TITLE

A Method of Signal Extraction from Noise-Added Signal

AUTHORS

† Masashi UNOKI

† Masato AKAGI

AFFILIATION

† Japan Advanced Institute of Science and Technology, Hokuriku Ishikawa-ken, 923-12 Japan

Abstract

This paper proposes a new method of signal extraction from noise-added signal, addressing on the problem of segregating two acoustic sources as a model of acoustic source segregation. This method models some constraints of auditory scene analysis (ASA) and makes it possible to segregate two acoustic sources using the amplitude envelope and the phase deviation (input and output phase) obtained from the output of a wavelet filterbank. Using these three physical clues, the amplitude envelope and the output phase are determined; then, the input phase is determined using the physical constraints translated from heuristic regularities, changes in an acoustic event and gradualness of change, as proposed by Bregman. As an example of segregation using the proposed method, we seek to provide a solution for the problem of segregating two acoustic sources in which a sinusoidal signals added to a bandpassed noise. In particular, if the parameters of the proposed model are set to the human auditory properties, it can be a computational model of co-modulation masking release, which makes extraction of sinusoidal signals when such signals are added to AM bandpassed noise simpler while making the extraction of sinusoidal signals difficult when such signals are added to bandpassed random noise.

Keyword

auditory scene analysis, two acoustic sources segregation, co-modulation masking release, gammatone filter wavelet filterbank

1 Introduction

Developments in recent years have caused the auditory system to be considered an active scene analysis system stimulating the study of acoustic source segregation based on auditory scene analysis (ASA) [1, 2]. If it becomes possible to solve the problem of acoustic source segregation, not only will it become possible to extract sounds required by the listener while rejecting others, but could find application in a robust speech recognition system [4]. We feel that constructing a computational theory of audition in an analogy to a computational theory of vision proposed by Marr [5] will require time complete; however, we feel that modeling based on ASA suggests a new approach in the construction of a computational theory of audition [6, 7, 8] since ASA shows a direction for constructing a computational theory.

Bregman reported that, for solving the problem of ASA in understanding an environment through acoustic events, the human auditory system uses four psychoacoustically heuristic regularities related to acoustic events[2, 3]:

(1) common onset and offset,

(2) gradualness of change,

(3) harmonicity,

(4) changes in an acoustic event.

There already exists ASA-based segregation models utilizing these four regularities: Brown and Cooke's segregation model based on acoustic events [9, 10, 11], Ellis's segregation model based on psychoacoustic grouping rules [12], and Nakatani *et al.*'s segregation model implementing a multi-agent system [13, 14]. Another model is a computational model of quantitative relationships between multiple features on the spectrogram and auditory segregation for auditory segregation of two frequency components, as proposed by Kashino *et al.*[15, **?**]. All these computational segregation models use regularities (1) and (3), as well as the amplitude or power spectrum as the acoustic feature. Because of this, they cannot extract completely the signal from a noise-added signal when signal and noise exist in the same frequency region.

We stress the need for considering not only the amplitude spectrum but also the phase spectrum, when attempting to extract completely the signal from a noise-added signal in which both exist in the same frequency region [17]; based on this stance, we seek to solve the problem of segregating two acoustic sources — basic problem of acoustic source segregation using regularities (2) and (4) as proposed by Bregman [18, 24].

This paper proposes a method of signal extraction from noise-added signal as a solution for the problem of segregating two acoustic sources. This method uses amplitude and phase spectra calculated by the wavelet transform from noise-added signal; it also shows that if the parameters of the proposed model are set to the human auditory properties, the proposed model can be a computational model of co-modulation masking release (CMR) [19].

The paper is organized as follows: Section 2 illustrates the proposed model and then formulates the problem of segregating two acoustic sources; Section 3 shows the design of

the wavelet filterbank and its characteristics; Section 4 shows calculation of the physical parameters and segregation algorithm; Section 5 carries out computer simulations for segregating two acoustic sources to show advantages of the proposed method; Section 6 shows that the proposed model can be a computational model of co-modulation masking release if the model parameters are set to the human auditory properties; Section 7 contains our conclusions.

2 Formulation of the problem of segregating two acoustic sources

In this paper, the problem of segregating two acoustic sources is defined as "the segregation of original signal components from a noise-added signal, where mixed signal is composed of two signals generated by two acoustic sources."

This problem is formulated as follows.

Firstly, we can observe only the signal $f(t)$:

$$f(t) = f_1(t) + f_2(t), \tag{1}$$

where $f_1(t)$ and $f_2(t)$ are the original acoustic signals. The observed signal is decomposed into its frequency components by an auditory filterbank as shown in Fig. 1. Secondly, outputs of the k-th channel, which correspond to $f_1(t)$ and $f_2(t)$, are assumed to be

$$f_1(t) : A_k(t)\sin(\omega_k t + \theta_{1k}(t)) \tag{2}$$

and

$$f_2(t) : B_k(t)\sin(\omega_k t + \theta_{2k}(t)), \tag{3}$$

respectively. Here, ω_k is a center frequency of the auditory filter and $\theta_{1k}(t)$ and $\theta_{2k}(t)$ are input phases of $f_1(t)$ and $f_2(t)$, respectively. Since the output of the k-th channel $X_k(t)$ is the sum of Eqs. (2) and (3), then it is represented by

$$X_k(t) = S_k(t)\sin(\omega_k t + \phi_k(t)). \tag{4}$$

Here,

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

and

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

where $\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$ and $\theta_k(t) \neq n\pi, n \in \mathbf{Z}$. Since the amplitude envelope $S_k(t)$ and the input phase are observable and if the input phases $\theta_{1k}(t)$ and $\theta_{2k}(t)$ are determined, the amplitude envelopes $A_k(t)$ 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

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

respectively. Finally, for each auditory filter, determining of $A_k(t)$ and $B_k(t)$, $f_1(t)$ and $f_2(t)$ are reconstructed from Eqs. (2) and (3) synthesizing each frequency components. Here, $\hat{f}_1(t)$ and $\hat{f}_2(t)$ are the reconstructed $f_1(t)$ and $f_2(t)$, respectively.

In this paper, the analysis synthesis system (filterbank) as shown in Fig. 1 is constructed using the wavelet transform. We assume that $\theta_{1k}(t) = 0$, $\theta_k(t) = \theta_{2k}(t)$, and that $f_1(t)$ is an amplitude modulation signal. We also assume that a center frequency of analytic filter matches a center frequency of $f_1(t)$, and that $f_2(t)$ is a bandpassed random noise which the center frequency is the same as $f_1(t)$. We will also consider the problem of segregating two acoustic sources in which the localized $f_1(t)$ is added to $f_2(t)$.

3 Wavelet filterbank with the gammatone filter

3.1 Definition of the wavelet transform

Firstly, to design a wavelet filterbank, the wavelet transform and the inverse wavelet transform are summarized as follows.

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

$$\tilde{f}(a,b) = \frac{1}{\sqrt{|a|}}\int_{-\infty}^{\infty} f(t)\overline{\psi\left(\frac{t-b}{a}\right)}dt, \tag{9}$$

where a is the "scale parameter," b is the "shift parameter," and $\overline{\psi}$ is the conjugate of ψ. The integral basis function is $\psi(t)$ scale-transformed by the parameter a and is shifted by the parameter b. The selection of $\psi(t)$ allows much mathematically freedom; however, in general, $\psi(t)$ is determined to be an integrable function satisfied by the following admissibility condition [20]:

$$D_\psi := \int_{-\infty}^{\infty}\frac{|\hat{\psi}(\omega)|^2}{|\omega|}d\omega \ < \infty \tag{10}$$

where $\hat{\psi}(\omega)$ is the Fourier transform of ψ. If the above equation is satisfied, ψ is called a "basic wavelet," and the inverse transform exists uniquely as follows[20]:

$$f(t) = \frac{1}{D_\psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\tilde{f}(a,b\ )\psi\left(\frac{t-b}{a}\right)\frac{dadb}{a^2}. \tag{11}$$

If ψ is absolutely integrable function, the admissibility condition implies $\hat{\psi}(0) = 0$.

If the basic wavelet is defined on a complex plane, it is possible that the wavelet transform is represented by the amplitude spectrum $|\tilde{f}(a,b)|$ and the phase spectrum $\arg(\tilde{f}(a,b))$ as follows [21]:

$$\tilde{f}(a,b) = |\tilde{f}(a,b)|e^{j\arg(\tilde{f}(a,b))}. \tag{12}$$

To construct an auditory filterbank simulating the auditory system, we selected as a basic wavelet the gammatone filter to simulate the response of the basilar membrane.

3.2 Characteristic of the gammatone filter

The gammatone filter is an auditory filter designed by Patterson[22], and is known to simulate well the response of the basilar membrane. The impulse response of the gammatone filter is defined by

$$gt(t) = At^{N-1}e^{-2\pi b_f t}\cos(2\pi f_0 t), \quad t \geq 0, \tag{13}$$

where $At^{N-1}e^{-2\pi b_f t}$ is the amplitude term represented by the Gamma distribution and f_0 is the center frequency. The amplitude characteristics of the gammatone filter are represented, approximately, by

$$GT(f) \approx \left[1 + \frac{j(f - f_0)}{b_f}\right]^{-N}, \quad 0 < f < \infty, \tag{14}$$

where $GT(f)$ is the Fourier transform of $gt(t)$ and represents bandpass filtering with a center frequency of f_0. Characteristics of impulse response and amplitude of the gammatone filter are shown in Fig. 2. Since it is clear from this figure that $GT(f) \approx 0$, the gammatone filter satisfies approximately the admissibility condition, meaning it can be used sufficiently as the basic wavelet.

3.3 Wavelet filterbank

To represent the wavelet transform as Eq. (12), we will redesign the gammatone filter by the Hilbert transform. As a results, basic wavelet becomes

$$\psi(t) = At^{N-1}e^{j2\pi f_0 t - 2\pi b_f t}, \tag{15}$$

wavelet filterbank is designed with a center frequency f_0 of 600 Hz, a bandpassed region from 60 Hz to 6000 Hz, and a number of filters K of 128. For convenience, we have used the integral (continuous) wavelet transform; however, when the wavelet filterbank is implemented on a computer, we will use a discrete wavelet transform with the following conditions[18]: sampling frequency $f_s = 20$ kHz, the scale parameter $a = \alpha^p, -\frac{K}{2} \leq p \leq \frac{K}{2}, \alpha = 10^{2/K}$, and the shift parameter $b = q/f_s$, where $p, q \in \mathbf{Z}$. Frequency characteristics of the wavelet filterbank are shown in Fig. 3. Amplitude characteristics of this filterbank overlap completely with the bandpassed region as shown in Fig. 3 while ERB (Equivalent Rectangular Bandwidth) of the filters are do not [18].

4 Calculation of physical parameters

4.1 Calculation of amplitude envelope $S_k(t)$ and output phase $\phi_k(t)$

The amplitude envelope $S_k(t)$ and the output phase $\phi_k(t)$ can be calculated using the following lemma.

[**Lemma1**] The amplitude envelope $S_k(t)$ is calculated by

$$S_k(t) = |\tilde{f}(\alpha^{k-\frac{K}{2}}, t)|, \tag{16}$$

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

$$\phi_k(t) = \int \left(\frac{d}{dt} \arg\left(\tilde{f}(\alpha^{k-\frac{K}{2}}, t) \right) - \omega_k \right) dt, \tag{17}$$

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

Proof. See Appendix 1. □

4.2 Calculation of input phase $\theta_k(t)$

Input phase $\theta_k(t)$ can be determined by applying three physical constraints: (i) gradualness of change, (ii) continuity (temporal proximity), and (iii) changes in an acoustic event. In particular, constraints (i) and (iii) are regularity (2) and (4) proposed by Bregman.

We will first apply regularity (i). This regularity states that "a single sound tends to change its properties smoothly and slowly (gradualness of change)"[2]. We will consider this in the following physical constraint.

[Constraint1] (gradualness change) Temporal differentiation of the amplitude envelope $A_k(t)$ must be represented by Rth-order differentiable polynomial $C_{k,R}(t)$ as follows:

$$\frac{dA_k(t)}{dt} = C_{k,R}(t). \tag{18}$$

□

Applying Physical constraint 1 into Eq. (7), a linear differential equation is obtained as follows:

$$y'(t) + \frac{P'(t)}{P(t)} y(t) = \frac{Q'(t) - C_{k,R}(t)}{P(t)}, \tag{19}$$

where $P(t) = S_k(t)\sin\phi_k(t)$, $Q(t) = S_k(t)\cos\phi_k(t)$, $y(t) = \cot\theta_k(t)$. $\theta_k(t)$ can be determined by solving the linear differential equation (19).

[Lemma2] By solving the linear differential equation, a general solution of the input phase $\theta_k(t)$ is determined by

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin\phi_k(t)}{S_k(t)\cos\phi_k(t) + C_k(t)} \right), \tag{20}$$

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

If $C_k(t)$ is determined, then $\theta_k(t)$ is uniquely determined by the above equation. Although it is possible to estimate the coefficients $C_{k,r}, r = 0, 1, \cdots, R$ by considering as an optimization problem, we will assume, in order to reduce the computational costs that in small segment Δt, $C_{k,R}(t) = C_{k,0}$. As this point, Eq. (18) is equivalent to $dA_k(t)/dt = 0$, and amplitude envelope $A_k(t)$ does not fluctuate in small segment Δt.

Next, we will use regularity (ii) to segregate each small segment Δt. Regularity (ii) means that "each physical parameter must retain temporal proximity in the bound ($t = T_r$) between pre-segment ($T_r - \Delta t \leq t < T_r$) and post-segment ($T_r \leq t < T_r + \Delta t$)." In order to apply this regularities to physical parameters, it is considered in the following physical constraint.

[Constraint2] (proximity) In the bound ($t = T_r$) between pre-segment and post-segment, each physical parameter $A_k(t)$, $B_k(t)$, and $\theta_k(t)$ must be connected within ΔA, ΔB, $\Delta\theta$, respectively. That is,

$$\begin{cases} |A_k(T_r + 0) - A_k(T_r - 0)| \leq \Delta A, \\ |B_k(T_r + 0) - B_k(T_r - 0)| \leq \Delta B, \\ |\theta_k(T_r + 0) - \theta_k(T_r - 0)| \leq \Delta\theta. \end{cases} \tag{21}$$

□

From Eqs. (7), (8) and (20), amplitude envelopes $A_k(t)$ and $B_k(t)$, and input phase $\theta_k(t)$ are functions of the unknown coefficient. Therefore, by considering the above relationships, we can interpret physical constraint 2 in order to determine $C_{k,0}$, which is restricted within

$$C_{k,\alpha} \leq C_{k,0} \leq C_{k,\beta}, \tag{22}$$

where $C_{k,\alpha}$ and $C_{k,\beta}$ are the upper limit and the lower limit $C_{k,0}$ in the bound between the two segments.

Finally, we will apply regularity (iii). This regularity states that "many changes take place in an acoustic event that affect all the components of the resulting sound in the same way and at the same time" [2]. This regularity is considered the following physical constraint.

[Constraint3] (Changes in an acoustic event) The amplitude envelope $B_k(t)$ must be highly correlated with the amplitude envelope $B_{k\pm1}(t)$ obtained by the output of adjacent channel:

$$B_k(t) \approx B_{k\pm1}(t) \tag{23}$$

□

Since an amplitude envelope $B_k(t)$ is a function of $C_{k,0}$ from Eqs. (8) and (20), and let $\hat{B}_k(t)$ be an amplitude envelope $B_k(t)$ determined by any $C_{k,0}$. In addition, a correlation between amplitude envelopes is defined by

$$\mathrm{Corr}(\hat{B}_k, \hat{\hat{B}}_k) = \frac{\langle \hat{B}_k, \hat{\hat{B}}_k \rangle}{||\hat{B}_k|| ||\hat{\hat{B}}_k||}, \tag{24}$$

where $\hat{\hat{B}}_k(t) = (\hat{B}_{k+1}(t) + \hat{B}_{k-1}(t))/2$. Let T_B be an arbitrary past time, where integral region is $T_B \leq t < T_r + \Delta t$. $\hat{B}_{k\pm1}(t)$ can be determined using $\hat{A}_k(t)$ and $\hat{B}_k(t)$ obtained by $C_{k,0}$, from Fig. 4. Here, physical constraint 3 can be considered as selecting $C_{k,0}$ when correlation (24) is a maximum. That is, by selecting $C_{k,0}$ as

$$\max_{C_{k,\alpha} \leq C_{k,0} \leq C_{k,\beta}} \mathrm{Corr}(\hat{B}_k, \hat{\hat{B}}_k), \tag{25}$$

the input phase $\theta_k(t)$ can be uniquely determined from Eq. (20).

4.3 Segregation algorithm

The theorem for segregation is obtained from previous lemma as follows.

[Theorem1] (Segregation algorithm) For the problem of segregating two acoustic sources where $\theta_{1k}(t) = 0$ and $\theta_k(t) = \theta_{2k}(t)$, the input phase $\theta_k(t)$ can be determined using physical constraints 1–3 from Eqs. (20) and (25). Therefore, this problem can be solved using the amplitude envelope $S_k(t)$ and the output phase $\phi_k(t)$ from Eqs. (1)–(8).

Proof. Refer to Lemma 1 and 2. □

Segregation algorithm based on Theorem 1 is shown in Fig. 5. When the problem of segregating two acoustic sources is to be solved using Theorem 1, signal duration which two signals exist in the same time region must be known. In Section 2, we assumed that when localized $f_1(t)$ is added to $f_2(t)$, the signal duration can be detected using onset and offset of $f_1(t)$. By focusing on the temporal deviation of $S_k(t)$ and $\phi_k(t)$, onset and offset of $f_1(t)$ can be determined as follows:

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

Therefore, the segregated duration is $t_{\rm on} \leq t \leq t_{\rm off}$.

5 Simulations of segregation

In this section, simulations are carried out using the previous method. For each mixed signal, the parameters of the proposed method are set so that the small segment is $\Delta t = 3/f_0$ and T_B is past time of about $100\Delta t$. In the segregated duration ($T_{\rm on} \leq t \leq T_{\rm off}$), let $S_{\max}$ be the maximum of $S_k(t)$, $\Delta B = 0.027 \cdot S_{\max}$ and $\Delta\theta = \pi/20$. However, ΔA is set to $\Delta A = |A_k(T_r - \Delta t) - A_k(T_r - 2\Delta t)|$ based on Eq. (21) because it is difficult to determine as a constant by as it is affected by $C_k(t) = C_{k,0}$, $T_r \leq t < T_r + \Delta t$.

5.1 Experimental stimuli

For experimental stimuli, we will assume $f_1(t)$ is a sinusoidal signal and $f_2(t)$ are two types of noise where $f_{21}(t)$ is an AM bandpassed noise and $f_{22}(t)$ is a bandpassed random noise, as follows:

$$f_1(t) = F_{\rm BP}(g_1), \tag{26}$$

$$g_1(t) = \begin{cases} 1200\sin(2\pi f_0 t), \\ \qquad 0.3 + T_m \leq t \leq 0.7 + T_m, \\ 0, \quad \text{otherwise} \end{cases} \tag{27}$$

$$f_{21}(t) = \sum_{f=f_0-500}^{f_0+500} E_M(t)\sin(2\pi ft + R(f)), \quad 0 \le t \le 1.0, \tag{28}$$

$$f_{22}(t) = \sum_{f=f_0-500}^{f_0+500} E_R(f,t)\sin(2\pi ft + R(f)), \quad 0 \le t \le 1.0, \tag{29}$$

where $F_{\mathrm{BP}}(\cdot)$ is a bandpass filter with a center frequency f_0 and a bandwidth of 23 Hz (as shown in Fig. 2), $f_0 = 600$ Hz, $T_m = 0.0125m, m = 0, 1, \cdots, 9$, and $R(f)$ is an uniform random within $[-\pi, \pi]$. Here, $E_M(t)$ and $E_R(f,t)$ are amplitudes in which white noise lowpass filtered at 30 Hz and added to the bias value to prevent over modulation. Note that the amplitude $E_R(f,t)$ fluctuates independently in frequency region f, and that the power of noise is adjusted so that $\sqrt{f_{21}(t)^2/f_{22}(t)^2} = 1$. Here, the bandwidth of noise is 1 kHz and the SNR between a sinusoidal signal and the bandpassed noise is -8.5 dB. These mixed signals are shown in Fig. 6.

The two types of mixed signals are $f_M(t) = f_1(t) + f_{21}(t)$ when $f_2(t) = f_{21}(t)$ and $f_R(t) = f_1(t) + f_{22}(t)$ when $f_2(t) = f_{22}(t)$. When a human hears these mixed signals, he can hear the sinusoidal signal from $f_M(t)$ caused by CMR; however cannot hear the sinusoidal signal from $f_R(t)$ caused by Masking. These mixed signals are shown in Fig. 7.

In this simulation, segregation is done using 10 mixed signals which are made by varying the onset of $f_1(t)$, $m = 0, 1, \cdots, 9$ in Eq. (27).

5.2 Results

First, simulation of segregation are conducted using 10 mixed signals $f_M(t)$. $f_M(t)$ is decomposed by a wavelet filterbank, and is then applied to the segregation algorithm as shown in Fig. 5. $S_k(t)$ and $\phi_k(t)$ are determined as shown in Fig. 8. Onset time and offset time of $f_1(t)$, T_{on} and T_{off}, are determined as shown in Fig. 9. The results of simulation for a sinusoidal signal ($m = 0$) as shown in Fig. 6 are shown in Fig. 10. From this figure, it can be seen that the proposed method can extract a sinusoidal signal from mixed signal.

Similarly, simulations of segregation are carried out using 10 mixed signals $f_R(t)$. The proposed model extracts so little of $f_1(t)$ that $f_2(t)$ becomes approximately the same as $f(t)$.

When the extracted signal corresponds to 10 mixed signals, the mean value of SNR in the time region is calculated as

$$\mathrm{SNR} = 10\log_{10}\frac{\int_0^T f_i^2(t)dt}{\int_0^T (f_i(t) - \hat{f}_i(t))^2 dt}, \qquad i = 1, 2. \tag{30}$$

It is shown that in the case of $f_M(t)$, the SNR of $\hat{f}_1(t)$ is 12.9 dB (standard deviation 2.58) and the SNR of $\hat{f}_2(t)$ is 10.1 dB (standard deviation 1.58). In the case of $f_R(t)$, the SNR of $\hat{f}_1(t)$ is 1.6 dB (standard deviation 1.58) and the SNR of $\hat{f}_2(t)$ is 8.7 dB (standard deviation 0.48). If $f(t)$ can also enhanced using a bandpassed filter (BPF) with a center frequency of 600 Hz and a bandwidth of 23 Hz and if $\hat{f}_1(t)$ is the enhanced signal, the SNR of $\hat{f}_1(t)$ is

only 8.1 dB. From these results, it can be stated that the proposed method is superior in improving SNR in comparison with method using BPF. In the case of $f_M(t)$, the proposed method can not only extract a sinusoidal signal more accurately than $f_R(t)$ but can also a bandpassed noise. While engineering application call for a method of segregating all signals from noisy signals, the proposed method successfully extracts signals by taking advantage of the constraints of ASA (c. f. $f_M(t)$). Note that the above results show of CMR which is as the human auditory phenomenon. To examine this similarity, we will carry out a simulation of CMR in next section by setting the model parameters to the human auditory properties.

6 Parameter setting for the human auditory properties

6.1 Re-design of the wavelet filterbank

The wavelet filterbank shown in Fig. 1 is adjusted to the human auditory properties. ERB (Equivalent Rectangular Bandwidth) of an auditory filter described in previous section is determined as a function of the number of auditory filters. By setting the parameters to the human auditory properties. The constant Q filterbank is constructed ERB equals 1 and when a center frequency of auditory filter is 600 Hz (K=128) [24].

6.2 Conditions and Results

We will consider as experimental stimuli ten pairs of two types of mixed signals used in previous section, $f_M(t)$ and $f_R(t)$. Although the simulation for CMR [19] was carried out for a function of the bandwidth of noise, this simulation is carried out for a function of a number of adjacent auditory filters L related to the bandwidth of a bandpassed noise. The amplitude envelope $\hat{\hat{B}}_k(t)$ in physical constraint 3 is determined by

$$\hat{\hat{B}}_k(t) = \frac{1}{2L} \sum_{\ell=-L,\ell\neq 0}^{L} \hat{B}_{k+\ell}(t) \tag{31}$$

while the input phase can be uniquely determined from Eqs. (20) and (25).

Relationship between the bandwidth and the improved SNR is shown in Fig. 11. In this figure, the vertical axis shows inversely the improved SNR of a sinusoidal signal and the horizontal axis shows the bandwidth in relation to L. The real line and the error-bar show mean and standard deviation of the SNR, respectively. It was shown that, for the mixed signal $f_M(t)$, the SNR of sinusoidal signal $\hat{f}_1(t)$ can be improved as the number of adjacent auditory filters L increase. In contrast, it is shown that, for the mixed signal $f_R(t)$, $\hat{f}_1(t)$ cannot be improved as L increases. Here, improvement of the extracted sinusoidal signal is equivalent to masking release. These results show that the masking of a sinusoidal signal can be released as a function of the bandwidth of bandpassed noise when noise is AM banpassed noise, and that it cannot be released as a function of the bandwidth when noise is bandpassed random noise. For reasons, the proposed model can be interpreted as a computational model of CMR.

7 Conclusions

This paper proposed a method of signal extraction from noise-added signal using the amplitude envelope, the output phase and the input phase obtained from a noise-added signal passed through the wavelet filterbank. Results of simulation using this method show that a sinusoidal signal can be easily extracted if it is added to AM bandpassed noise, and that it cannot be extracted if it is added to a bandpassed noise. In applications, we feel that any signal can be segregated from a noisy signal. However, because it is impossible to segregate for a signal without using certain constraints, the proposed model extracts desired signal from a noisy signal using constraints based on auditory scene analysis.

If the parameters of wavelet filterbank are set to human auditory properties, it can be shown that the masking of a sinusoidal signal can be released as a function of the bandwidth of bandpassed noise when noise is AM bandpassed noise, and that it can not be released as a function of the bandwidth when noise is a bandpassed random noise. From these results, it can be interpreted that the proposed model is a computational model of co-modulation masking release.

In this paper two of the four heuristic regularities proposed by Bregman, (2) gradualness of change and (4) changes in an acoustic event, were considered as physical constraints in order to determine the input phase. Future work includes determining the input phases $\theta_{1k}(t)$ and $\theta_{2k}(t)$. We feel that these parameters can be determined using physical constraints related the remain regularities, (1) common onset and offset and (3) harmonicity. We also feel that the proposed model can extract the desired complex tone and speech from noisy signals using the above constraints.

References

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

[2] A. S. Bregman, “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.

[3] Hideki Kawahara, “Auditory Scene Analysis in Speech Communications,” The 1994 Autumn Meeting of ASJ, 2-7-13, 1994 (in Japanese).

[4] Masato Akagi, “Cocktail Party Effect and Its Modeling,” The Journal of IEICE. vol.78 no.5 pp.450-453, May 1995 (in Japanese).

[5] D.Marr, “Vision, A Computational Investigation into the Human Representation and Processing of Visual Information,”Freeman, New York 1982.

[6] Hideki Kawahara, “Toward The Computational Theory of Audition,” ASJ Tech. Rep., H-94-63, Nov. 1994 (in Japanese).

[7] Toshio Irino and Roy D. Patterson, “A Computational Theory of Auditory Event Detection and Enhancement,” ASJ Tech. Rep., H-94-64, Nov. 1994 (in Japanese).

[8] Toshio Irino, "A Computational Theory of the Peripheral Auditory System," IEICE Tech. Rep., SP95-40, July 1995 (in Japanese).

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

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

[11] M.P.Cooke and G.J.Brown, "Computational auditory scene analysis : Exploiting principles of perceived continuity," Speech Communication, pp. 391-399, North Holland,13, Dec. 1993.

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

[13] Tomohiro Nakatani, Takeshi Kawabata, and Hiroshi G. Okuno, "Computational Approach to Auditory Stream Segregation," ASJ Tech. Rep., H-93-83, Dec. 1993 (in Japanese).

[14] T.Nakatani, H.G.Okuno and T.Kawabata, "Unified Architecture for Auditory Scene Analysis and Spoken Language Processing," ICSLP'94, vol.24, no.3, Sept. 1994.

[15] Kunio Kashino, Hidehiko Tanaka, "A Computational Model of Auditory Segregation of Two Frequency Components — Evaluation and Integration of Multiple Cues —," IEICE Trans. A, vol.J77-A, no.5, pp. 731-740, May 1994 (in Japanese).

[16] Kunio Kashino, "Toward computational auditory scene analysis — A first step —" J. Acoust. Soc. Jpn (J), vol.50 no.12, pp. 1023-1028, Dec. 1994.

[17] Hiroyuki Ikawa and Masato Akagi, "Auditory segregation," The Spring Meeting of ASJ, 3-4-15, Mar. 1994 (in Japanese).

[18] Masashi Unoki and Masato Akagi, "An Extraction Method of the signal from Noise-added Signals," ASJ Tech. Rep., H-95-79, Nov. 1995 (in Japanese).

[19] Brian C.J.Moore, "An Introduction to the Psychology of Hearing," Academic Press, London, 1989.

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

[21] Hideki Kawahara, "Wavelet analysis in auditory perception research," J. Acoust. Soc. Jpn (J), vol.47, no.6, pp.424-429, June 1991.

[22] Roy D.Patterson and John Holdsworth, "A Functional Model of Neural Activity Patterns and Auditory Images," Advances in speech, Hearing and Language Processing, vol.3, JAI Press, London, 1991.

[23] Masato Akagi, "Auditory Filter and Its Modeling," The Journal of IEICE, vol.77, no.9, pp. 948-956, Sept. 1994 (in Japanese).

[24] Masashi Unoki, Masato Akagi, "A Study on Computational model of Co-modulation Masking Release," IEICE Tech. Rep., SP96-37, July 1996 (in Japanese).

Appendix 1. Proof of Lemma 1

The wavelet transform in Eq. (12) is a complex representation for the output of analytic filter in Eq. (4).

$$\begin{aligned} X_k(t) &= S_k(t)e^{j(\omega_k t+\phi_k(t))} \\ &:= \tilde{f}(a,b), \qquad a=\alpha^{k-\frac{K}{2}}, b=t. \end{aligned} \tag{32}$$

Representing an absolute value for both terms, we obtain

$$|X_k(t)| = S_k(t) = |\tilde{f}(\alpha^{k-\frac{K}{2}}, t)|. \tag{33}$$

Similarly, comparing the phase terms between Eqs. (32) and (12), we obtain

$$\omega_k t + \phi_k(t) = \arg(\tilde{f}(a,b)). \tag{34}$$

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

$$\arg(\tilde{f}(a,b)) = \tan^{-1}\frac{Im\{\tilde{f}(a,b)\}}{Re\{\tilde{f}(a,b)\}}, \tag{35}$$

it becomes a periodical ramp function within $-\pi \le \arg(\tilde{f}(a,b)) \le \pi$. Differentiating both terms in Eq. (34), it becomes

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

After clearing, we obtain

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

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

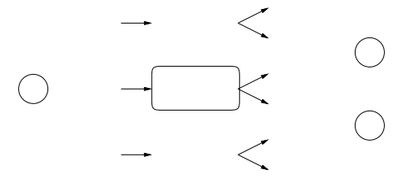
$$\phi_k(t) = \int\left(\frac{d}{dt}\arg\left(\tilde{f}(\alpha^{k-\frac{K}{2}}, t)\right) - \omega_k\right)dt.$$

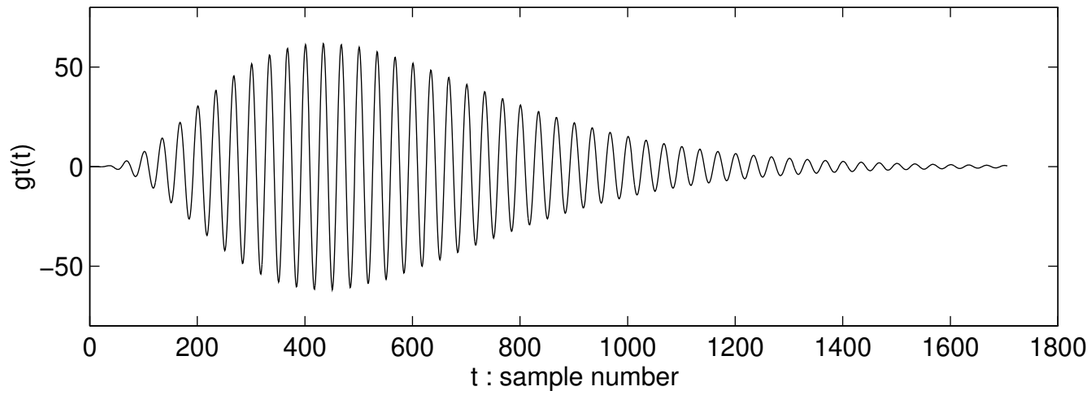
□

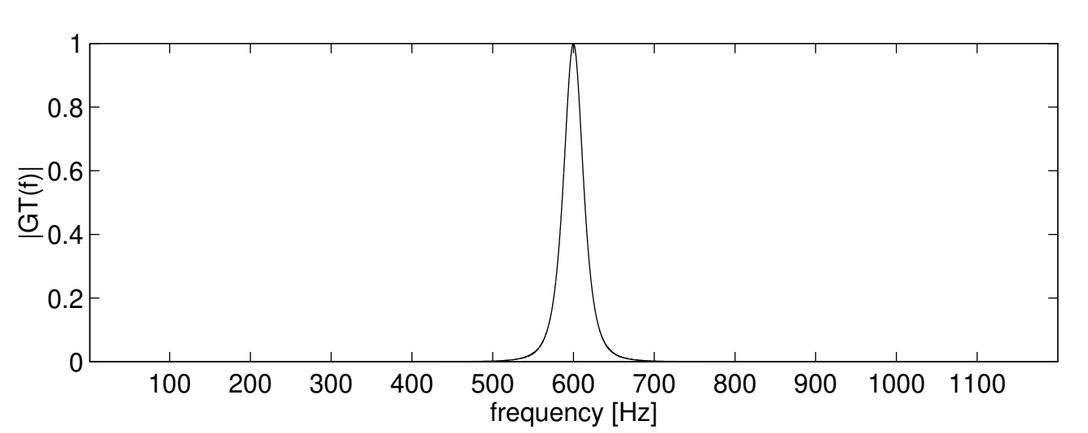
Biography

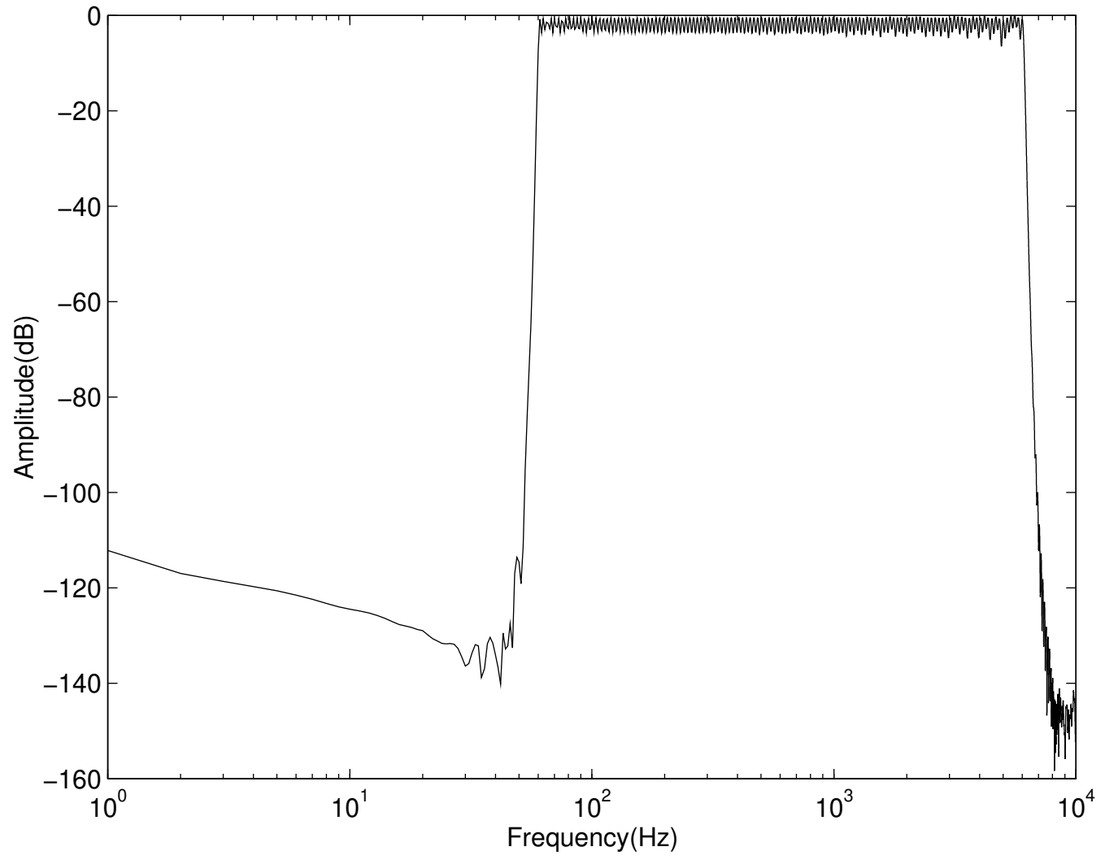
- **Masashi Unoki** was born in Akita, Japan on June 26, 1969. He Graduated the School of Information Engineer, the Polytechnic University in 1994. He received master degree (Information Science) from Japan Advanced Institute of Science and Technology (JAIST) in 1996. Since April 1996, he has been an doctoral program in the School of Information Science, Japan Advanced Institute of Science and Technology. His main research interests are in digital signal processing for acoustic signals and modeling of the auditory system. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Acoustical Society of Japan (ASJ).

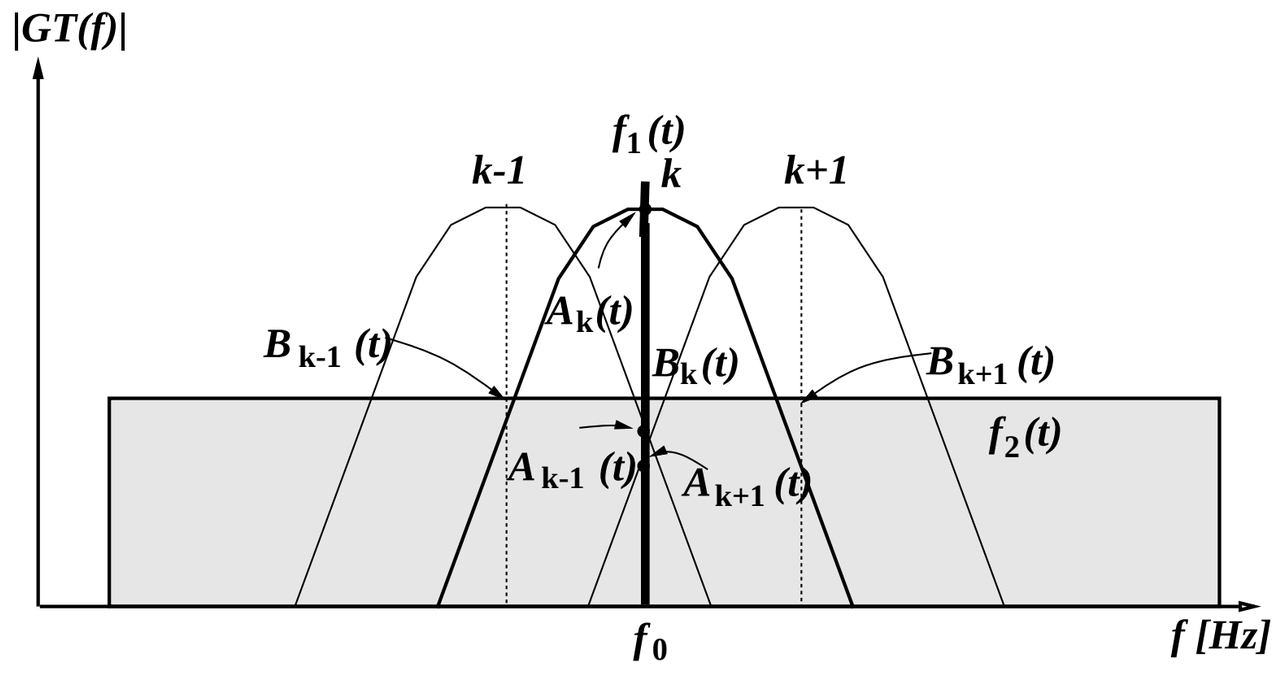
- **Masato Akagi** was born in Okayama, Japan on September 12, 1956. He received a B.E. degree from Nagoya Institute of Technology in 1979, and M.E. and D.E. degrees from Tokyo Institute of Technology in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been with School of Information Science, Japan Advanced Institute of Science and Technology (JAIST) and now he is an Associate Professor of JAIST. His research interests include speech perception, modeling of speech perception mechanisms of humans, and signal processing of speech. During 1988, he joined the Research Laboratories of Electronics, MIT as a visiting researcher and in 1993, he studied at the Institute of Phonetic Science, Univ. of Amsterdam. Dr. Akagi is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of Japan (ASJ), the Institute of Electrical and Electronic Engineering (IEEE), the Acoustical Society of America (ASA), and the European Speech Communication Association (ESCA).

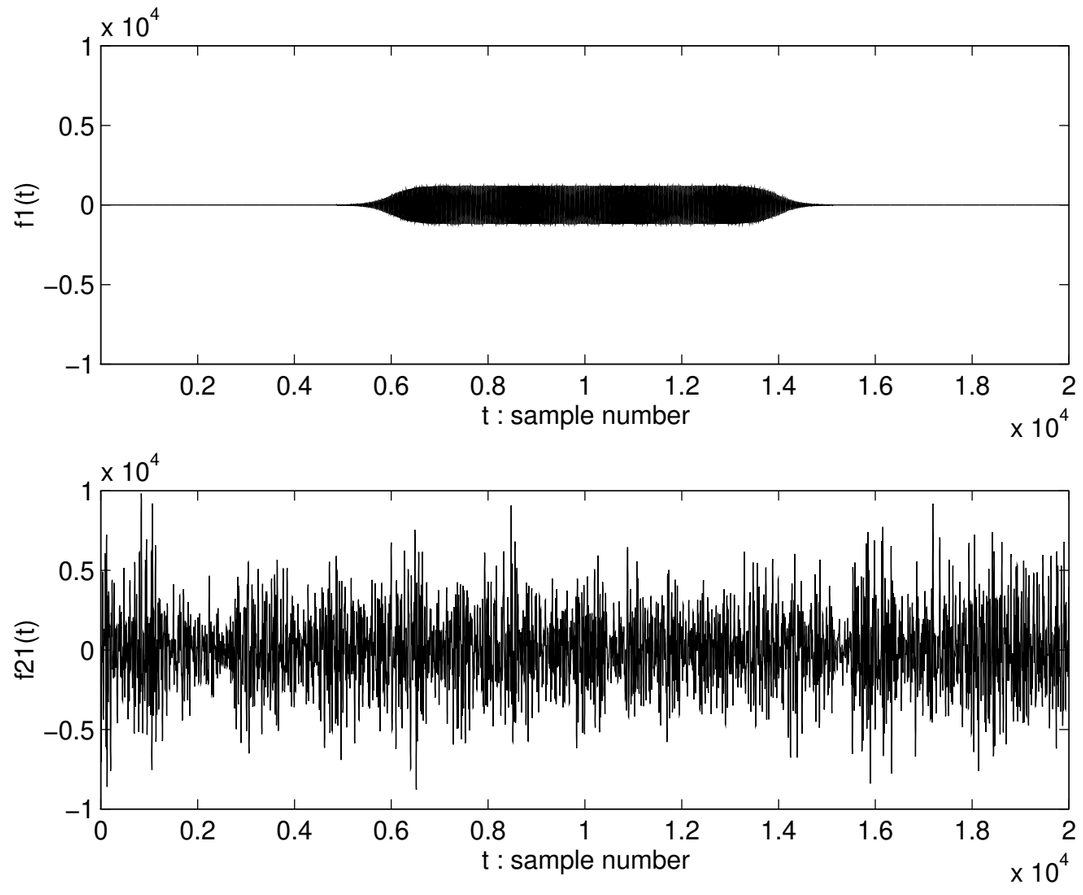


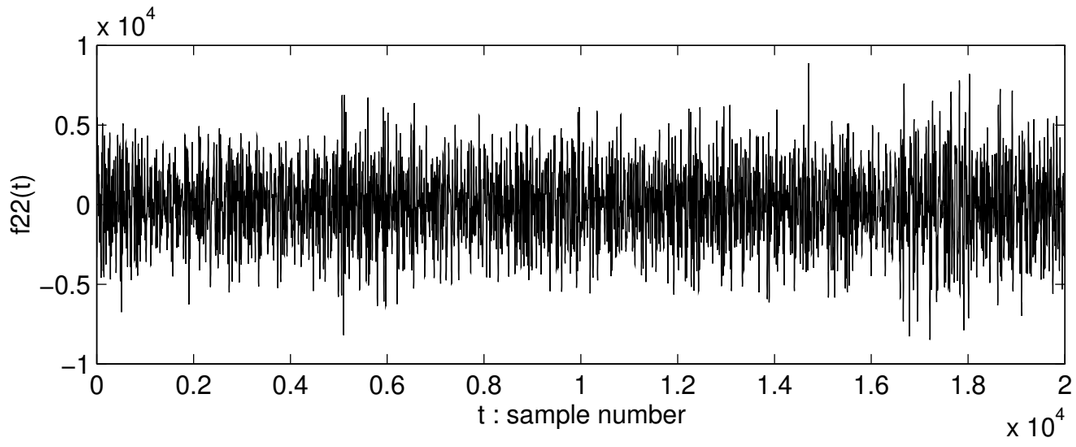

50
0
−50
gt(t)
0
200
400
600
800
1000
1200
1400
1600
1800
t : sample number

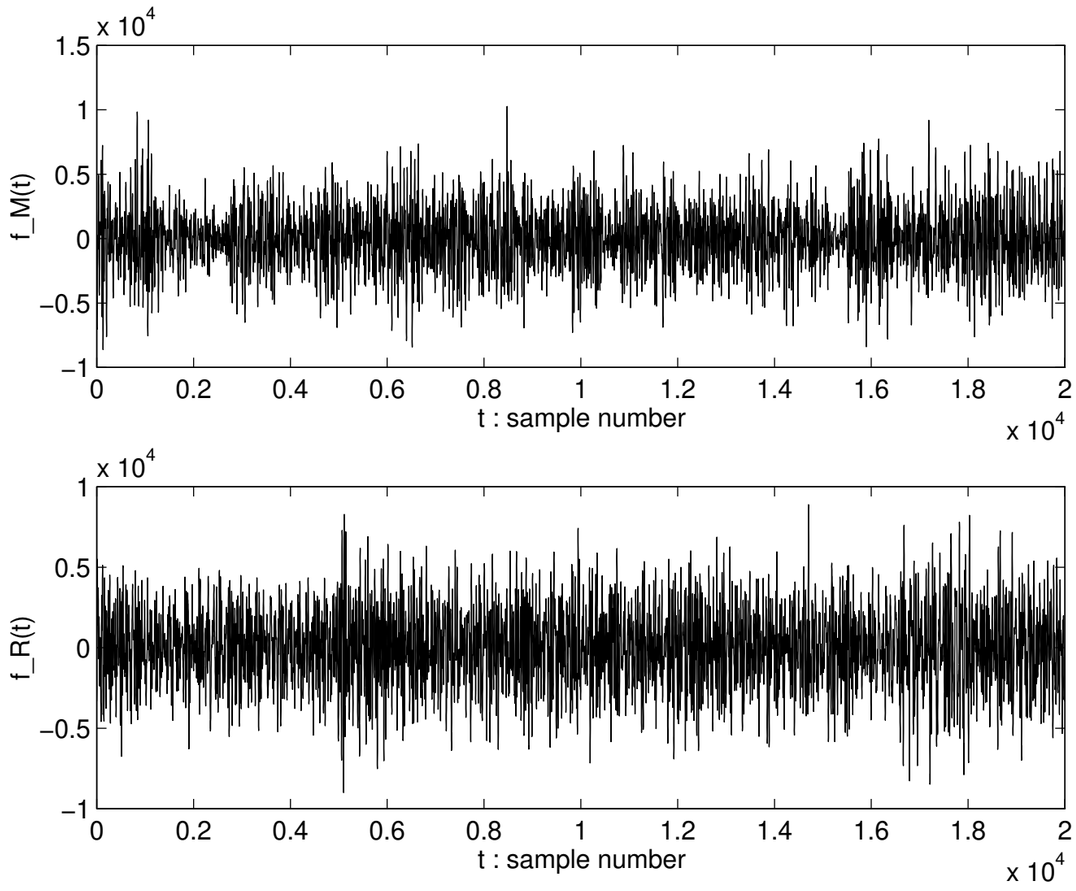

1
0.8
0.6
0.4
0.2
0
|GT(f)|
100
200
300
400
500
600
700
800
900
1000
1100
frequency [Hz]

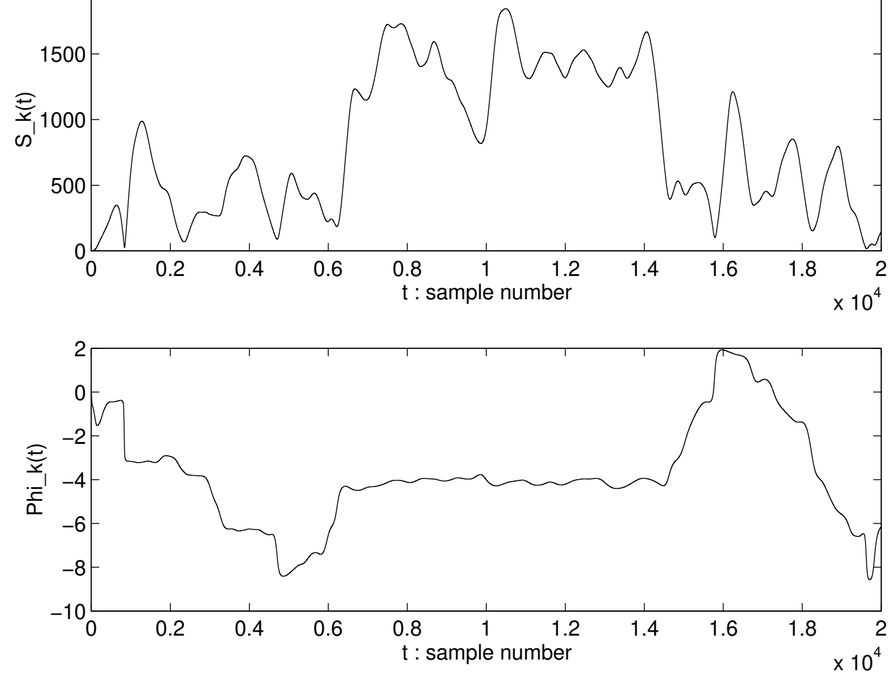

0
−20
−40
−60
−80
−100
−120
−140
−160
Amplitude(dB)
10^0
10^1
10^2
10^3
10^4
Frequency(Hz)

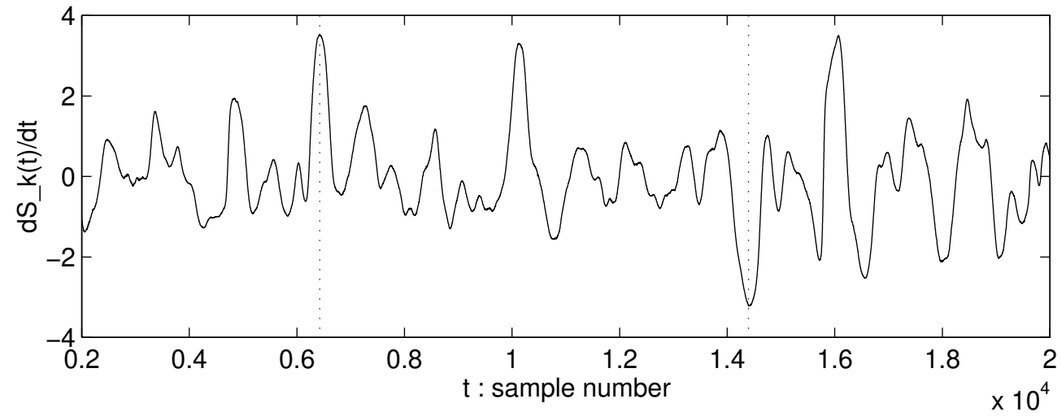

|GT(f)|
f1 (t)
k-1
k
k+1
Ak (t)
B k-1 (t)
Bk (t)
B k+1 (t)
f2 (t)
A k-1 (t)
A k+1 (t)
f 0
f [Hz]

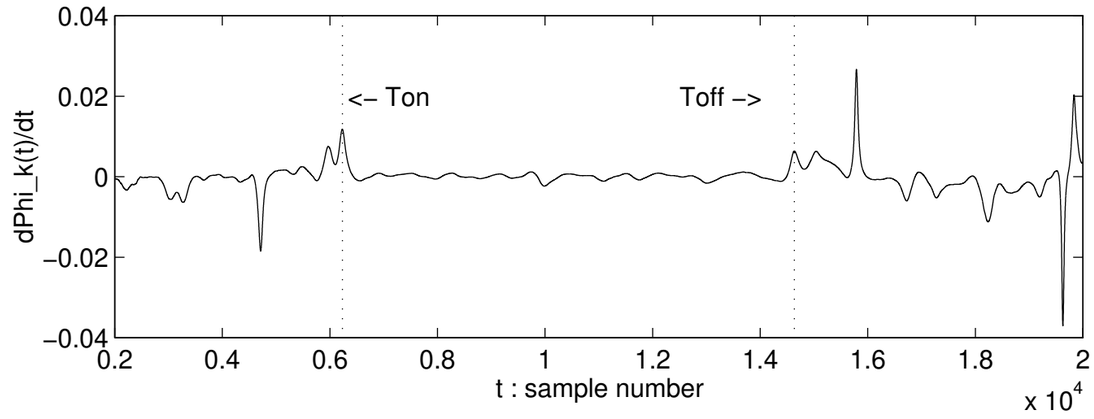

x 10^4
1
0.5
0
−0.5
−1
f1(t)
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
t : sample number
x 10^4
x 10^4
1
0.5
0
−0.5
−1
f21(t)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
t : sample number
x 10^4

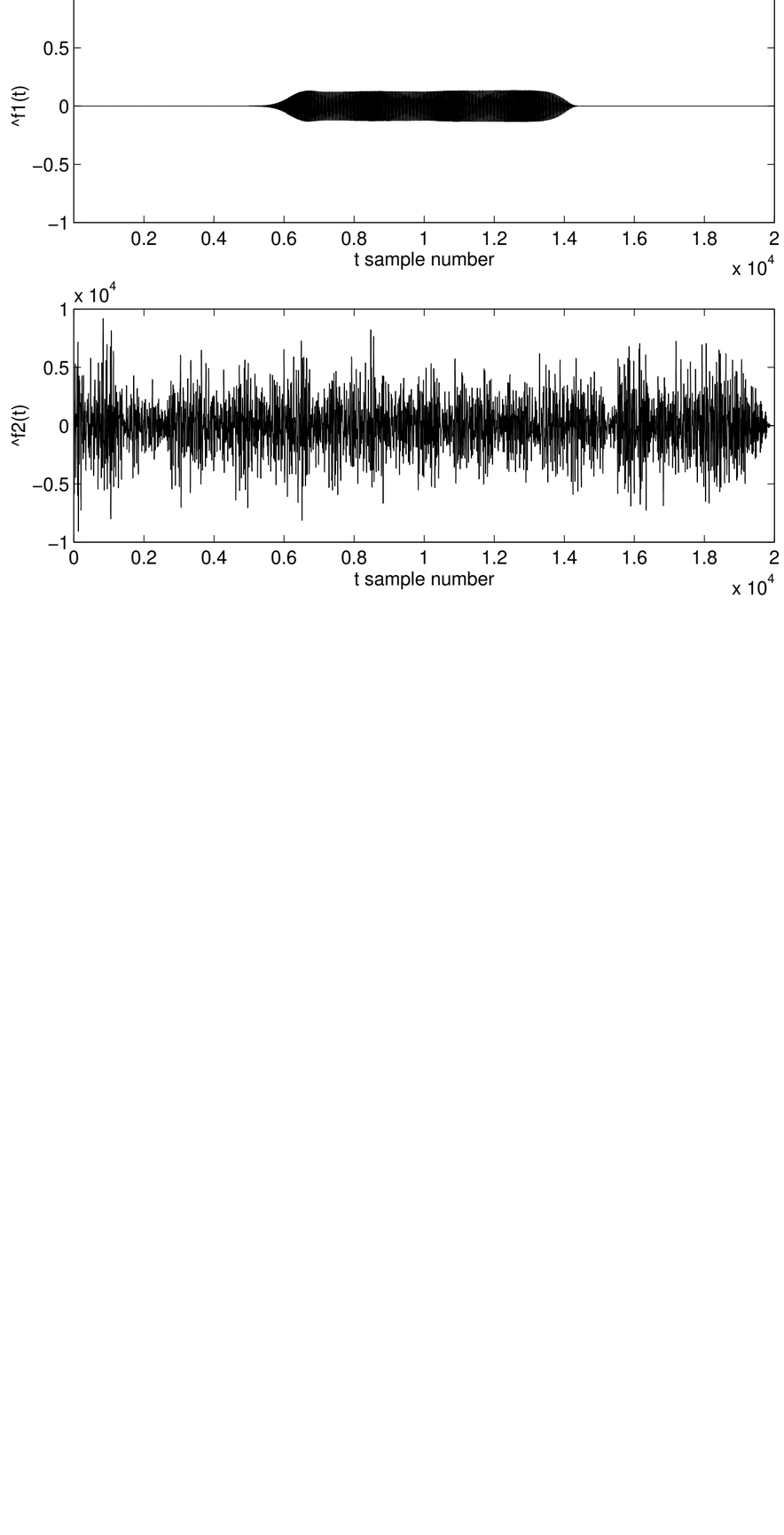

x 10^4
1
0.5
0
−0.5
−1
f22(t)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
t : sample number
x 10^4

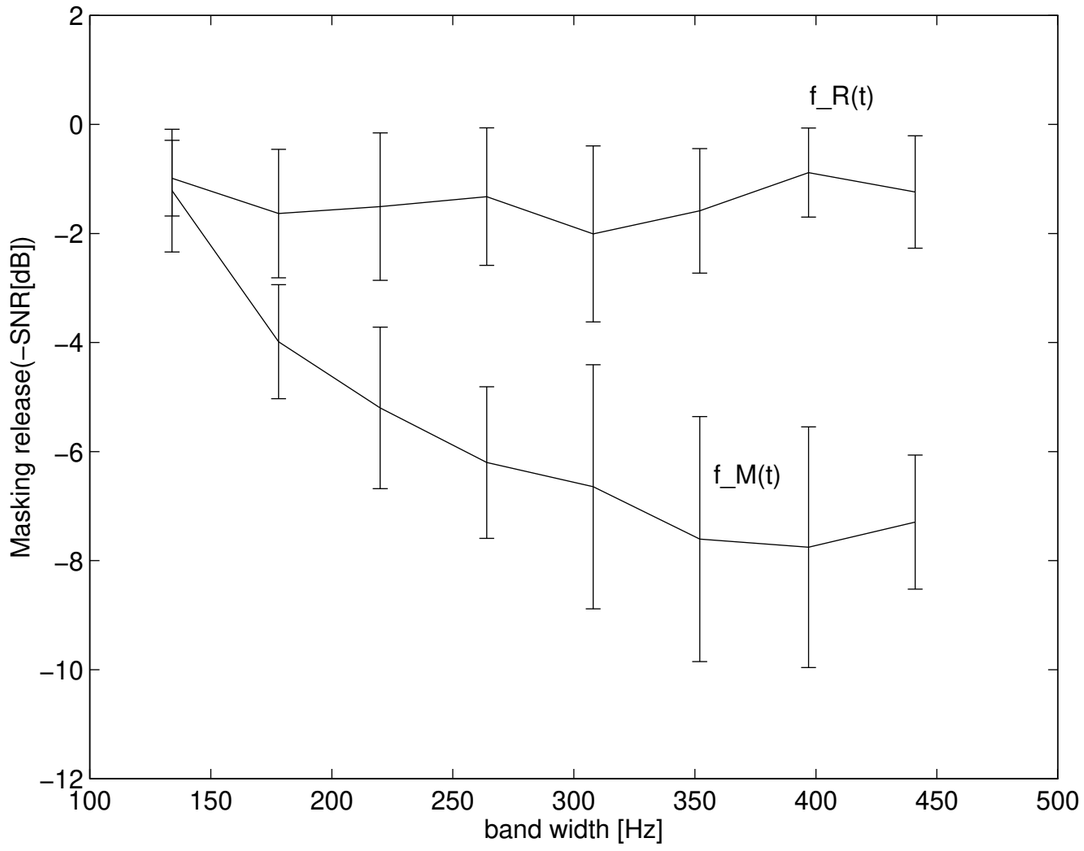

x 10^4
1.5
1
0.5
0
−0.5
−1
f_M(t)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
t : sample number
x 10^4
x 10^4
1
0.5
0
−0.5
−1
f_R(t)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
t : sample number
x 10^4


S_k(t)
0
500
1000
1500
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
t : sample number
x 10^4
Phi_k(t)
2
0
−2
−4
−6
−8
−10
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
t : sample number
x 10^4


dS_k(t)/dt
t : sample number
x 10^4


dPhi_k(t)/dt
<– Ton
Toff –>
t : sample number
x 10^4


0.5
0
−0.5
−1
^f1(t)
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
t sample number
x 10^4
x 10^4
1
0.5
0
−0.5
−1
^f2(t)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
t sample number
x 10^4


f_R(t)
f_M(t)
Masking release(−SNR[dB])
band width [Hz]
2
0
−2
−4
−6
−8
−10
−12
100
150
200
250
300
350
400
450
500