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

Masashi Unoki
School of Information Science, JAIST
1-1 Asahidai, Tatsunokuchi,
Ishikawa 923-12
Japan

Masato Akagi
School of Information Science, JAIST
1-1 Asahidai, Tatsunokuchi,
Ishikawa 923-12
Japan

Abstract

This paper presents a method of extracting the desired signal from a noise-added signal as a model of acoustic source segregation. Using physical constraints related to the four regularities proposed by Bregman, the proposed method can solve the problem of segregating two acoustic sources. These physical constraints correspond to the regularities, which we have translated from qualitative conditions into quantitative conditions. Three simulations were carried out using the following signals: (a) noise-added AM complex tone, (b) mixed AM complex tones, and (c) noisy synthetic vowel. The performance of the proposed method has been evaluated using two measures: precision, that is, likely SNR, and spectrum distortion (S-D). As results using the signals (a) and (b), the proposed method can extract the desired AM complex tone from noise-added AM complex tone or mixed AM complex tones, in which signal and noise exist in the same frequency region. In particular, the average of the reduced SD is about 20 dB. Moreover, as the result using the signal (c), the proposed method can also extract the speech signal from noisy speech.

1 Introduction

Recently, the term "Auditory Scene Analysis: ASA" has become widely known due to Bregman's book[Bregman, 1990]. ASA is understanding a real environment using acoustic events. Although the real environment, that we experience everyday, consists of speech, noise and reflection, simultaneously, it seems that the human auditory system can solve the problem of ASA. But, in solving the problem of ASA using acoustic signals received from the same environment, a unique solution can not be derived without constraints on acoustic sources and the real environment.

Bregman reported that, for solving 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 taken in the acoustic event[Bregman, 1993].

We think that, by translating these heuristic regularities into physical constraints and by using these physical constraints, it is possible to solve the problem of computational auditory scene analysis. As a first step, if it is possible to solve an acoustic source segregation problem, where the sounds required by the listener are extracted selectively while the other sounds are rejected, this solution can be used not only to construct a preprocessor for a robust speech recognition system but also to simulate cocktail party effects. And, it seems that the solution can be a computational model of auditory phenomena such as Co-modulation Masking Release (CMR).

On the one hand, there are two types of typical models of auditory segregation using some of the four regularities, based on either bottom-up or top-down processes. An example of the former type is Brown and Cooke's segregation model based on acoustic events[Brown, 1992; Cooke, 1993]. And as for the later type, there are Ellis' segregation model based on psychoacoustic grouping rules[Ellis, 1994] and Nakatani *et al.*'s stream segregation agents[Nakatani *et al.*, 1994]. All these segregation models use regularities (i) and (iii), and an amplitude (or power) spectrum as the acoustic feature. Thus they can not extract the desired signal from a noisy signal completely when the signal and noise exist in the same frequency region. And, if the power of background noise increases, it seems that these proposed models can not extract the desired signal with high precision.

In contrast, we have discussed the need for using not only the amplitude spectrum but also the phase spectrum, for completely extracting the desired signal from a noisy signal in which signal and noise exist in the same frequency region[Unoki *et al.*, 1997]. We have proposed a method for solving the problem of segregating a sinusoidal signal from noise-added signal, using physical constraints related to regularities (ii) and (iv). As a result of computer simulations, it was found that the proposed model can segregate a sinusoidal signal from noise-added signal. 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[Unoki *et al.*, 1997].

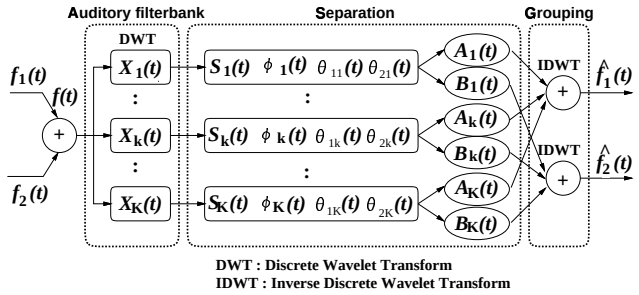

Figure 1: Auditory segregation model.

In this paper, we present a method for extracting the desired signal from noisy signal by using physical constraints related to regularities (i) – (iv), as an auditory segregation model. In particular, we consider that the problem of extracting the desired signal from the following signals: (a) noise-added AM complex tone, (b) mixed AM complex tones, and (c) noisy synthetic vowel.

2 Auditory segregation model

The auditory segregation model shown in Fig. 1 consists of three parts: (a) auditory filterbank, (b) separation, and (c) grouping. The auditory filterbank is constructed using a gammatone filter as an "analyzing wavelet". The separation block uses physical constraints related to heuristic regularities (ii) and (iv). The grouping block uses physical constraints related to heuristic regularities (i) and (iii), and signal reconstruction in the grouping block is done with the inverse wavelet transform. In this model, the separation block follows the formulation of the problem of segregating two acoustic sources.

2.1 Formulation of the problem of segregating two acoustic sources

In this paper, we define the problem of segregating two acoustic sources as "the segregation of the mixed signal into original signal components, where mixed signal is composed of two signals generated by any two acoustic sources". The problem of segregating two acoustic sources is formulated as follows.

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

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

where $f_1(t)$ is the desired signal and $f_2(t)$ is a noise. The observed signal $f(t)$ is decomposed into its frequency components by an auditory filterbank. 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. Since the output of the k-th channel $X_k(t)$ is represented by

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

where

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

and

$$\phi_k(t) = \tan^{-1}\left(\frac{A_k(t)\sin\theta_{1k}(t) + B_k(t)\sin\theta_{2k}(t)}{A_k(t)\cos\theta_{1k}(t) + B_k(t)\cos\theta_{2k}(t)}\right), \tag{6}$$

then the amplitude envelopes 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

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

respectively, where $\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$ and $\theta_k(t) \neq n\pi, n \in \mathbf{Z}$. Thus, if the four parameters, $S_k(t)$, $\phi_k(t)$, $\theta_{1k}(t)$, and $\theta_{2k}(t)$ are calculated, $A_k(t)$ and $B_k(t)$ can be determined by the above equations. Finally, $f_1(t)$ and $f_2(t)$ can be reconstructed by grouping constraints. $\hat{f}_1(t)$ and $\hat{f}_2(t)$ are the reconstructed $f_1(t)$ and $f_2(t)$, respectively.

In this paper, we assume $\theta_{1k}(t) = 0$ and $\theta_k(t) = \theta_{2k}(t)$. Moreover, we consider the problem of segregating two acoustic sources in which the localized $f_1(t)$ is added to $f_2(t)$.

2.2 Auditory filterbank

Firstly, we describe the wavelet transform and the inverse wavelet transform to design an auditory filterbank.

If $\psi \in L^2(\mathbf{R})$ satisfies the "admissibility" condition:

$$D_\psi := \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty, \tag{9}$$

where $\hat{\psi}$ is Fourier transform of ψ, then ψ is called a "basic wavelet". Relative to every basic wavelet ψ, the integral wavelet transform on $L^2(\mathbf{R})$ is defined by

$$\tilde{f}(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t)\overline{\psi\left(\frac{t-b}{a}\right)} dt, \tag{10}$$

where a is the "scale parameter", b is the "shift parameter", and $a, b \in \mathbf{R}$ with $a \neq 0$. In addition, under this additional assumption, it follows that $\hat{\psi}$ is a continuous function, so that finiteness of D_ψ in Eq. (9) implies $\hat{\psi}(0) = 0$, or equivalently, $\int_{-\infty}^{\infty} \psi(t)dt = 0$.

If $\psi(t)$ is a basic wavelet, then the inverse wavelet transform exist for all t as follows:

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

Moreover, if we let $\psi(t)$ be a complex basic wavelet, then the integral wavelet transform can be represented by

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

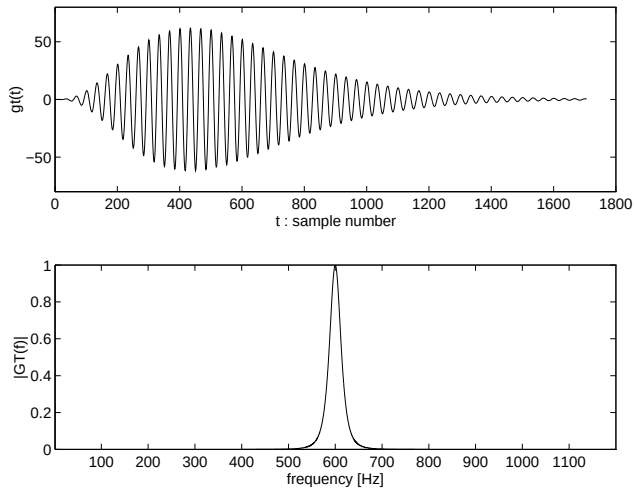

Figure 2: Impulse response and amplitude characteristics of the gammatone filter ($f_0 = 600$ Hz, $N = 4$, $b_f = 22.99$).

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

Secondly, to construct an auditory filterbank, we use the gammatone filter as an analyzing wavelet. The gammatone filter is an auditory filter designed by Patterson[Patterson *et al.*, 1994], and simulates the response of the basilar membrane. The impulse response of the gammatone filter is given by

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

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

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

where $GT(f)$ is the Fourier transform of $gt(t)$. The characteristics of the gammatone filter are shown in Fig. 2. To determine phase information, we extend the impulse response of the gammatone filter, which is a basic wavelet. This basic wavelet is represented by

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

using the Hilbert transform. This analyzing wavelet satisfies the admissibility condition approximately, because $GT(0) \approx 0$.

Finally, an auditory 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. This auditory filterbank is implemented on computer, using a discrete wavelet transform with the following conditions: sampling frequency $f_s = 20$ kHz, the scale parameter $a = \alpha^p, -K/2 \leq p \leq K/2, \alpha = 10^{2/K}$, 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.

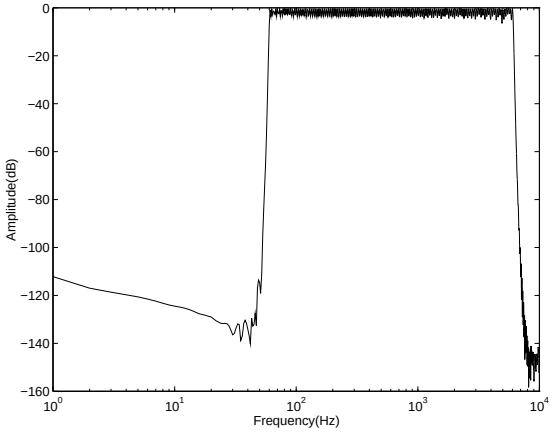

Figure 3: Frequency characteristics of the wavelet filterbank.

3 Calculation of the four physical parameters

3.1 Calculation of $S_k(t)$ and $\phi_k(t)$

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

Lemma 1 *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 in [Unoki *et al.*, 1997]. □

3.2 Calculation of $\theta_k(t)$

In this paper, we assume $\theta_{1k}(t) = 0$ and $\theta_k(t) = \theta_{2k}(t)$. Therefore, we must know the input phase $\theta_k(t)$. The input phase $\theta_k(t)$ can be determined by applying three physical constraints derived from regularities(ii) and (iv) as follows.

Firstly, we use regularity (ii). This regularity means that "a single sound tends to change its properties smoothly and slowly (gradualness of change)". We consider this regularity as the following physical constraint, to apply it to the amplitude envelope $A_k(t)$.

Physical constraint 1 *Temporal differentiation of the amplitude envelope $A_k(t)$ must be represented by Rth-order differentiable polynomial $C_{k,R}(t)$ as follows:*

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

□

A general solution of the input phase $\theta_k(t)$ is determined by solving the linear differential equation obtained by applying **Physical constraint 1** to Eq. (7).

Lemma 2 *A general solution of the input phase $\theta_k(t)$ is determined by*

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin\phi_k(t)}{S_k(t)\cos\phi_k(t) + C_k(t)}\right), \qquad (19)$$

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

Therefore, if $C_k(t)$ is determined, then $\theta_k(t)$ is uniquely determined by Eq. (19). In this paper, we estimate $C_k(t)$ using the Kalman filter.

Estimation of $C_k(t)$ using the Kalman filter

We formulate the problem of estimating $C_k(t)$ by using the Kalman filter.

A complex representation of the output of the kth channel $X_k(t)$ represented by Eq. (4) is the wavelet transform given by Eq. (10) as follows.

$$\begin{aligned} X_k(t) &= S_k(t)e^{j(\omega_k t+\phi_k(t))} \\ &:= \tilde{f}(a,b), \quad a = \alpha^{k-\frac{K}{2}}, b = t_m, \end{aligned} \qquad (20)$$

where $t_m = m/f_s, m = 0, 1, \cdots, M$. From Eq. (1), this is expressed as the sum of the wavelet transforms of $f_1(t)$ and $f_2(t)$. Hence,

$$\tilde{f}(\alpha^{k-\frac{K}{2}}, t_m) = \tilde{f}_1(\alpha^{k-\frac{K}{2}}, t_m) + \tilde{f}_2(\alpha^{k-\frac{K}{2}}, t_m) \qquad (21)$$

where

$$\tilde{f}_1(\alpha^{k-\frac{K}{2}}, t_m) = A_k(t_m)e^{(j\omega_k t_m+\theta_{1k}(t))} \qquad (22)$$

and

$$\tilde{f}_2(\alpha^{k-\frac{K}{2}}, t_m) = B_k(t_m)e^{(j\omega_k t_m+\theta_{2k}(t))}. \qquad (23)$$

On the other hand, from Eqs. (18) and (19), we obtain the following relation.

$$C_k(t) = -A_k(t). \qquad (24)$$

Suppose that a displacement of $C_k(t)$ in discrete time t_m is represented by

$$C_k(t_{m+1}) = C_k(t_m)\Delta C_k + w_m, \qquad (25)$$

where

$$\Delta C_k = 1 + \frac{C_k(t_m) - C_k(t_{m-1})}{C_k(t_m)\cdot f_s}. \qquad (26)$$

That is, $C_k(t_{m+1})$ is represented by $C_k(t_m)$ times ΔC_k, and represented-error w_m follows a white Gaussian probability process with average 0 and variance σ_w.

In this paper, the problem is to estimate unknown function $C_k(t)$ from the observed information $X_k(t)$.

It is required to represent probability system composed of state equation determined by Eq. (21) and the observation equation, to apply the Kalman filter to the estimation problem. If the observed signal is $\mathbf{y}_m = \tilde{f}(\alpha^{k-\frac{k}{2}}, t_m)$, state variable is $\mathbf{x}_m = -C_k(t)$, observed noise is $\mathbf{v}_m = \tilde{f}_2(\alpha^{k-\frac{k}{2}}, t_m)$, and system noise is

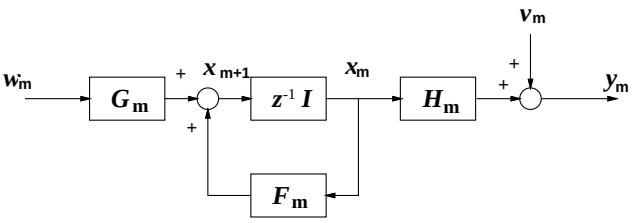

Figure 4: Basic system of the Kalman filter.

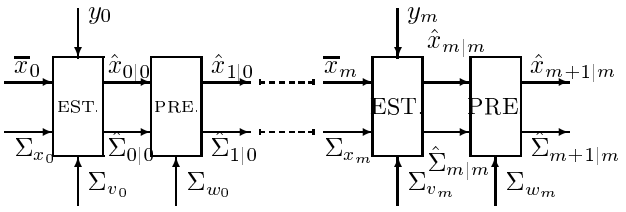

Figure 5: Algorithm for the Kalman filter. "EST." and "PRE." denote "estimation" and "prediction", respectively.

$\mathbf{w}_m = w_m$, then, Eqs. (25) and (21) can be represented by complex probability system as follows.

$$\mathbf{x}_{m+1} = \mathbf{F}_m\mathbf{x}_m + \mathbf{G}_m\mathbf{w}_m \quad (\textbf{state}) \qquad (27)$$

$$\mathbf{y}_m = \mathbf{H}_m\mathbf{x}_m + \mathbf{v}_m \quad (\textbf{observation}), \qquad (28)$$

where state transition matrix $\mathbf{F}_m = \Delta C_k$, observation matrix $\mathbf{H}_m = e^{j\omega_k t_m}$, and driving matrix $\mathbf{G}_m = -1$. These equations are called the "basic system" and are shown in Fig. 4. A complex Kalman filter is represented by the following equations, and is applied to the estimation problem shown in Fig. 5.

1. **Filtering equation**

$$\hat{\mathbf{x}}_{m|m} = \hat{\mathbf{x}}_{m|m-1} + \mathbf{K}_m(\mathbf{y}_m - \mathbf{H}_m\hat{\mathbf{x}}_{m|m-1}) \qquad (29)$$

$$\hat{\mathbf{x}}_{m+1|m} = \mathbf{F}_m\hat{\mathbf{x}}_{m|m} \qquad (30)$$

2. **Kalman gain**

$$\mathbf{K}_m = \frac{\hat{\mathbf{\Sigma}}_{m|m-1}\mathbf{H}_m^{*T}}{\mathbf{H}_m\hat{\mathbf{\Sigma}}_{m|m-1}\mathbf{H}_m^{*T} + \mathbf{\Sigma}_{v_m}} \qquad (31)$$

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

$$\hat{\mathbf{\Sigma}}_{m|m} = \hat{\mathbf{\Sigma}}_{m|m-1} - \mathbf{K}_m\mathbf{H}_m\hat{\mathbf{\Sigma}}_{m|m-1} \qquad (32)$$

$$\hat{\mathbf{\Sigma}}_{m+1|m} = \hat{\mathbf{F}}_m\mathbf{\Sigma}_{m|m}\mathbf{F}_m^{*T} + \mathbf{G}_m\mathbf{\Sigma}_{w_m}\mathbf{G}_m^{*T} \qquad (33)$$

Initial values of parameters are as follows: $\hat{\mathbf{x}}_{0|-1} = 0$, $\hat{\mathbf{\Sigma}}_{0|-1} = S_k(t_0)$, $\hat{\mathbf{\Sigma}}_{w_m} = 0.01$, and $\hat{\mathbf{\Sigma}}_{v_m}$ is the covariance of $\tilde{f}_2(\alpha^{k-\frac{K}{2}}, t_m)$. We remark that $\hat{\mathbf{\Sigma}}_{v_m}$ is given by the variance of $X_k(t_m)$ for the duration in which only $f_2(t)$ exists.

In this manner, the minimum value of the estimation $\hat{C}_k(t)$ and the estimated-error $P_k(t)$ are determined by

$$\hat{C}_k(t) = -|\hat{\mathbf{x}}_{m|m}| \qquad (34)$$

and

$$P_k(t) = |\hat{\mathbf{\Sigma}}_{m|m}|. \tag{35}$$

Although a unique solution for $\theta_k(t)$ is obtained with the minimum value of the estimated $\hat{C}_k(t)$, $A_k(t)$ obtained by the estimated $\theta_k(t)$ does not necessarily satisfy this "smoothness" of $A_k(t)$. So, we define the smoothness of $A_k(t)$ using the following physical constraint.

Definition of the smoothness using spline interpolation

Suppose that $\hat{A}_k(t)$ is the amplitude envelope of $f_1(t)$ given by any unknown function $C_k(t)$, and $t_1, t_2, \cdots, t_i$ are within the opened-duration (t_a, t_b), where $t_a < t_1 < \cdots < t_i < t_b$. In addition, suppose that $\hat{A}_{k,i} := \hat{A}_k(t_i)$ is the value of the amplitude envelope at time t_i. To determine the smoothest interpolation function $A_k(t_i) = \hat{A}_{k,i}, i = 1, 2, \cdots, I$ means that we determine the interpolation function such that integral $\sigma = \int_{t_a}^{t_b} [A_k^{(r)}(t)]^2 dt$ is the smallest, where $A_k(t)$ is defined in the closed-duration $[t_a, t_b]$ and is rth-order differentiable.

We consider the smoothness in regularity (ii) as the following physical constraint, to define the smoothness of the amplitude envelope $A_k(t)$.

Physical constraint 2 *Suppose that the amplitude envelope $A_k(t)$ is defined in the closed-duration $[t_a, t_b]$ and satisfies* **Physical constraint 1**. *If $A_k(t)$ is as smooth as possible, then the following integral must be minimized:*

$$\sigma = \int_{t_a}^{t_b} [A_k^{(R+1)}(t)]^2 dt \quad \Rightarrow \min. \tag{36}$$

□

According to **Physical constraint 2**, the smoothest of the interpolation function is the $(2R+1)$th-order spline function. This spline function exists uniquely.

By considering the relationship between $A_k(t)$ and $C_k(t)$ from Eqs. (7) and (19), we can interpret **Physical constraint 2** in order to determine $C_k(t)$, which is interpolated by using the spline function within the estimated-error region:

$$\hat{C}_k(t) - P_k(t) \le C_k(t) \le \hat{C}_k(t) + P_k(t). \tag{37}$$

Therefore, by calculating the candidates of $C_k(t)$ interpolated using the spline function within the estimated error, and by calculating a correct solution from the candidates of $C_k(t)$, the smoothest $A_k(t)$ can be determined uniquely. For example, $C_k(t)$ as interpolated by the spline interpolation function in time t_i is shown in Fig. 6. In this figure, each candidate of $C_k(t)$ is determined by fixing $C_k(t_1), \cdots, C_k(t_{i-1})$ for $t_1, \cdots, t_{i-1}$, and by interpolating $C_k(t)$ for changes in $C_k(t_i)$, where $\hat{C}_k(t_i) - P_k(t_i) \le C_k(t_i) \le \hat{C}_k(t_i) + P_k(t_i)$.

In this paper, we use the cubic spline function ($R = 1$). The interpolated duration is $\Delta t = 15/f_0$.

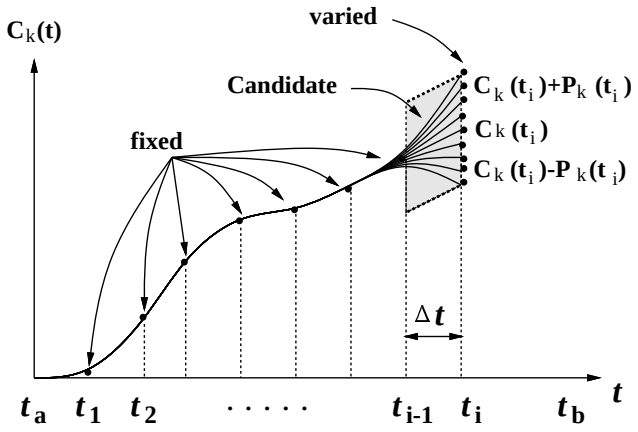

Figure 6: Candidates for $C_k(t)$ interpolated by the spline function

Determination of $C_k(t)$ using correlation between the amplitude envelopes

Finally, we use regularity (iv) to narrow down the candidates for $C_k(t)$, which is interpolated by spline function. Regularity (iv) means 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"[Bregman, 1993]. Therefore, we consider this regularity as the following physical constraint.

Physical constraint 3 *The normalized amplitude envelope of the output of the kth channel must approximate that of ℓth channel as follows:*

$$\frac{A_k(t)}{\|A_k(t)\|} \approx \frac{A_\ell(t)}{\|A_\ell(t)\|}, \qquad k \neq \ell. \tag{38}$$

□

To select an optimal function $C_k(t)$ when the correlation between $A_k(t)$ and $A_\ell(t)$ becomes maximum at any $C_k(t)$ within the estimated-error, we interpret **Physical constraint 3** as follows:

$$\max_{\hat{C}_k - P_k \le C_k \le \hat{C}_k + P_k} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||}, \tag{39}$$

where $\hat{A}_k(t)$ is the amplitude envelope given by interpolated $C_k(t)$, and $\hat{\hat{A}}_k(t)$ is the amplitude envelope in other channel. We explain the amplitude envelope $\hat{\hat{A}}_k(t)$ in the next section.

Hence, $\theta_k(t)$ is uniquely determined using the optimized $C_k(t)$ from Eq. (19).

4 Segregation and Grouping

In this section, we describe the grouping constraints. The aim of grouping constraints is to extract the desired signal from the noise-added signal using regularities (i) and (iii) proposed by Bregman. Therefore, the grouping block takes a solution for the problem of segregating two acoustic sources and applies to $X_k(t)$, in which two

acoustic signals exist in the same time region. In other words, it applies the solution to $X_k(t)$, if either of the two physical constraints are satisfied as follows.

As a first regularity, we use regularity (iii). This regularity means that "when a body vibrates with a repetitive period, its vibrations give rise to an acoustic pattern in which the frequency components are multiples of a common fundamental". In order to use regularity (iii), we consider it as the following physical constraint.

Physical constraint 4 *Suppose that F_0 is the fundamental frequency, and N_{F_0} is the order of harmonics. If the harmonic component exists in $X_\ell(t)$, then the channel number ℓ must satisfy*

$$\ell = \frac{K}{2} - \left\lceil \frac{\log(n \cdot F_0/f_0)}{\log \alpha} \right\rceil \pm 1, \quad n = 1, 2, \cdots, N_{F_0}, \quad (40)$$

where α is the scale parameter. □

As a second regularity, we use regularity (i). This regularity means that "unrelated sounds seldom start or stop at exactly the same time". Therefore, we consider this regularity as the following physical constraint.

Physical constraint 5 *Suppose that $T_{\rm S}$ and $T_{\rm E}$ are onset and offset of $f_1(t)$, which is generated by one acoustic source. If an acoustic event obtained by a channel is component of $f_1(t)$, then onset $T_{k,\rm on}$ and offset $T_{k,\rm off}$ determined for the same channel must satisfy*

$$|T_{\rm S} - T_{k,\rm on}| \le 50 \text{ ms} \quad (41)$$

and

$$|T_{\rm E} - T_{k,\rm off}| \le 100 \text{ ms}. \quad (42)$$

□

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 $|\frac{dS_k(t)}{dt}|$.

In addition, onset $T_{\rm S}$ and offset $T_{\rm E}$ are obtained by determining $T_{k,\rm on}$ and $T_{k,\rm off}$ of the channel corresponding to the fundamental frequency F_0.

Moreover, the amplitude envelope $\hat{\hat{A}}_k(t)$ in **Physical constraint 3**, is determined by

$$\hat{\hat{A}}_k(t) = \frac{1}{N_{F_0}} \sum_{\ell \in \mathbf{L}} \frac{\hat{A}_\ell(t)}{\|\hat{A}_\ell(t)\|}, \quad (43)$$

where $\mathbf{L}$ is the set of ℓ satisfying Eq. (40).

The algorithm for solving the problem of segregating two acoustic sources using physical constraints related to the four regularities is shown in Fig. 7.

```
decompose f(t) into its frequency components using the
  wavelet filterbank (wavelet transform) as Eq. (4);
for k := 1 to K do
    θ_1k(t) = 0 and θ_k(t) = θ_2k(t);
    determine S_k(t) and φ_k(t) from Lemma 1;
    determine onset T_k,on and offset T_k,off;
    the segregated duration is T_k,on ≤ t ≤ T_k,off;
    if Physical constraint 4 or 5 is satisfied then
        estimate C_k(t) using the Kalman filter;
        determine the interpolated duration;
        let I be the number of the interpolated samples;
        for i = 1 to I do
            determine the candidates for C_k(t), which
              interpolated by the spline function within
            Ĉ_k(t_i) − P_k(t_i) ≤ C_k(t_i) ≤ Ĉ_k(t_i) + P_k(t_i);
            determine θ̂_k(t) from Eq. (19);
            determine Â_k(t) from Eq. (7);
            determine Â^_k(t) from Eq. (43);
            determine Corr(Â_k(t), Â^_k(t)) from Eq. (39);
        end
        determine C_k(t) when Corr(Â_k(t), Â^_k(t))
          becomes a maximum within the estimated
          -error;
        determine θ_k(t) from Eq. (19);
    else
        set A_k(t) = 0, B_k(t) = S_k(t) and θ_k(t) = φ_k(t);
    end
    determine A_k(t) and B_k(t) from Eqs. (7) and (8);
    determine each frequency components of f_1(t)
      and f_2(t) from Eqs. (2) and (3);
end
reconstruct f̂_1(t) and f̂_2(t) using the wavelet filterbank
  (inverse wavelet transform) from Eqs. (7) and (8);
```

Figure 7: Segregation algorithm

5 Simulations

We have 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 it. These simulations are composed as follows:

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

We use two types of measures to evaluate the performance of segregation using the proposed method.

One is the power ratio in terms of the amplitude envelope $A_k(t)$, i.e., likely SNR. The aim of using this measure is 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

$$\text{Precision}(k) := 10 \log_{10} \frac{\int_0^T A_k^2(t) dt}{\int_0^T (A_k(t) - \hat{A}_k(t))^2 dt}, \quad (44)$$

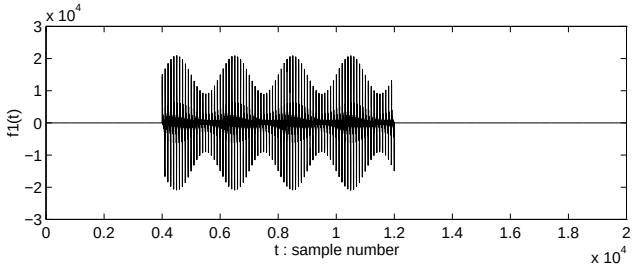

Figure 8: AM complex tone $f_1(t)$.

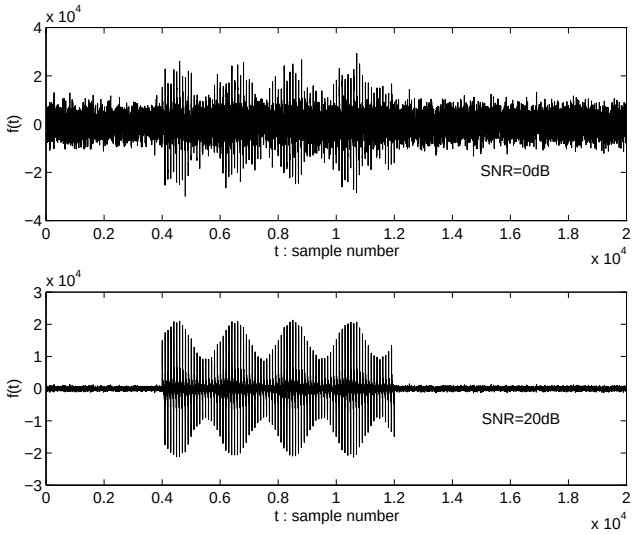

Figure 9: Mixed signals $f(t)$.

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

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

$$\mathrm{SD} := \sqrt{\frac{1}{W}\sum_{\omega}^{W}\left(20\log_{10}\frac{\tilde{F}_1(\omega)}{\tilde{\hat{F}}_1(\omega)}\right)^2}, \qquad (45)$$

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

Reduced SD of $f_1(t)$ is the SD difference between $f(t)$ and $\hat{f}_1(t)$.

5.1 Simulation 1

This simulation assumes that $f_1(t)$ is an AM complex tone as shown in Fig. 8, where $F_0 = 200$ Hz, $N_{F_0} = 10$,

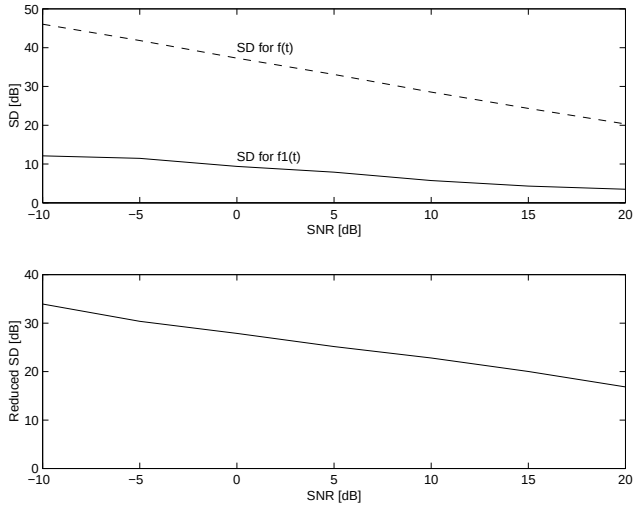

Figure 10: SD for $\hat{f}_1(t)$ and the reduced SD of $\hat{f}_1(t)$.

and whose amplitude envelope is sinusoidal (10 Hz), and $f_2(t)$ is a bandpassed random noise, where bandwidth of about 6 kHz. Seven types of $f(t)$ are used as simulation stimuli, where the SNRs of $f(t)$ are from -10 to 20 dB in 5-dB steps. Mixed signals in cases of SNR= 0 dB and SNR= 20 dB are plotted in Fig. 9.

The simulations were carried out using the seven mixed signals. The average SDs of $f_1(t)$ and $f(t)$, and the mean of the reduced SD of $f_1(t)$ are shown in Fig. 10. Hence, it is possible to reduce the SD by about 20 dB as noise reduction, using the proposed method. For example, when the SNR of $f(t)$ is 20 dB, the proposed method can segregate $A_k(t)$ with a high precision as shown in Fig. 11, and can extract the $\hat{f}_1(t)$ shown in Fig. 12 from the $f(t)$ as shown in Fig. 9. Moreover, when the SNR of $f(t)$ is 0 dB, the proposed method can also segregate $A_k(t)$ as shown in Fig. 13, and can extract the $\hat{f}_1(t)$ shown in Fig. 14 from the $f(t)$ as shown in Fig. 9. Hence, the proposed model can extract the amplitude information of signal $f_1(t)$ from a noise-added signal $f(t)$ with a high precision in which signal and noise exist in the same frequency region.

5.2 Simulation 2

This simulation assumes that $f_1(t)$ is an AM complex tone as the same as Fig. 8 and $f_2(t)$ is another AM complex tone as shown in Fig. 15, where $F_0 = 300$ Hz, $N_{F_0} = 10$, and whose amplitude envelope is sinusoidal (15 Hz). Therefore, harmonics of $f_1(t)$ and $f_2(t)$ in the multiple of 600 Hz, for example, third harmonic of $f_1(t)$ and second harmonic of $f_2(t)$, exist in the same frequency region. Seven types of $f(t)$ are used as simulation stimuli, where the SNRs of $f(t)$ are from -10 to 20 dB in 5-dB steps. Mixed signal in case of SNR= 10 dB is plotted in Fig. 16.

The simulations were carried out using the seven

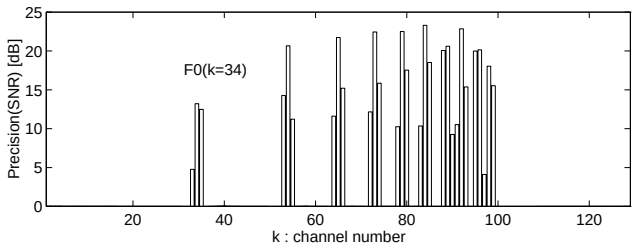

Figure 11: Precision for $A_k(t)$ (SNR= 20 dB).

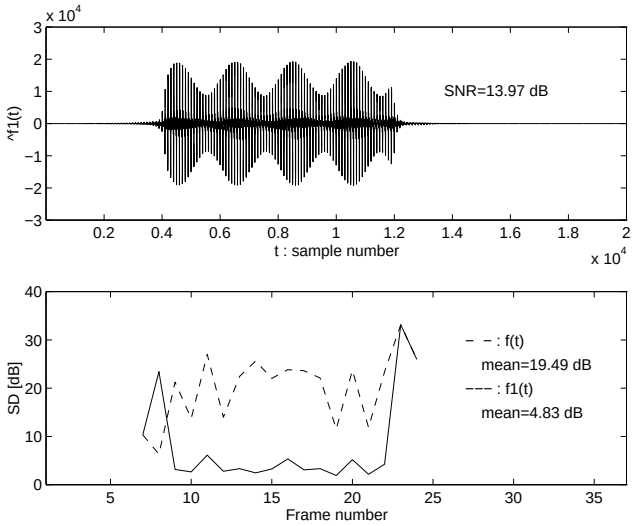

Figure 12: Extraction property for $\hat{f}_1(t)$ (SNR= 20 dB).

mixed signals. The average SDs of $f_1(t)$ and $f(t)$, and the mean of the reduced SD of $f_1(t)$ are shown in Fig. 17. Hence, it is possible to reduce the SD by about 20 dB as noise reduction, using the proposed method. For example, when the SNR of $f(t)$ is 10 dB, the proposed method can segregate $A_k(t)$ with a high precision as shown in Fig. 18, and can extract the $\hat{f}_1(t)$ shown in Fig. 19 from the $f(t)$ as shown in Fig. 16. Hence, just as the result of previous simulations, the proposed model can also extract the amplitude information of signal $f_1(t)$ from a noise-added signal $f(t)$ with a high precision in which two AM complex tones exist in the same frequency region.

5.3 Simulation 3

This simulation assumes that $f_1(t)$ is a synthetic vowel as shown in Fig. 20, where $F_0 = 125$ Hz, $N_{F_0} = 40$, and it is a vowel /a/ synthesized by the LMA, and $f_2(t)$ is a bandpassed random noise, where bandwidth of about 6 kHz. Three types of $f(t)$ are used as simulation stimuli, where the SNRs of $f(t)$ are from 0 to 20 dB in 10-dB steps. Mixed signal in case of SNR= 10 dB is plotted in Fig. 21.

The simulations were carried out using the three mixed signals. The average SDs of $f_1(t)$ and $f(t)$, and the mean

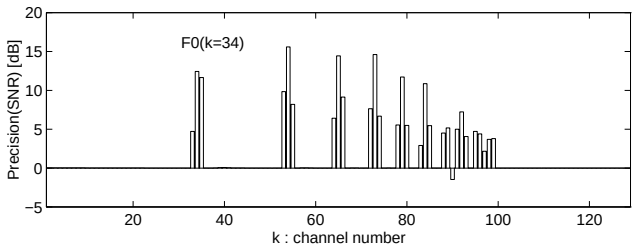

Figure 13: Precision for $A_k(t)$ (SNR= 0 dB).

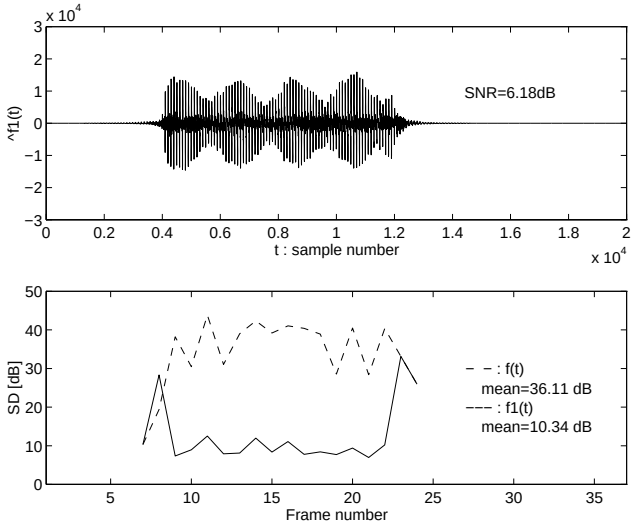

Figure 14: Extraction property for $\hat{f}_1(t)$ (SNR= 0 dB).

of the reduced SD of $f_1(t)$ are shown in Fig. 22. Hence, it is possible to reduce the SD by about 15 dB as noise reduction, using the proposed method. For example, when the SNR of $f(t)$ is 10 dB, the proposed method can segregate $A_k(t)$ with a high precision as shown in Fig. 23, and can extract the $\hat{f}_1(t)$ shown in Fig. 24 from the $f(t)$ as shown in Fig. 21. Therefore, the proposed model can also extract the amplitude information of speech $f_1(t)$ from a noisy speech $f(t)$ with a high precision in which speech and noise exist in the same frequency region. Hence, this method can be applied in case where a speech signal is to be extracted from noisy speech.

6 Conclusion

In this paper, we proposed a method of extracting the desired signal from a noise-added signal, using physical constraints related to the four regularities proposed by Bregman, and by solving the problem of segregating two acoustic sources. We have 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 it. These simulations are:

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

2. Extracting one AM complex tone from mixed AM complex tones.
3. Extracting a speech signal from noisy speech.

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

Future work includes as follows: (1) to determine the input phases $\theta_{k1}(t)$ and $\theta_{2k}(t)$, and (2) to evolve the grouping constraints for the deviation of F_0. If the above subjects are cleared, then the proposed model can be used not only to extend the problems of extracting the desired FM tone from noise-added FM tone and the desired AM-FM tone from noise-added AM-FM tone, but also to extend problems of extracting the desired speech signal from noisy signal in a real environment, such as cocktail party effects.

References

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

[Bregman, 1993] 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.

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

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

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

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

[Unoki *et al.*, 1997] Masashi Unoki and Masato Akagi. "A Method of Signal Extraction from Noise-Added Signal," IEICE, vol. J80-A, no. 3, March 1997(in Japanese).

[Patterson *et al.*, 1994] 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.

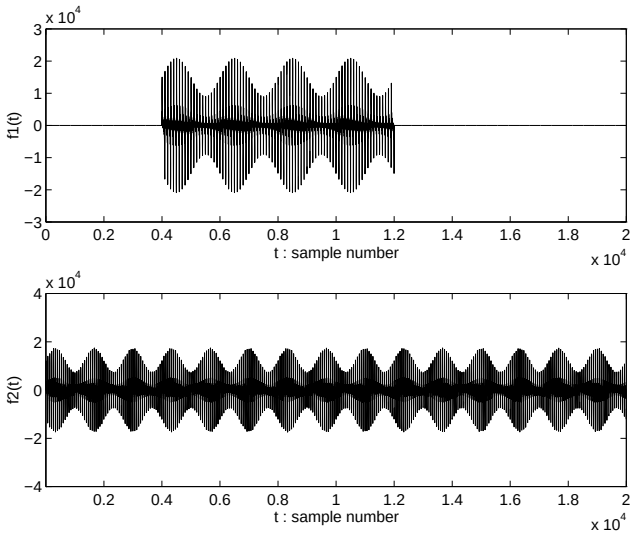

Figure 15: Two AM complex tones, $f_1(t)$ and $f_2(t)$.

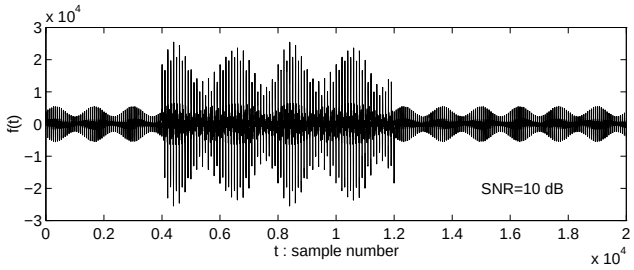

Figure 16: Mixed signals $f(t)$.

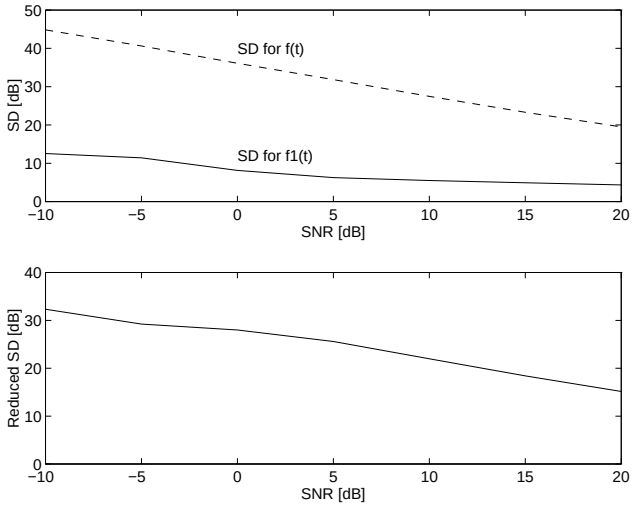

Figure 17: SD for $\hat{f}_1(t)$ and the reduced SD of $\hat{f}_1(t)$.

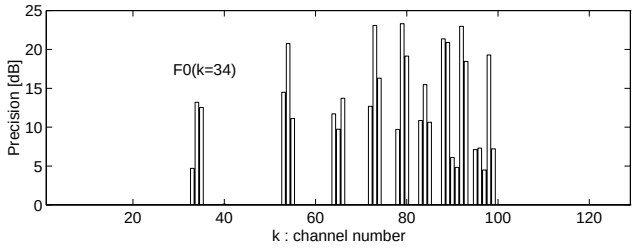

Figure 18: Precision for $A_k(t)$ (SNR= 10 dB).

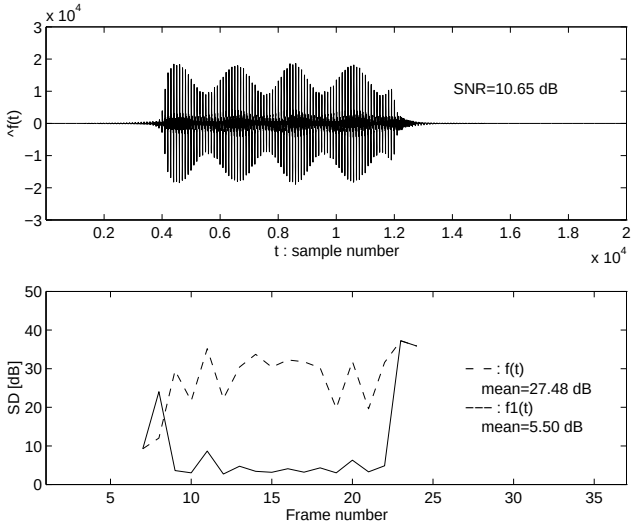

Figure 19: Extraction property for $\hat{f}_1(t)$ (SNR= 10 dB).

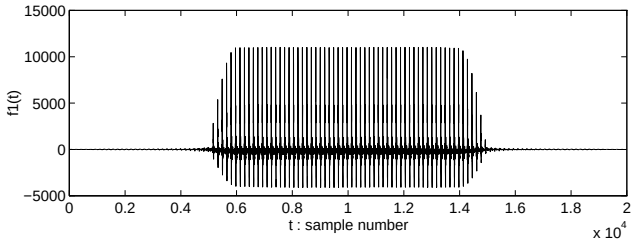

Figure 20: Synthetic vowel $f_1(t)$.

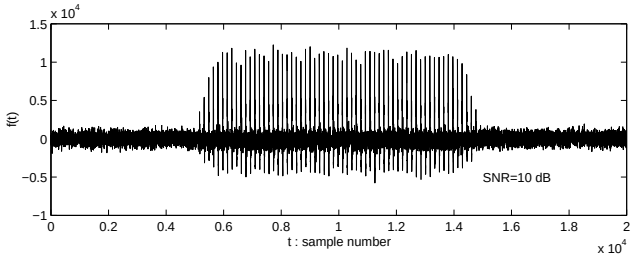

Figure 21: Mixed signal $f(t)$.

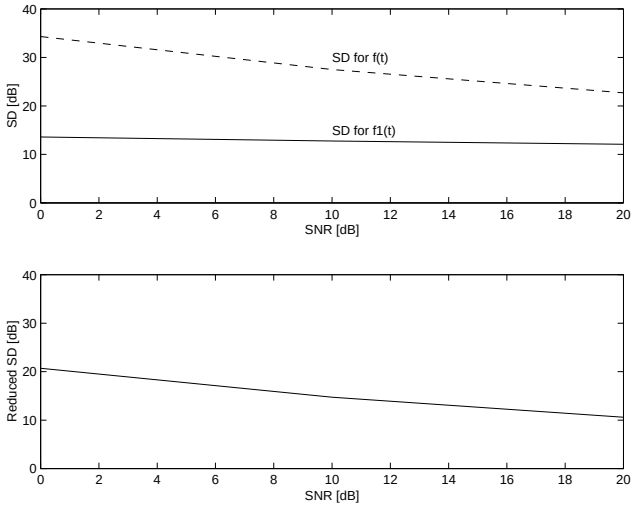

Figure 22: SD for $\hat{f}_1(t)$ and the reduced SD of $\hat{f}_1(t)$.

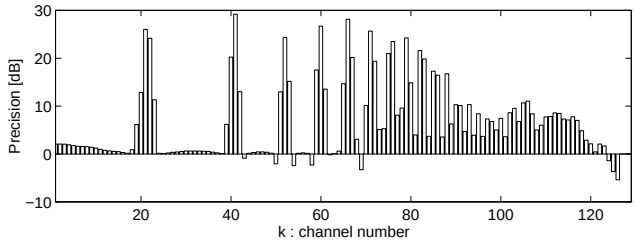

Figure 23: Precision for $A_k(t)$ (SNR= 10 dB).

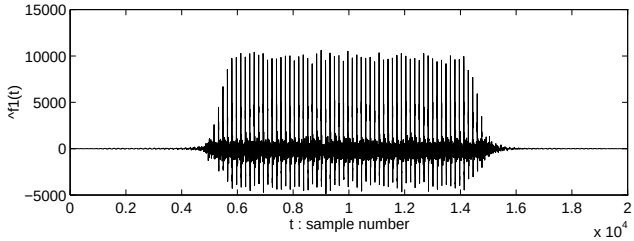

Figure 24: Segregated signal $\hat{f}_1(t)$ (SNR= 10 dB).