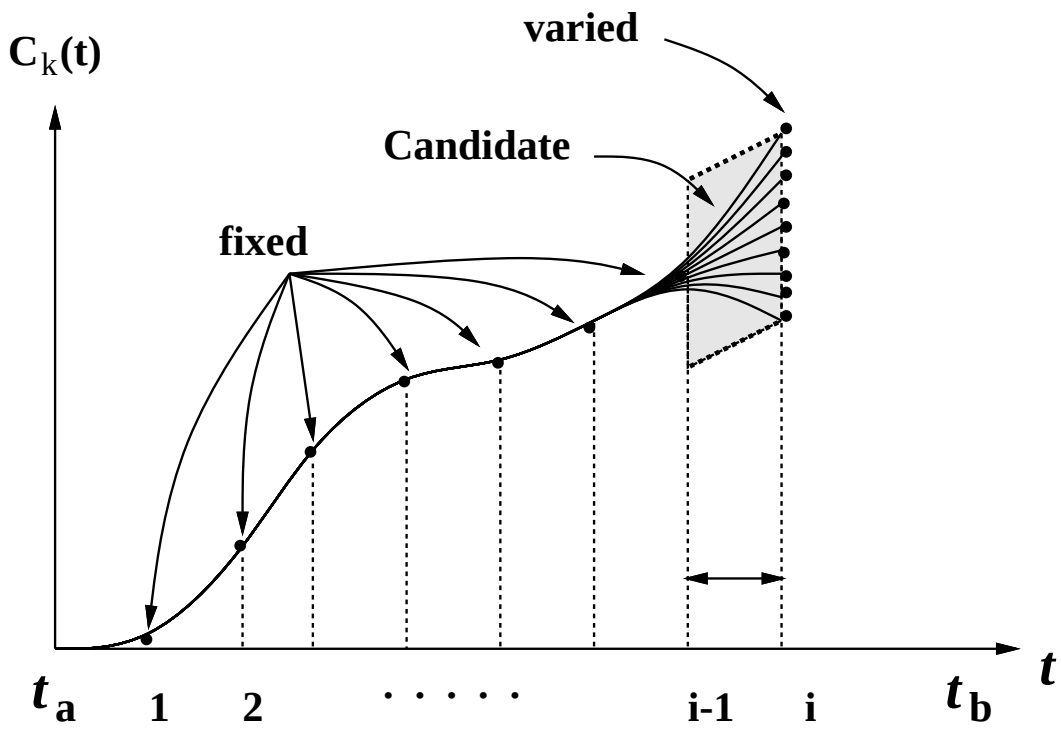

C_k(t)
varied
Candidate
fixed
t
t_a
1
2
. . . . .
i-1
i
t_b

Figure 7: Candidates for $C_k(t)$ interpolated by the spline function.

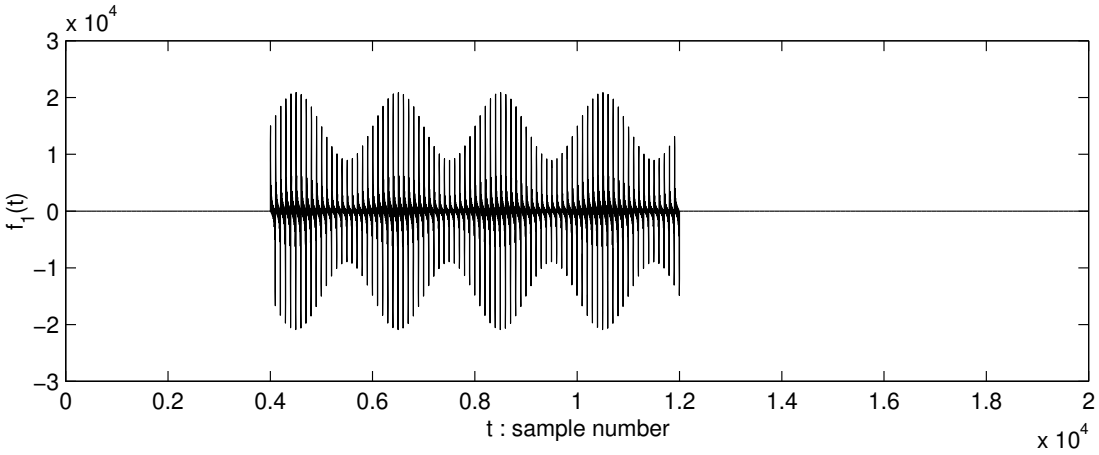


Figure 8: AM complex tone $f_1(t)$.

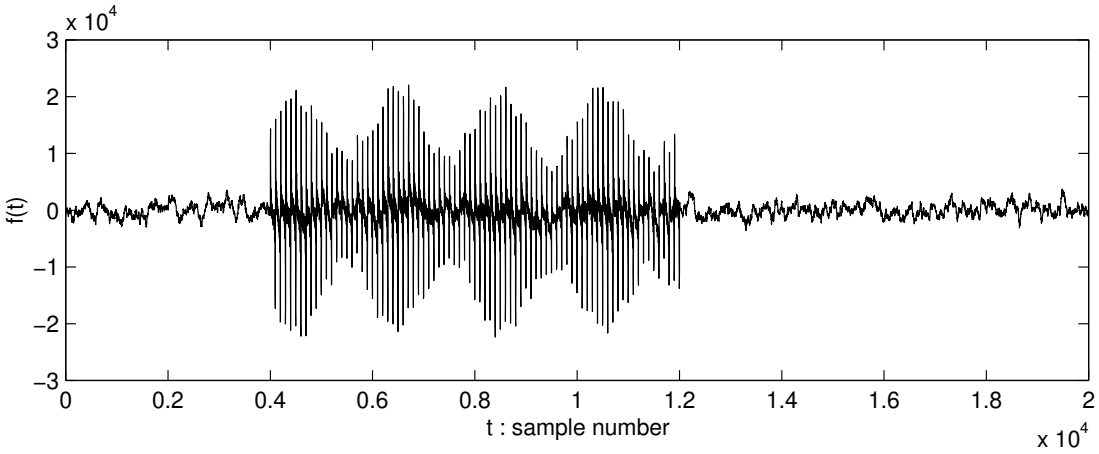


Figure 9: Mixed signals $f(t)$ (SNR= 10 dB).

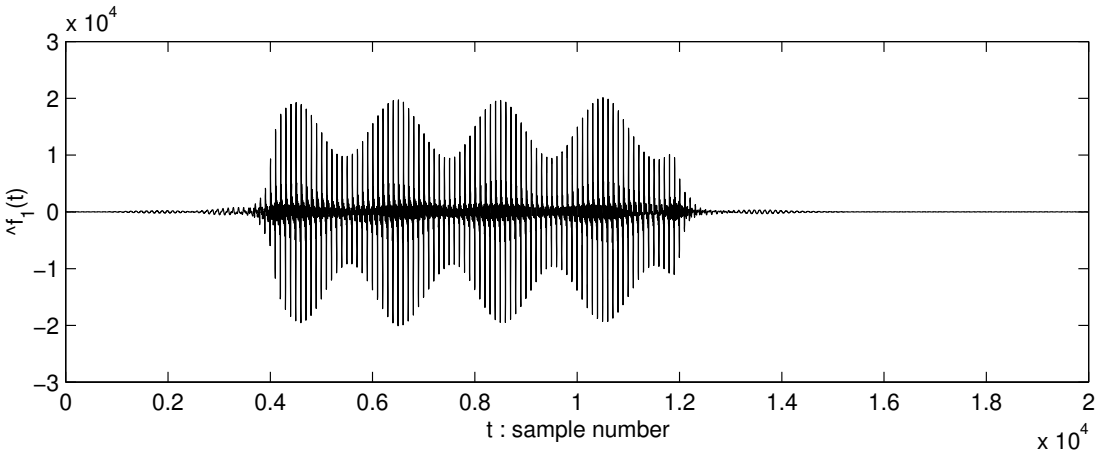


Figure 10: Extracted signal $\hat{f}_1(t)$.

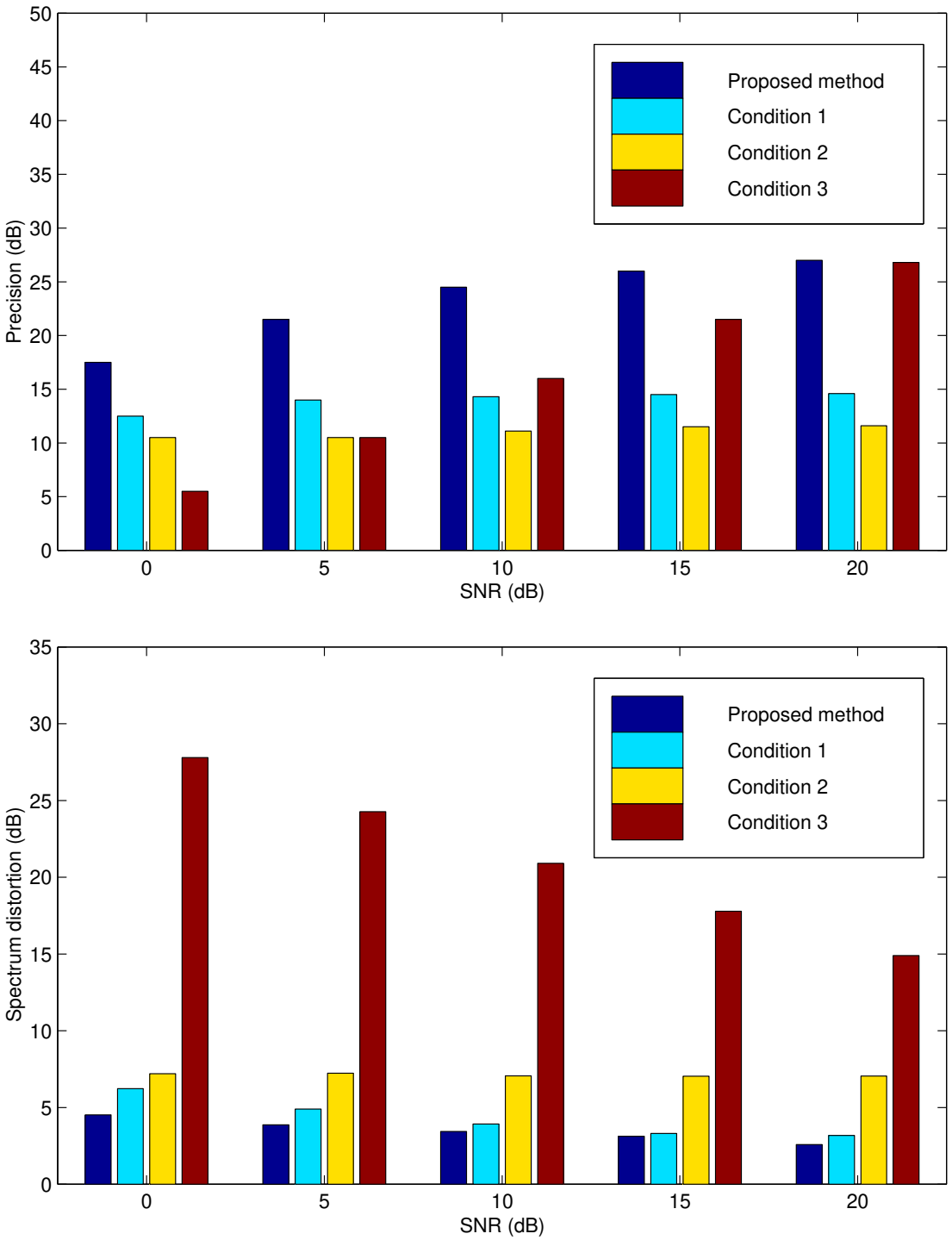


Figure 11: Segregation accuracy for simulation 1: (top panel) precision for the $A_k(t)$, (bottom panel) spectrum distortion for the extracted signal $\hat{f}_1(t)$.

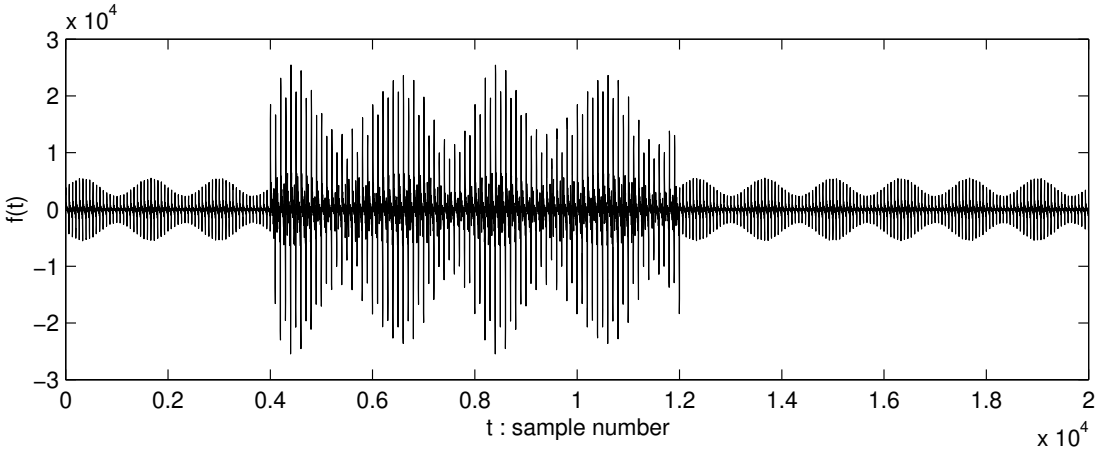


Figure 12: Mixed signals $f(t)$ (SNR= 10 dB).

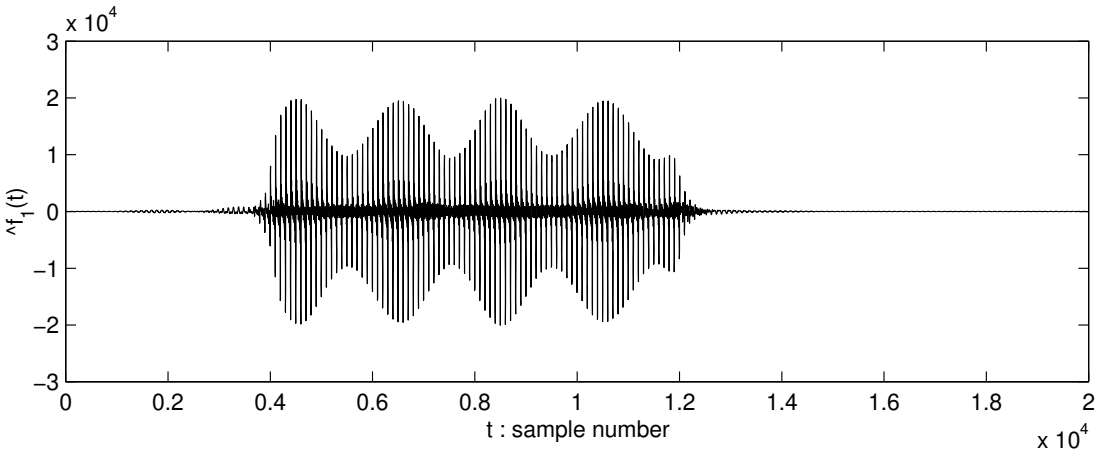


Figure 13: Extracted signal $\hat{f}_1(t)$.

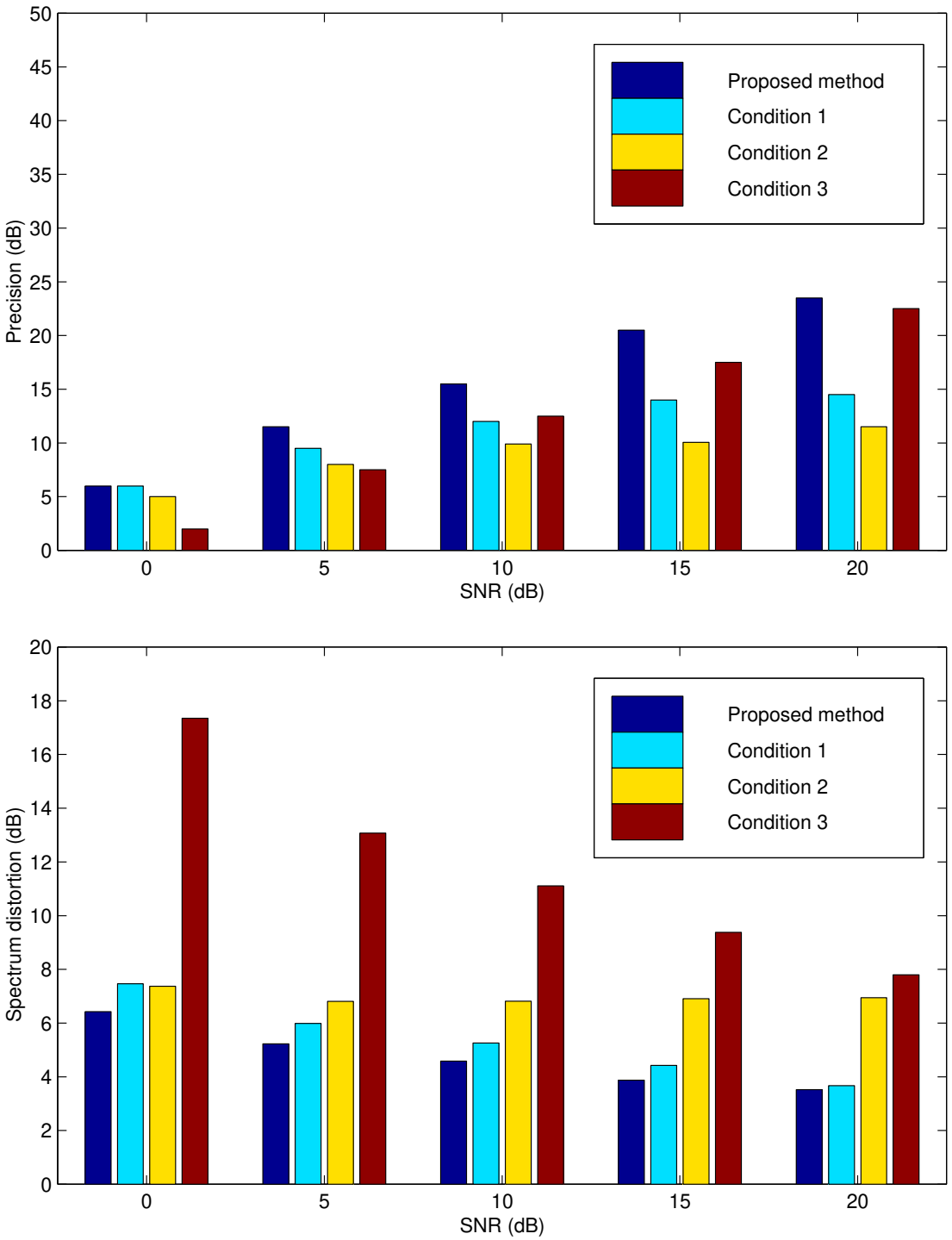


Figure 14: Segregation accuracy for simulation 2: (top panel) precision for the $A_k(t)$, (bottom panel) spectrum distortion for the extracted signal $\hat{f}_1(t)$.

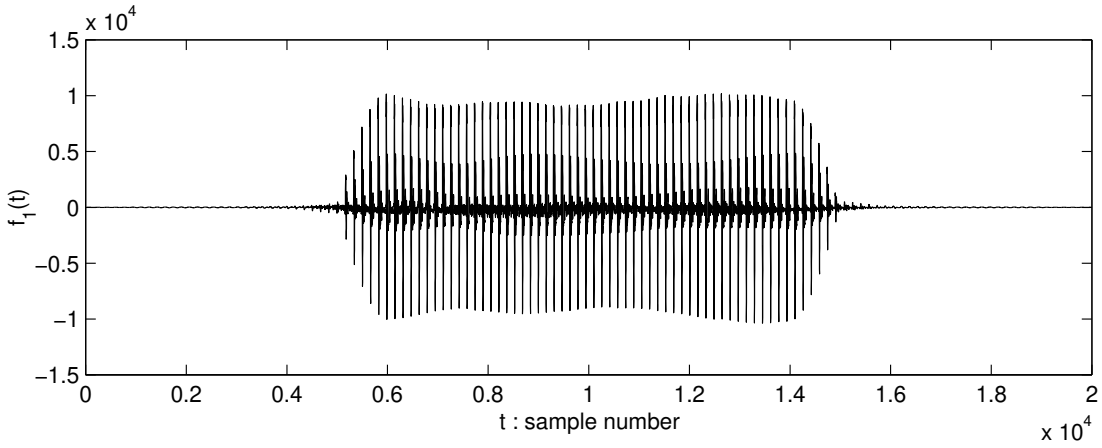


Figure 15: LMA-synthesized vowel /a/ $f_1(t)$.

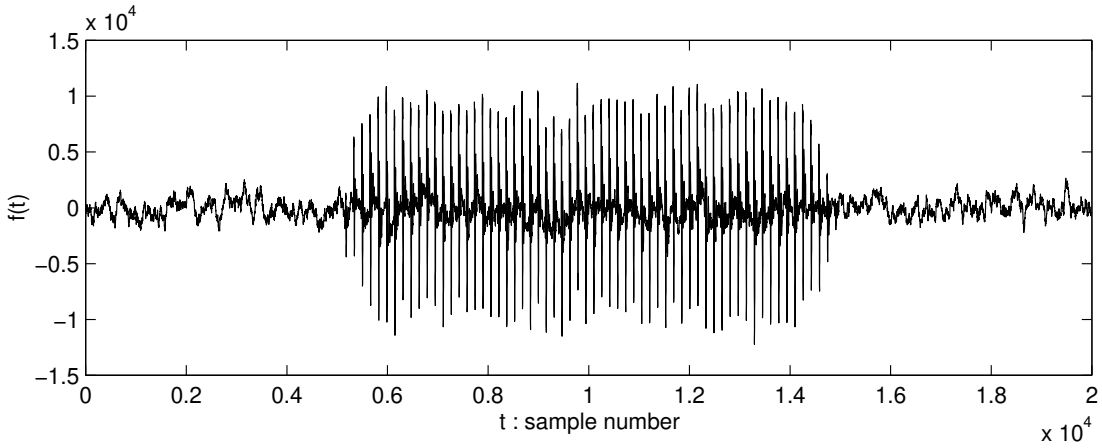


Figure 16: Mixed speech $f(t)$ (SNR= 10 dB).

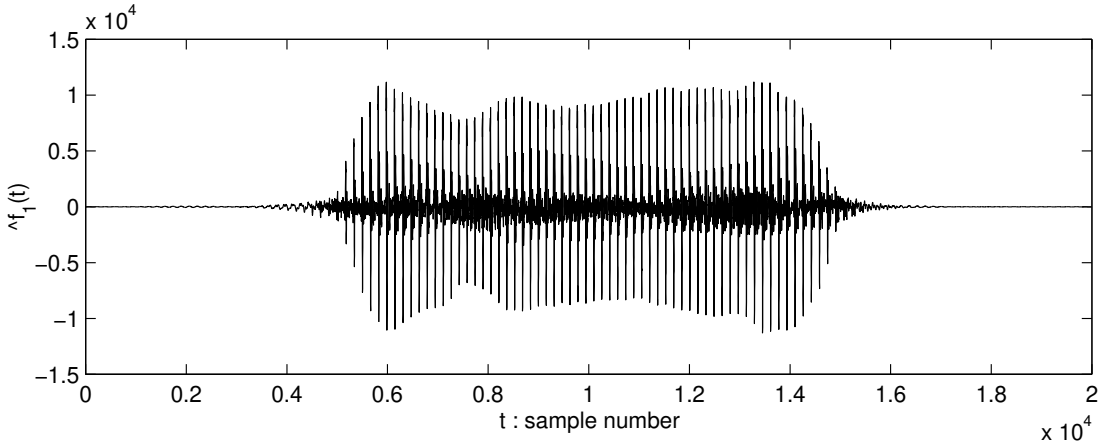


Figure 17: Extracted vowel /a/ $\hat{f}_1(t)$.

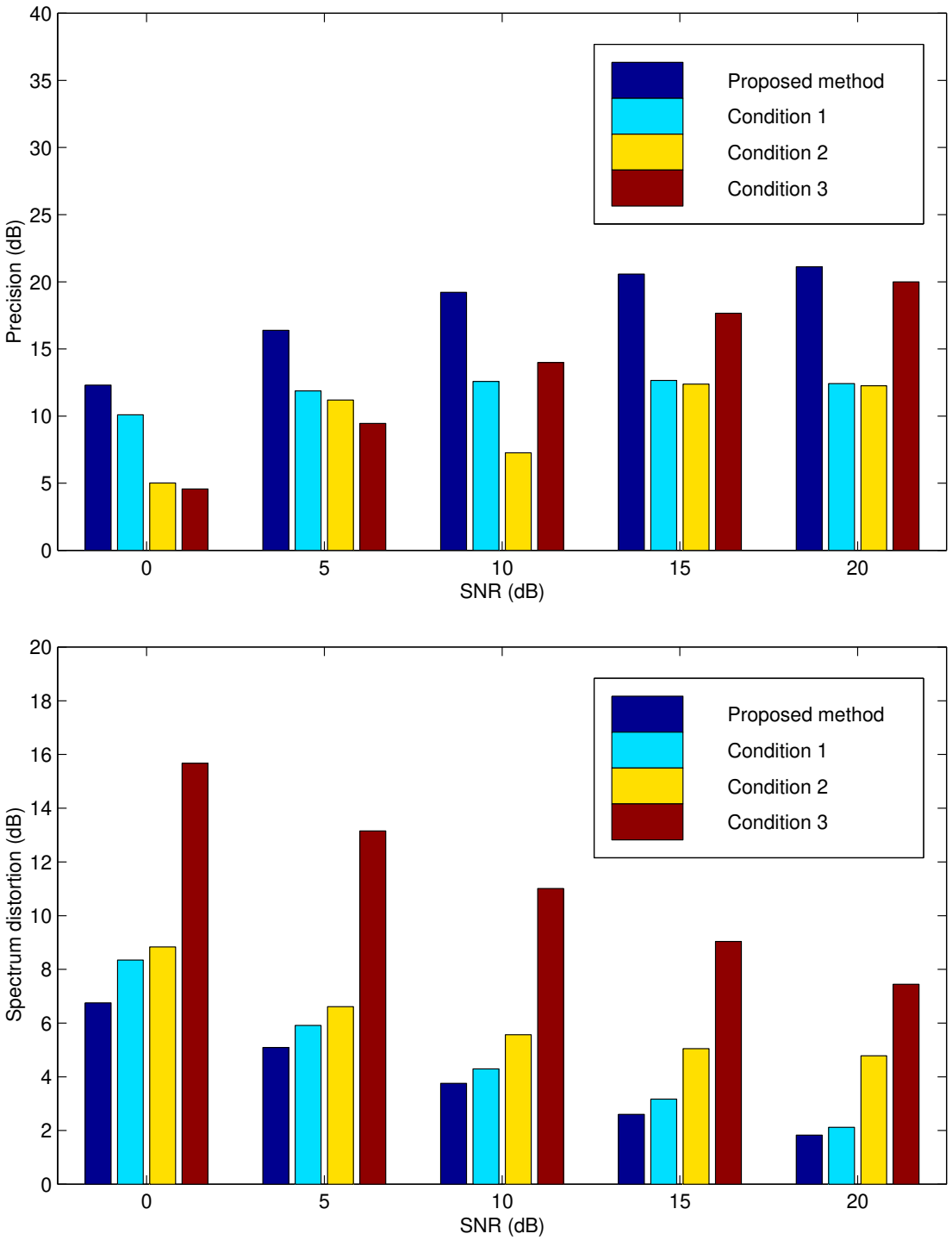


Figure 18: Segregation accuracy for simulation 3: (top panel) precision for the $A_k(t)$, (bottom panel) spectrum distortion for the extracted signal $\hat{f}_1(t)$.

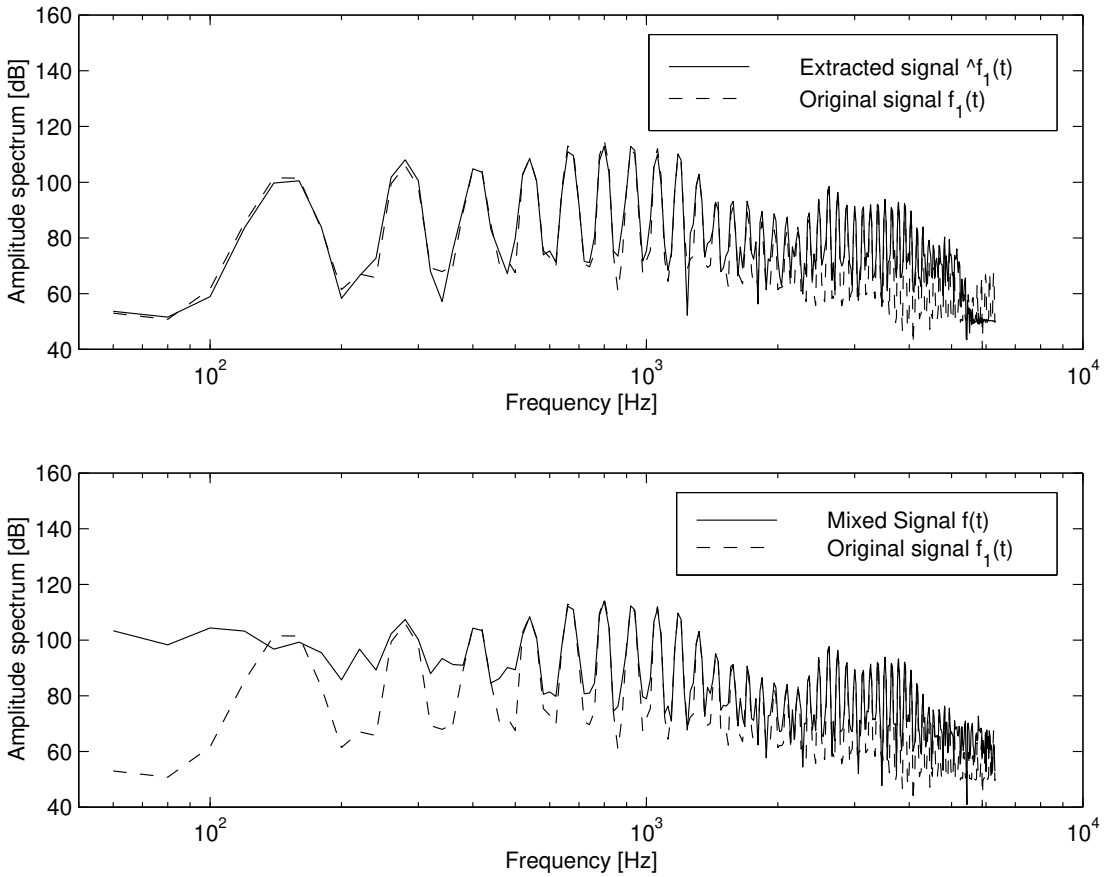


Figure 19: Comparison of the amplitude spectrum of $\hat{f}_1(t)$ with the amplitude spectrum of $f_1(t)$.

Table 1: Definitions of symbols for the segregation problem.

Symbol	Definition
$S_k(t)$	instantaneous amplitude
$\phi_k(t)$	instantaneous output phase
$A_k(t), B_k(t)$	instantaneous amplitude
$\theta_{1k}(t), \theta_{2k}(t)$	instantaneous input phases
$\theta_k(t)$	input phase difference
$F_0(t)$	Fundamental frequency
$C_{k,R}(t), D_{k,R}(t),$ $E_{0,R}(t)$	R-th-order polynomial (differentiable, piecewise)
$C_k(t)$	undetermined function

Table 2: Constraints corresponding to Bregman's regularities.

Heuristic regularity (Bregman, 1993)	Constraint
(i) Unrelated sounds seldom start or stop at exactly the same time.	Synchronous of onset and offset
(ii) Gradualness of change.	Gradualness of change
(a) A single sound tends to change its properties smoothly and slowly.	(piecewise–differentiable polynomial approximation and smoothness)
(b) A sequence of sounds from the same source tends to change its properties slowly.	(piecewise–differentiable polynomial approximation and smoothness)
(iii) When a body vibrates with a repetitive period, this vibrations give rise to an acoustic pattern in which the frequency components are multiples of a common fundamental.	Harmonicity
(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

Table 3: Specifications of the filterbank design

Symbol	Definition	
f_s	sampling frequency	20 kHz
K	channel number	128
W	bandwidth	60 Hz ~ 6 kHz
a	scale parameter	α^p
α	scale	$10^{2/K}$
p	index	$-\frac{K}{2} \leq p \leq \frac{K}{2}$
b	dilation	q/f_s
q	index	$q \in \mathbf{Z}$

Table 4: Definitions of symbol for the Kalman filter.

Symbol	Definition
observed signal	$\mathbf{y}_m = X_k(t_m)$
state variable	$\mathbf{x}_m = -C_{k,0}(t_m)$
observed noise	$\mathbf{v}_m = X_{2,k}(t_m)$
system noise	$\mathbf{w}_m = w_m$
state transition matrix	$\mathbf{F}_m = \Delta C_k(t_m)$
observation matrix	$\mathbf{H}_m = e^{j\omega_k t_m}$
driving matrix	$\mathbf{G}_m = -1$