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

Masashi UNOKI and Masato AKAGI

6 Feb. 1998
IS-RR-98-0005P

School of Information Science
Japan Advanced Institute of Science and Technology, Hokuriku
Asahidai 1–1, Tatsunokuchi
Nomi, Ishikawa, 923-12, JAPAN
unoki@jaist.ac.jp, akagi@jaist.ac.jp

ISSN 0918-7553

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

Masashi UNOKI and Masato AKAGI
School of Information Science, JAIST
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-12 Japan

Abstract

This paper presents a method of extracting the desired signal from a noisy signal by using physical constraints as a model of acoustic source segregation. Using physical constraints, which are related to the four regularities proposed by Bregman, the proposed method can solve the problem of segregating two acoustic sources. The physical constraints are obtained from the regularities by translating from qualitative to quantitative conditions. Three simulations were carried out using the following signals: (a) a noise-added AM complex tone, (b) a mixed AM complex tones, and (c) noisy synthesized vowel. The performance of the proposed method was evaluated using two measures: precision, that is, likely SNR, and spectrum distortion (SD). We found that using signals (a) and (b), it could extract the desired AM complex tone from a 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 was about 20 dB. Moreover, using signal (c), it could also extract the desired speech signal from noisy speech.

Key words: auditory scene analysis, two acoustic source segregation, gammatone filter, wavelet filterbank

1 Introduction

Recently, the term "Auditory Scene Analysis (ASA)" has become widely known due to Bregman's book [Bregman, 1990]. ASA means understanding a real environment by 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 easily solve the problem of ASA. However, using acoustic signals received from the same environment, it is not possible to derive a unique solution to ASA without constraints on both acoustic sources and the real environment.

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].

We think that by translating these heuristic regularities into physical constraints and using them it should be possible to solve the problem of computational auditory scene

analysis. As the first step, if it is possible to solve an acoustic source segregation problem, where the sounds required by the listener are extracted selectively while 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. Moreover, the solution should be a computational model of auditory phenomena, such as Co-modulation Masking Release (CMR).

There are two main types of 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 examples of the latter type include 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 the amplitude (or power) spectrum as the acoustic feature. Thus they cannot completely extract the desired signal from a noisy signal if the signal and noise exist in the same frequency region. Moreover, as the power of the background noise increases, the precision with which these proposed models can extract the desired signal decreases.

In contrast, 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 signal and noise exist in the same frequency region [Unoki *et al.*, 1997a , Unoki *et al.*, 1997b]. We have proposed a method for segregating a sinusoidal signal from a noisy signal, using physical constraints related to regularities (ii) and (iv), and have demonstrated its ability to do this by computer simulations. If the parameters of this model are set to the human auditory characteristics, it can act as a computational model for Co-modulation Masking Release [Unoki *et al.*, 1997a].

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

2 Auditory segregation model

The auditory segregation model shown in Fig. 1 consists of three blocks: (a) auditory filterbank, (b) a separation block, and (c) a grouping block. 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 Auditory filterbank

First, 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:

$$D_\psi := \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty, \tag{1}$$

where $\hat{\psi}$ is the 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

$$\tilde{f}(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t) \overline{\psi\left(\frac{t-b}{a}\right)} dt, \tag{2}$$

where a is the "scale parameter", b is the "shift parameter", and $a, b \in \mathbf{R}$ with $a \neq 0$. Moreover, under this additional assumption, it follows that $\hat{\psi}$ is a continuous function, so that the finiteness of D_ψ in Eq. (1) implies $\hat{\psi}(0) = 0$, or equivalently, $\int_{-\infty}^{\infty} \psi(t) dt = 0$.

If $\psi(t)$ is a basic wavelet, then for all t there exists the following inverse wavelet transform:

$$f(t) = \frac{1}{D_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{f}(a,b) \psi\left(\frac{t-b}{a}\right) \frac{dadb}{a^2} \tag{3}$$

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

$$\tilde{f}(a,b) = |\tilde{f}(a,b)| e^{j \arg(\tilde{f}(a,b))}, \tag{4}$$

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

Second, 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 it simulates the response of the basilar membrane. Its impulse response is given by

$$gt(t) = At^{N-1} e^{-2\pi b_f t} \cos(2\pi f_0 t), \qquad t \geq 0, \tag{5}$$

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. In addition, amplitude characteristics of the gammatone filter are represented approximately by

$$GT(f) \approx \left[1 + \frac{j(f-f_0)}{b_f}\right]^{-N}, \qquad 0 < f < \infty, \tag{6}$$

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 its impulse response, which is a basic wavelet. This basic wavelet is represented by

$$\psi(t) = At^{N-1} e^{j2\pi f_0 t - 2\pi b_f t}, \tag{7}$$

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 bandpass region from 60 Hz to 6000 Hz, and 128 filters. 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 $p, q \in \mathbf{Z}$ and K is the number of filters. Frequency characteristics of the wavelet filterbank are shown in Fig. 3.

2.2 Formulation of the problem of segregating two acoustic sources

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

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

$$f(t) = f_1(t) + f_2(t), \tag{8}$$

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

$$f_1(t) : A_k(t)\sin(\omega_k t + \theta_{1k}(t)) \tag{9}$$

and

$$f_2(t) : B_k(t)\sin(\omega_k t + \theta_{2k}(t)). \tag{10}$$

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 represented by

$$X_k(t) = S_k(t)\sin(\omega_k t + \phi_k(t)), \tag{11}$$

where

$$S_k(t) = \sqrt{A_k^2(t) + 2A_k(t)B_k(t)\cos\theta_k(t) + B_k^2(t)} \tag{12}$$

and

$$\phi_k(t) = \tan^{-1}\left(\frac{A_k(t)\sin\theta_{1k}(t) + B_k(t)\sin\theta_{2k}(t)}{A_k(t)\cos\theta_{1k}(t) + B_k(t)\cos\theta_{2k}(t)}\right), \tag{13}$$

the amplitude envelopes of the two signals $A_k(t)$ and $B_k(t)$ can be determined by

$$A_k(t) = \frac{S_k(t)\sin(\theta_{2k}(t) - \phi_k(t))}{\sin\theta_k(t)} \tag{14}$$

and

$$B_k(t) = \frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{\sin\theta_k(t)}, \tag{15}$$

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 output phase $\phi_k(t)$ are observable, and if the 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 constraints. $\hat{f}_1(t)$ and $\hat{f}_2(t)$ are the reconstructed $f_1(t)$ and $f_2(t)$, respectively.

In the above formulation, it is difficult to determine the input phases $\theta_{1k}(t)$ and $\theta_{2k}(t)$ uniquely. However, it can be considered that each frequency component of the signal closes to the center frequency of auditory filter if the bandwidth of the auditory filter is narrow and the number of channels is large. Therefore, 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 $f_2(t)$.

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 $\phi_k(t)$ represented by Eqs. (12) and (13) 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.*, 1997a]. □

3.2 Calculation of $\theta_{1k}(t)$ and $\theta_{2k}(t)$

In this paper, we assume $\theta_{1k}(t) = 0$. Therefore, since $\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$, we must find 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.

First, we use regularity (ii), which is the gradualness of change. This regularity means that "a single sound tends to change its properties smoothly and slowly". We consider this regularity as the following physical constraint, in order 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 an R-th-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 $\theta_{2k}(t)$ is determined by solving the linear differential equation obtained by applying Physical constraint 1 to Eq. (14).

Lemma 2 *A general solution of the input phase $\theta_{2k}(t)$ is determined by*

$$\theta_{2k}(t) = \arctan\left(\frac{S_k(t)\sin\phi_k(t)}{S_k(t)\cos\phi_k(t) + C_k(t)} \right), \tag{19}$$

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

Therefore, if $C_k(t)$ is determined, then $\theta_{2k}(t)$ is uniquely determined by Eq. (19). In this paper, we estimate $C_k(t)$ using the Kalman filter.

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

Let us 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. (11) is the wavelet transform given by Eq. (2) as follows.

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

where $t_m = m/f_s, m = 0, 1, \cdots, M$. From Eq. (8), this is expressed as the sum of the wavelet transforms of $f_1(t)$ and $f_2(t)$. Hence,

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

where

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

and

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

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

$$C_k(t) = -A_k(t). \tag{24}$$

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

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

where

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

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 necessary to represent a probability system composed of the state equation determined by Eq. (21) and the observation equation, in order to apply the Kalman filter to the estimation. 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 $\mathbf{w}_m = w_m$, then Eqs. (25) and (21) can be represented by the following complex probability system.

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

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

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**

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

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

2. **Kalman gain**

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

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

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

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

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

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

and

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

Although a unique solution for $\theta_{2k}(t)$ is obtained with the estimated $\hat{C}_k(t)$, $A_k(t)$ obtained by the estimated $\theta_{2k}(t)$ does not necessarily satisfy the "smoothness" of $A_k(t)$. Therefore, we define the smoothness of $A_k(t)$ using the following physical constraint.

3.2.2 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. Determining the smoothest interpolation function $A_k(t_i) = \hat{A}_{k,i}, i = 1, 2, \cdots, I$ means determining the interpolation function such that integral $\sigma = \int_{t_a}^{t_b}[A_k^{(r)}(t)]^2dt$ is the smallest, where $A_k(t)$ is defined in the closed-duration $[t_a, t_b]$ and is r-th-order differentiable.

We consider the smoothness in regularity (ii) as the following physical constraint, in order 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:*

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

□

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

By considering the relationship between $A_k(t)$ and $C_k(t)$ from Eqs. (14) 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:

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

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 $C_k(t)$, we can determine the smoothest $A_k(t)$ 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 $\hat{C}_k(t_i) - P_k(t_i) \leq C_k(t_i) \leq \hat{C}_k(t_i) + P_k(t_i)$.

In this paper, we use the cubic spline function ($R = 1$). The interpolated duration Δt is $15/(f_0 \cdot \alpha^{k-\frac{k}{l}2})$.

3.2.3 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 k-th channel must approximate that of the ℓ-th channel as follows:*

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

□

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

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

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, $\theta_{2k}(t)$ 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 Bregman's regularities (i) and (iii). Therefore, the grouping block applies the solution of the problem of segregating two acoustic sources not to all $X_k(t)$ but to only the $X_k(t)$ in which two acoustic signals exist in the same time region. In other words, if either of the two physical constraints is satisfied, it applies the solution to $X_k(t)$ as follows.

4.1 Estimation for the fundamental frequency

In this paper, the fundamental frequency of the complex tones is estimated using TEMPO (Time-domain Excitation extraction based on a minimum perturbation operator) [Kawahara, 1997] proposed by Kawahara. The TEMPO procedure is to estimate the output of the analysis-filter including the fundamental component from outputs of the constant Q filterbank. Therefore, this procedure can be implemented in the proposed auditory filterbank.

In general, since the fundamental frequency varies temporally, a procedure dealing with temporal variation must be applied when the separation block is done using the grouping constraints. The procedure corresponding to temporal variation of the fundamental frequency is given below.

Let $F_0(t)$ be the fundamental frequency estimated using TEMPO. If we use regularity (ii) for the fundamental frequency again, it can be interpreted that $F_0(t)$ tends to vary smoothly and slowly. In order to use regularity (ii) for $F_0(t)$, we regard it as the following physical constraint.

Physical constraint 4 *Temporal variation of the fundamental frequency in a small segment is constant:*

$$\frac{dF_0(t)}{dt} = 0. \tag{40}$$

□

In each small segment, it can be interpreted that the small segment has a constant duration for which the temporal variation of $F_0(t)$ has the same variance of $F_0(t)$. The small segment can be determined as follows:

$$\frac{1}{t_h - t_{h-1}} \int_{t_{h-1}}^{t_h} \left| F_0(t) - \overline{F_0(t)} \right|^2 dt \leq \Delta F_0^2, \tag{41}$$

where the length of the small segment is $t_h - t_{h-1}$ and ΔF_0^2 is the variance of $F_0(t)$.

The relationship between $F_0(t)$ and the small segments using Physical constraint 4 is shown in Fig. 7. For $F_0(t)$, as shown by dotted line in Fig. 7, segregated duration ($F_0(t)$ duration) is applied to small segments from Eq. (41).

The next section presents the grouping constraints for the fundamental frequency $F_0(t)$.

4.2 Grouping constraints

As the 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 regard it as the following physical constraint.

Physical constraint 5 *Suppose that $f_1(t)$ is a complex tone, $F_0(t)$ is the estimated fundamental frequency by Eq. (41), 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*

$$\ell = \frac{K}{2} - \left\lceil \frac{\log(n \cdot F_0(t)/f_0)}{\log \alpha} \right\rceil, \quad n = 1, 2, \cdots, N_{F_0}, \tag{42}$$

where α is the scale parameter. □

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

Physical constraint 6 *Let $f_1(t)$ be a complex tone. Suppose that $T_{\rm S} = t_{h-1}$ and $T_{\rm E} = t_h$ are the onset and offset of the fundamental component determined using Physical constraint 4, which is generated by one acoustic source. If an acoustic event obtained by a channel is a harmonic component of $f_1(t)$, then onset $T_{k,\rm on}$ and offset $T_{k,\rm off}$ determined for the same channel must satisfy*

$$|T_{\rm S} - T_{k,\rm on}| \leq 50 \text{ ms} \tag{43}$$

and

$$|T_{\rm E} - T_{k,\rm off}| \leq 100 \text{ ms}. \tag{44}$$

□

In this paper, onset $T_{k,\rm on}$ and offset $T_{k,\rm off}$ of the harmonic component 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 $|\frac{dS_k(t)}{dt}|$.

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

$$\hat{\hat{A}}_k(t) = \frac{1}{N_{F_0}} \sum_{\ell \in \mathbf{L}} \frac{\hat{A}_\ell(t)}{\|\hat{A}_\ell(t)\|}, \tag{45}$$

where $\mathbf{L}$ is the set of ℓ satisfying Eq. (42).

Here, in the two grouping constraints, physical constraint 5 works by segregating harmonic components of $f_1(t)$ and physical constraint 6 works by segregating non-harmonic components of $f_1(t)$. The algorithm for solving the problem of segregating two acoustic sources using physical constraints related to the four regularities is shown in Fig. 8.

5 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 it. 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.

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

One was the power ratio in terms of the amplitude envelope $A_k(t)$, i.e., likely SNR. 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

$$\text{Precision}(k) := 10 \log_{10} \frac{\int_0^T A_k^2(t) dt}{\int_0^T (A_k(t) - \hat{A}_k(t))^2 dt}, \tag{46}$$

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 measure was the spectrum distortion (SD). 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

$$\text{SD} := \sqrt{\frac{1}{W} \sum_{\omega}^{W} \left(20 \log_{10} \frac{\tilde{F}_1(\omega)}{\tilde{\hat{F}}_1(\omega)} \right)^2}, \tag{47}$$

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. In the above equation, the frame length is 51.2 ms, the frame shift is 25.6 ms, W is the analyzable bandwidth of filterbank (about 6 kHz), and the window function is Hamming.

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

5.1 Simulation 1

This simulation assumed that $f_1(t)$ was an AM complex tone as shown in Fig. 9, where $F_0 = 200$ Hz, $N_{F_0} = 10$, and whose amplitude envelope was sinusoidal (10 Hz), and $f_2(t)$ was a bandpassed pink noise, where the bandwidth was about 6 kHz. Five types of $f(t)$ were used as simulation stimuli, where the SNRs of $f(t)$ were from 0 to 20 dB in 5-dB steps.

For example, when the SNR of $f(t)$ was 10 dB, as shown in Fig. 10, the proposed method could segregate $A_k(t)$ with high precision and could extract $\hat{f}_1(t)$, shown in Fig. 11, from the $f(t)$. In this case, the precision for $A_k(t)$ is shown in Fig. 12. In addition,

for five simulations, the average SDs of $\hat{f}_1(t)$ and $f(t)$ are shown in Fig. 13. It was possible to reduce the SD by about 15 dB as noise reduction, using the proposed method. 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 signal and noise existed in the same frequency region.

5.2 Simulation 2

This simulation assumed that $f_1(t)$ was an AM complex tone the same as Fig. 9 and that $f_2(t)$ was another AM complex tone, where $F_0 = 300$ Hz, $N_{F_0} = 10$, and whose amplitude envelope 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 $f(t)$ were from 0 to 20 dB in 5-dB steps.

For example, when the SNR of $f(t)$ was 10 dB, as shown in Fig. 14, the proposed method could segregate $A_k(t)$ with high precision and could extract $\hat{f}_1(t)$, shown in Fig. 15, from the $f(t)$, even when two components of the signals existed in the same frequency region (e.g. the number of channel was 65, i.e. 600 Hz. In this case, the precision for $A_k(t)$ is shown in Fig. 16. In addition, for five simulations, the average SDs of $\hat{f}_1(t)$ and $f(t)$ are shown in Fig. 17. It was possible to reduce the SD by about 20 dB as noise reduction, using the proposed method. 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.

5.3 Simulation 3

This simulation assumed that $f_1(t)$ was a vowel /a/ synthesized by the LAM as shown in Fig. 18, where averaged $\overline{F_0(t)} = 125$ Hz, dynamic range was 5 Hz (from 123 to 128 Hz), and $f_2(t)$ was a bandpassed pink noise, where the bandwidth was about 6 kHz. Five types of $f(t)$ were used as simulation stimuli, where the SNRs of $f(t)$ were from 0 to 20 dB in 5-dB steps.

For example, when the SNR of $f(t)$ was 10 dB as shown in Fig. 19, the proposed method could segregate $A_k(t)$ with high precision and could extract $\hat{f}_1(t)$, shown in Fig. 20, from $f(t)$. In this case, the precision for $A_k(t)$ is shown in Fig. 21. In addition, for five simulations, the average SDs of $\hat{f}_1(t)$ and $f(t)$ are shown in Fig. 22. It was possible to reduce the SD by about 15 dB as noise reduction, using the proposed method. Hence, the proposed model could also extract with high precision 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. 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. 23. Hence, this method can be applied in cases where a speech signal is to be extracted from noisy speech.

Finally, the noise-reduction characteristics of the three simulations of the proposed method are shown in Fig. 24.

6 Conclusion

In this paper, we proposed a method of extracting the desired signal from a noisy signal, using physical constraints related to the four regularities proposed by Bregman, and by solving the problem of segregating two acoustic sources. 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 it. 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.

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 a noisy speech.

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.

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

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

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

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

[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.

[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.

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

[Unoki *et al.*, 1997b] Masashi Unoki and Masato Akagi. "A Method of Signal Extraction from Noisy Signal," In Proc. EuroSpeech'97, vol. 5, pp. 2583-2586, RHODOS-GREECE, Sept. 1997.

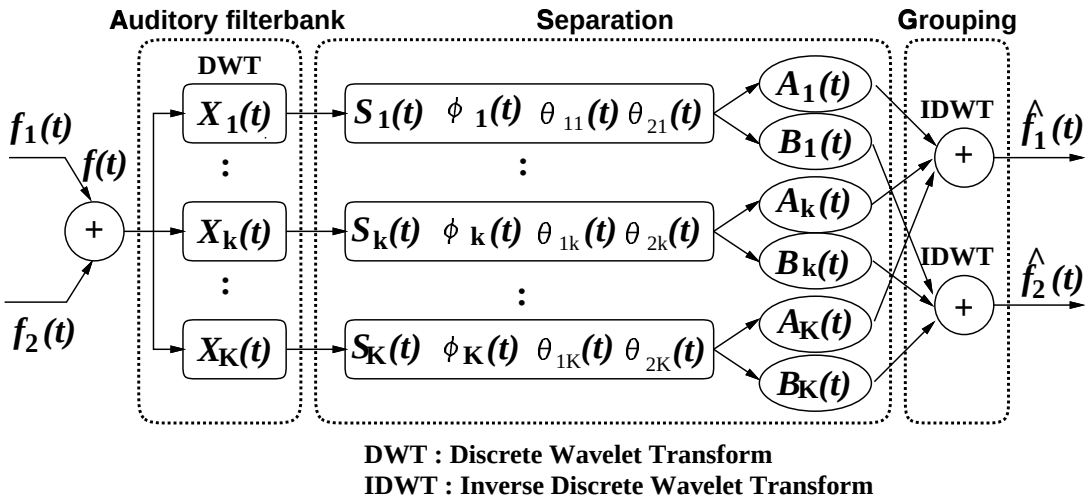

Figure 1: Auditory segregation model.

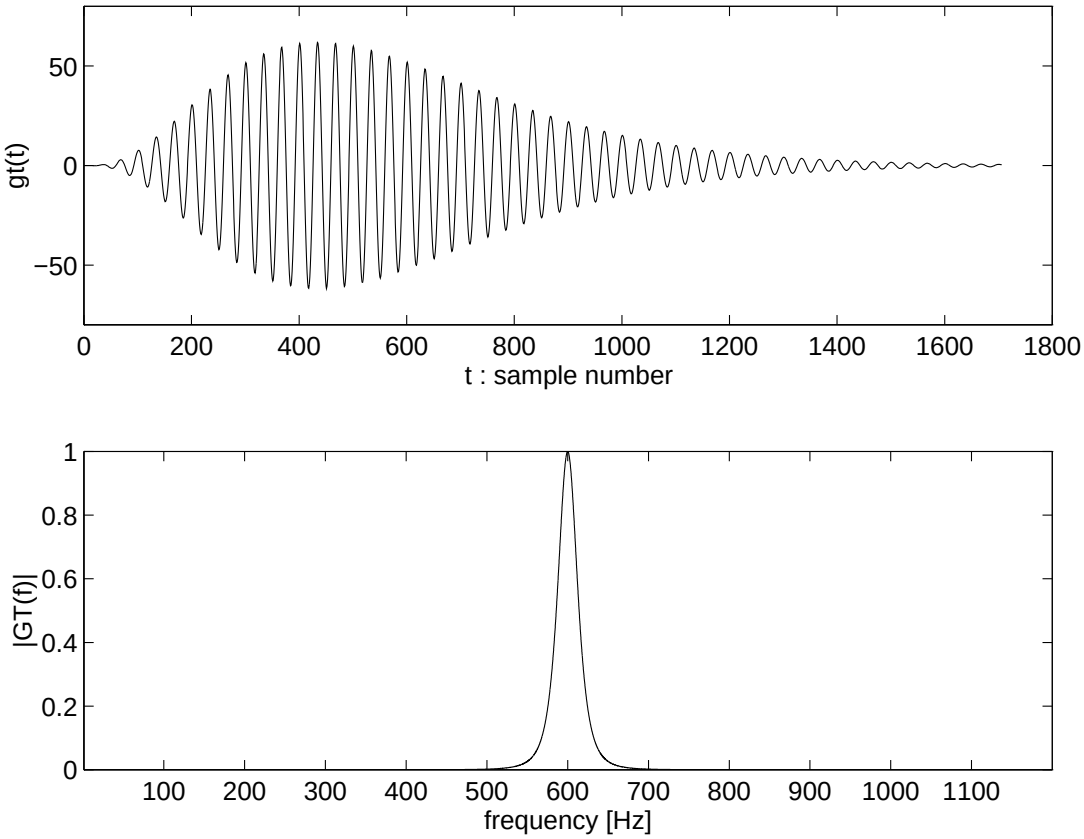

Figure 2: Impulse response and amplitude characteristics of the gammatone filter ($f_0 = 600$ Hz, $N = 4$, $b_f = 22.99$).

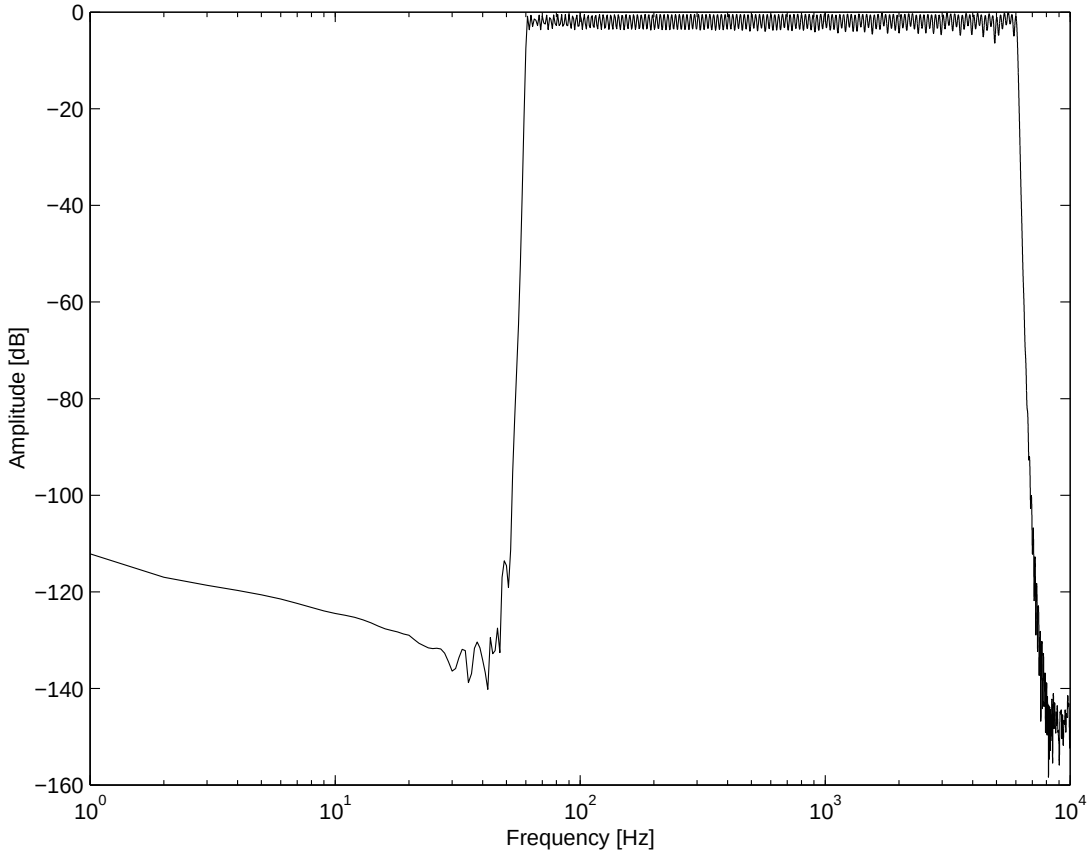

Figure 3: Frequency characteristics of the wavelet filterbank.

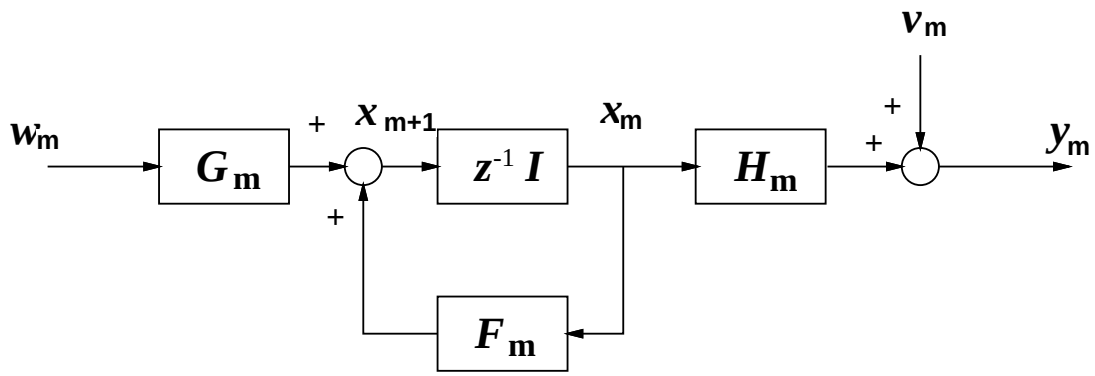

Figure 4: Basic system of the Kalman filter.

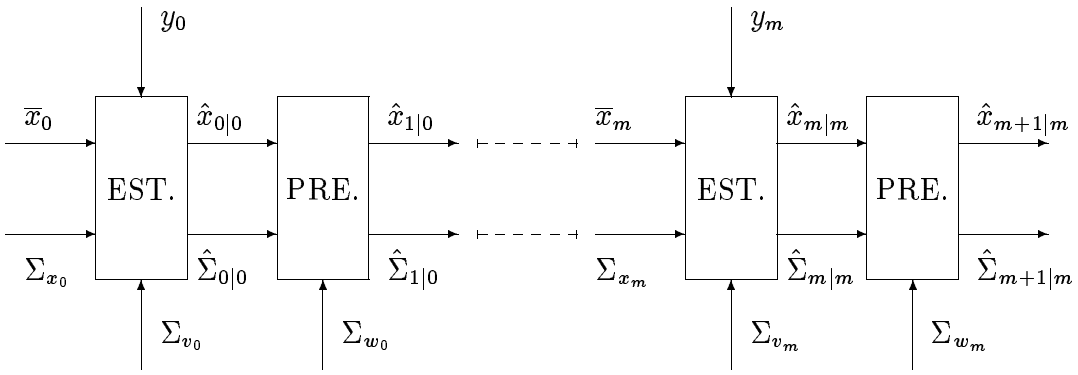

Figure 5: Algorithm for the Kalman filter. "EST." and "PRE." denote "estimation" and "prediction", respectively.

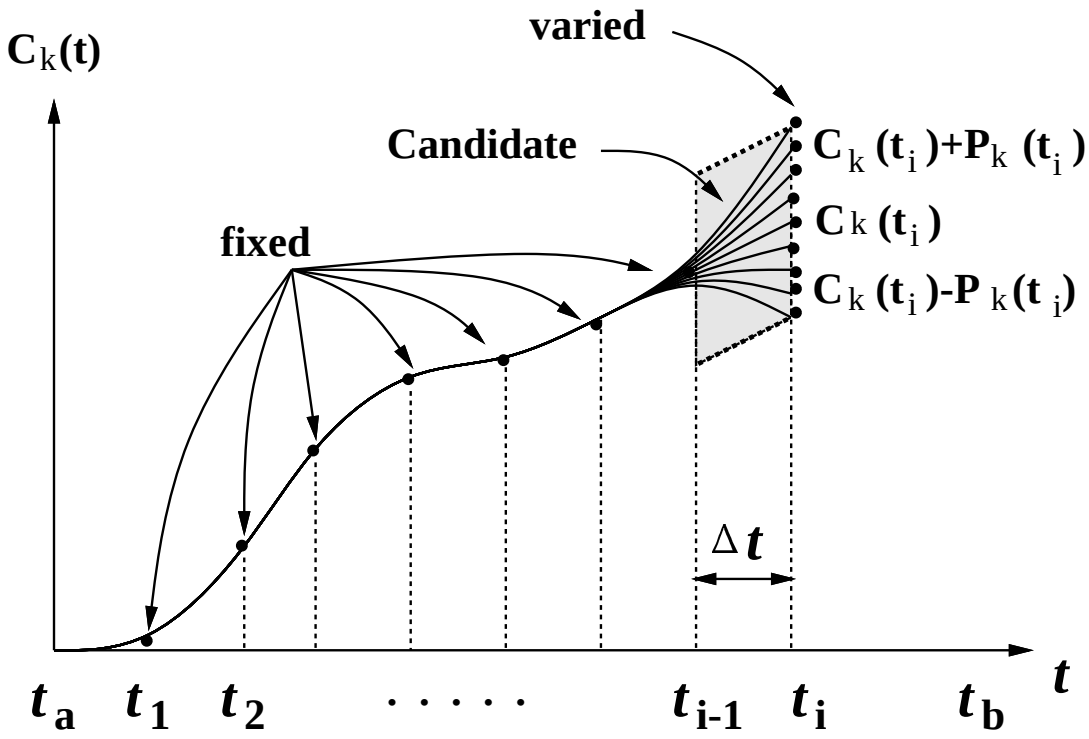

Figure 6: Candidates for $C_k(t)$ interpolated by the spline function.

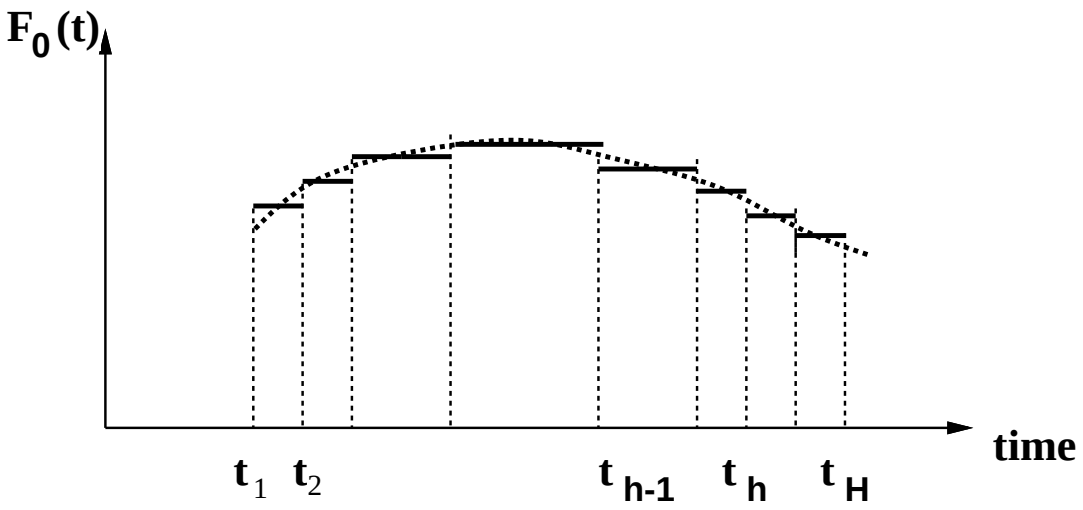

Figure 7: Temporal variation of the fundamental frequency.

decompose $f(t)$ into its frequency components using the wavelet filterbank (wavelet transform) as Eq. (11);
determine the fundamental frequency $F_0(t)$ using TEMPO;
let H be the number of dulations from Eq. (40);
for $k := 1$ **to** K **do**
  $\theta_{1k}(t) = 0$;
  determine $S_k(t)$ and $\phi_k(t)$ from Lemma 1;
  for $h := 2$ **to** H **do**
    $T_{\rm S} = t_{h-1}$ and $T_{\rm E} = t_h$;
    the segregated duration is $t_{h-1} \leq t \leq t_h$;
    determine onset $T_{k,\rm on}$ and offset $T_{k,\rm off}$;
    if Physical constraint 5 or 6 is satisfied **then**
      estimate $C_k(t)$ using the Kalman filter;
      determine the interpolated duration;
      let I be the number of interpolated samples;
      for $i = 1$ **to** I **do**
        determine the candidates for $C_k(t)$, which are interpolated by the spline function within $\hat{C}_k(t_i) - P_k(t_i) \leq C_k(t_i) \leq \hat{C}_k(t_i) + P_k(t_i)$;
        determine $\hat{\theta}_{2k}(t)$ from Eq. (19);
        determine $\hat{A}_k(t)$ from Eq. (14);
        determine $\hat{\hat{A}}_k(t)$ from Eq. (45);
        determine $\mathrm{Corr}(\hat{A}_k(t), \hat{\hat{A}}_k(t))$ from Eq. (39);
      end
      determine $C_k(t)$ when $\mathrm{Corr}(\hat{A}_k(t), \hat{\hat{A}}_k(t))$ becomes a maximum within the estimated -error;
      determine $\theta_{2k}(t)$ from Eq. (19);
    else
      set $A_k(t) = 0$, $B_k(t) = S_k(t)$ and $\theta_{2k}(t) = \phi_k(t)$;
    end
    determine $A_k(t)$ and $B_k(t)$ from Eqs. (14) and (15);
  end
  determine each frequency component of $f_1(t)$ and $f_2(t)$ from Eqs. (9) and (10);
end
reconstruct $\hat{f}_1(t)$ and $\hat{f}_2(t)$ using the wavelet filterbank (inverse wavelet transform) from Eqs. (14) and (15);

Figure 8: Segregation algorithm.

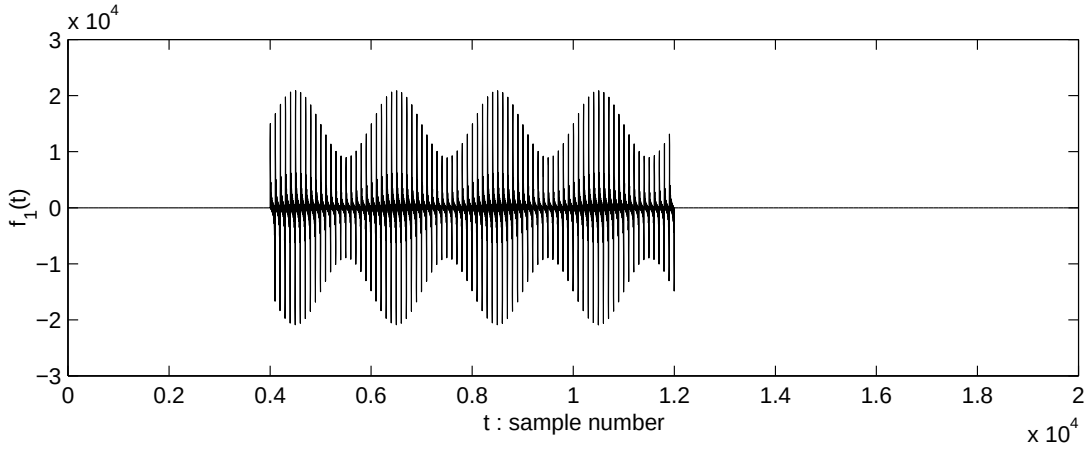

Figure 9: AM complex tone $f_1(t)$.

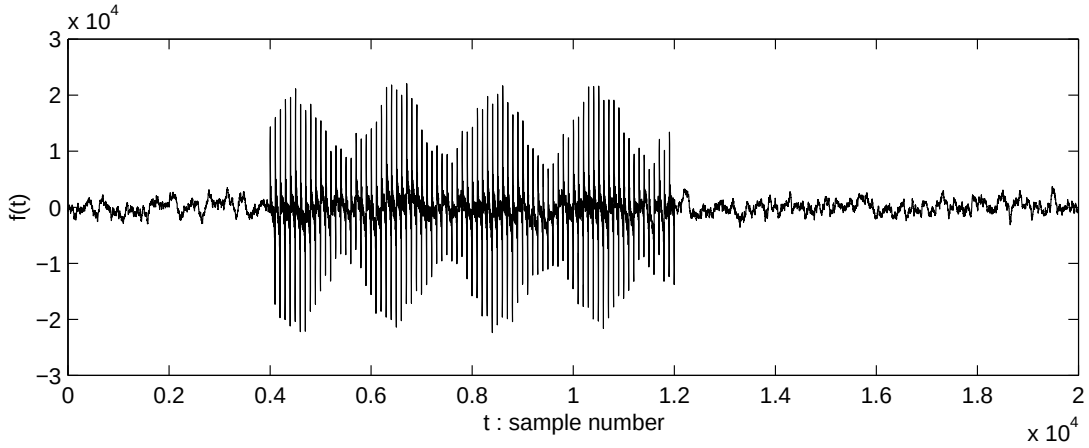

Figure 10: Mixed signals $f(t)$ (SNR= 10 dB).

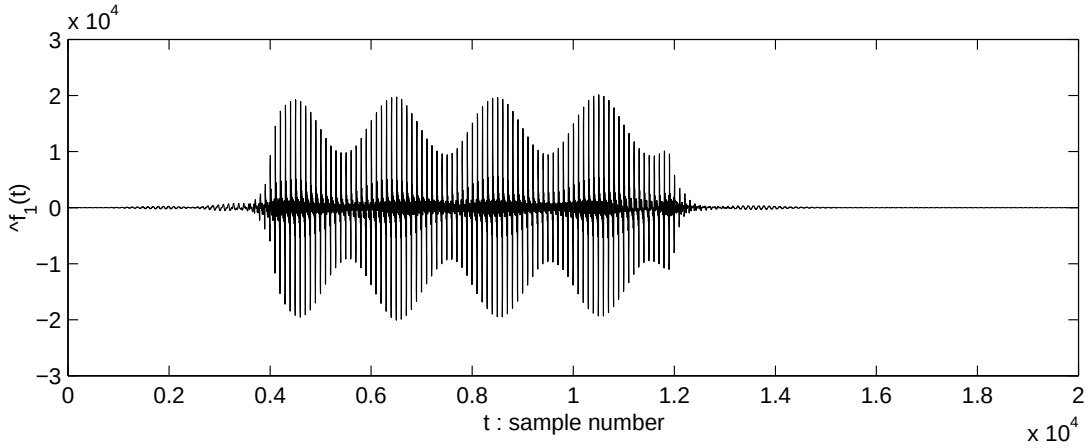

Figure 11: Extracted signal $\hat{f}_1(t)$ (SNR= 10 dB).

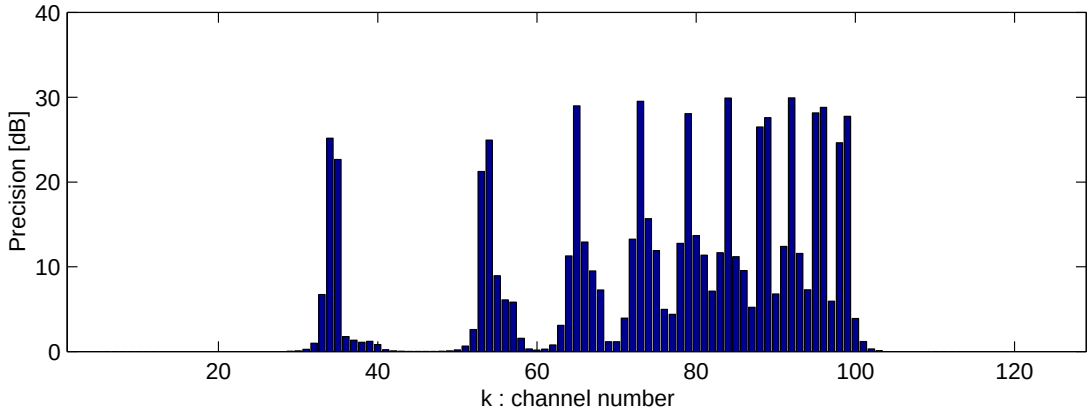

Figure 12: Precision for $A_k(t)$ (SNR= 10 dB).

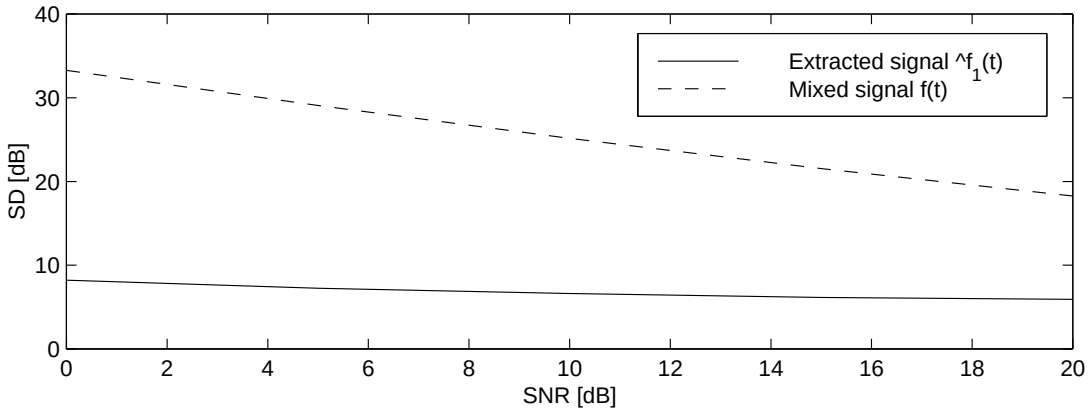

Figure 13: SD for the extracted signal $\hat{f}_1(t)$.

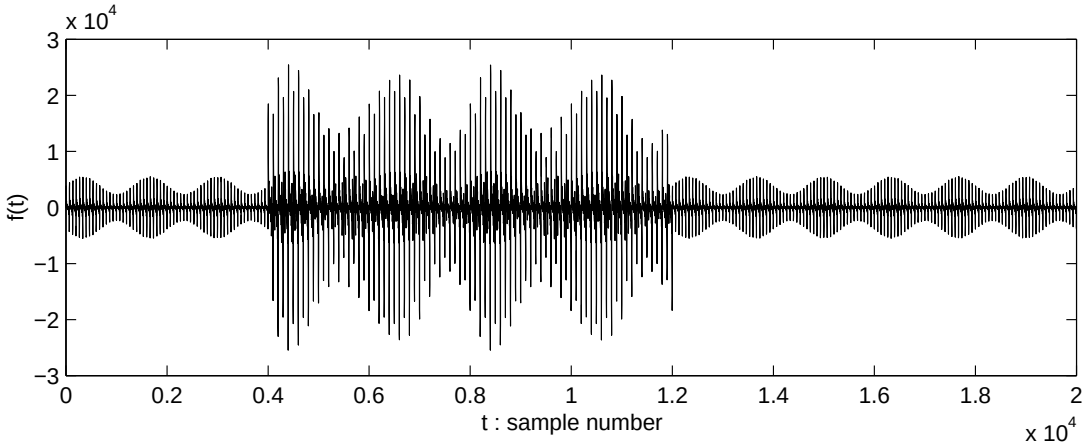

Figure 14: Mixed signals $f(t)$ (SNR= 10 dB).

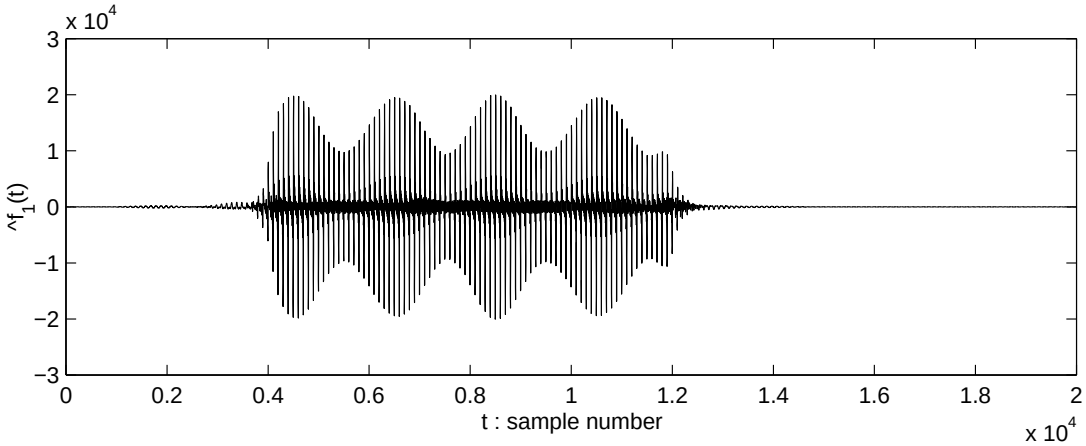

Figure 15: Extracted signal $\hat{f}_1(t)$ (SNR= 10 dB).

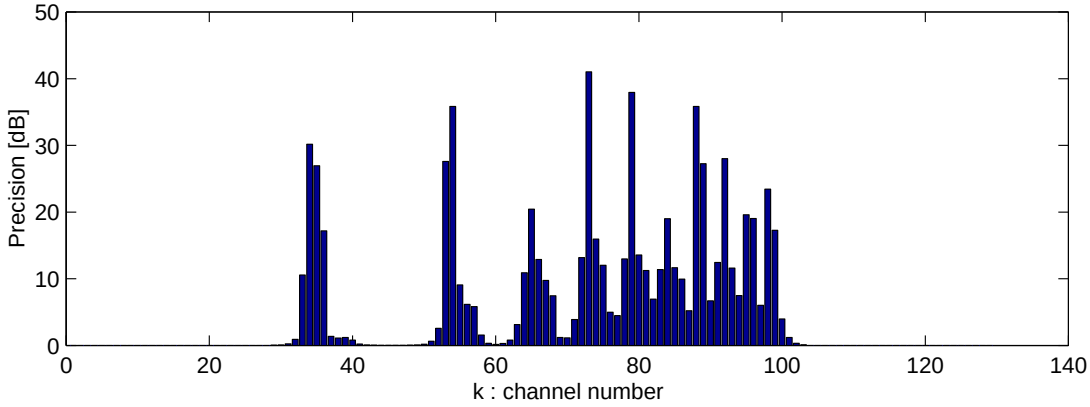

Figure 16: Precision for $A_k(t)$ (SNR= 10 dB).

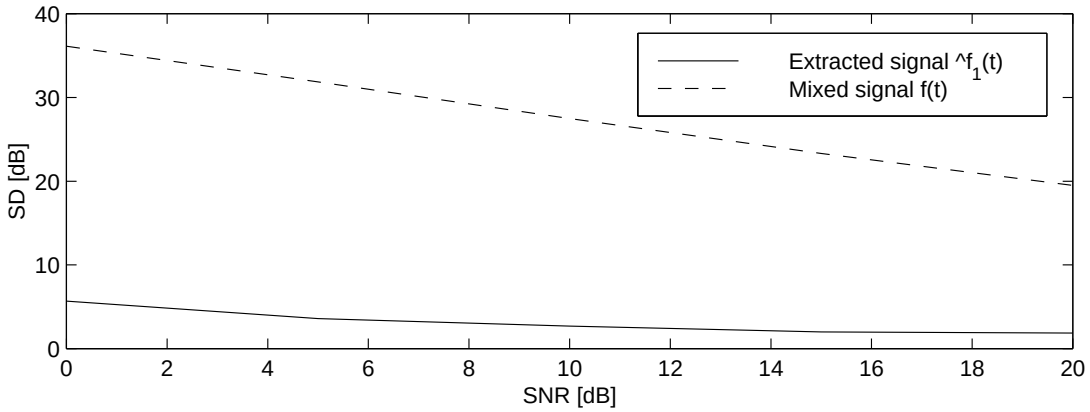

Figure 17: SD for the extracted signal $\hat{f}_1(t)$.

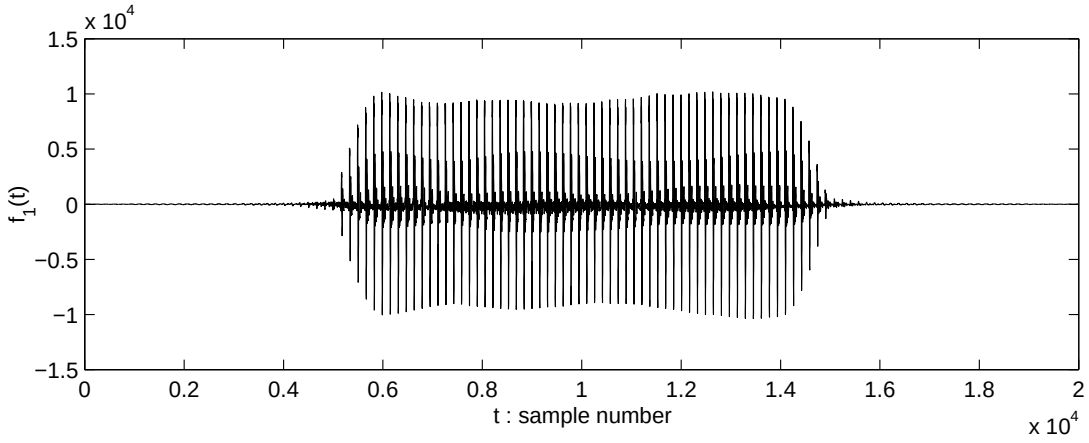

Figure 18: LMA-synthesized vowel /a/ $f_1(t)$.

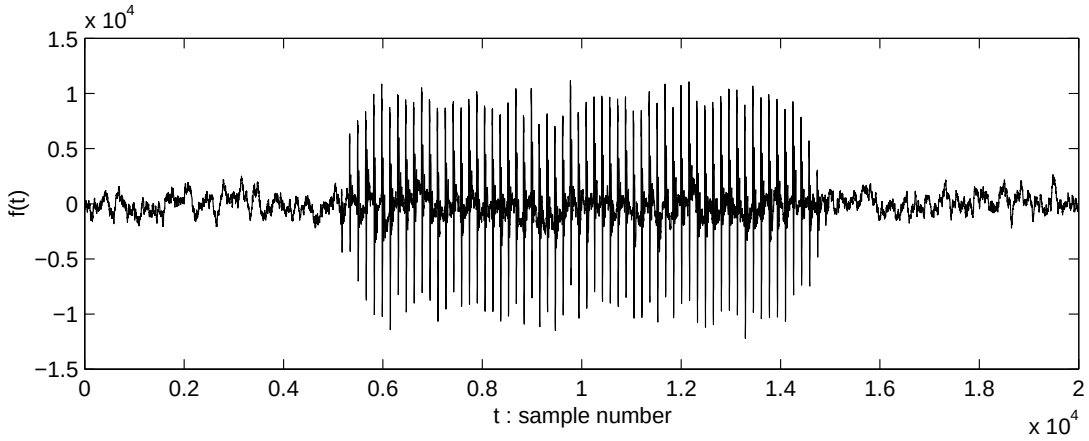

Figure 19: Mixed speech $f(t)$ (SNR= 10 dB).

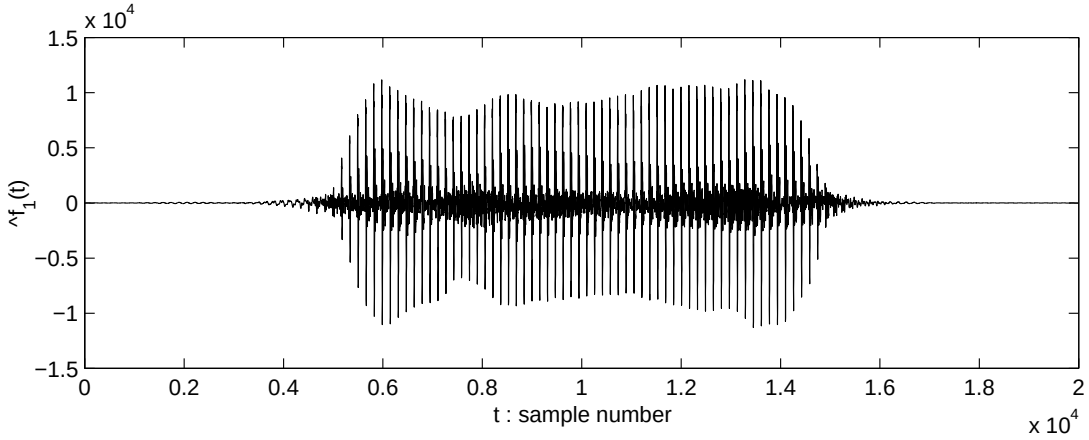

Figure 20: Extracted vowel /a/ $\hat{f}_1(t)$ (SNR= 10 dB).

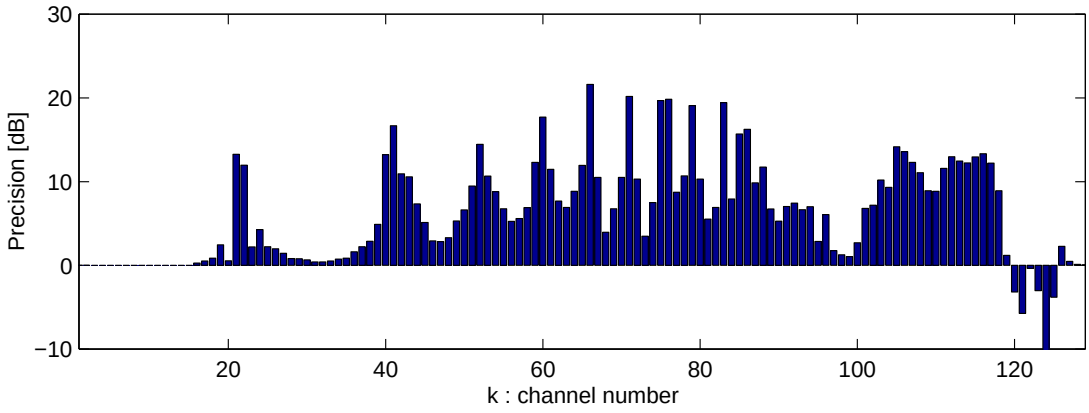

Figure 21: Precision for $A_k(t)$ (SNR= 10 dB).

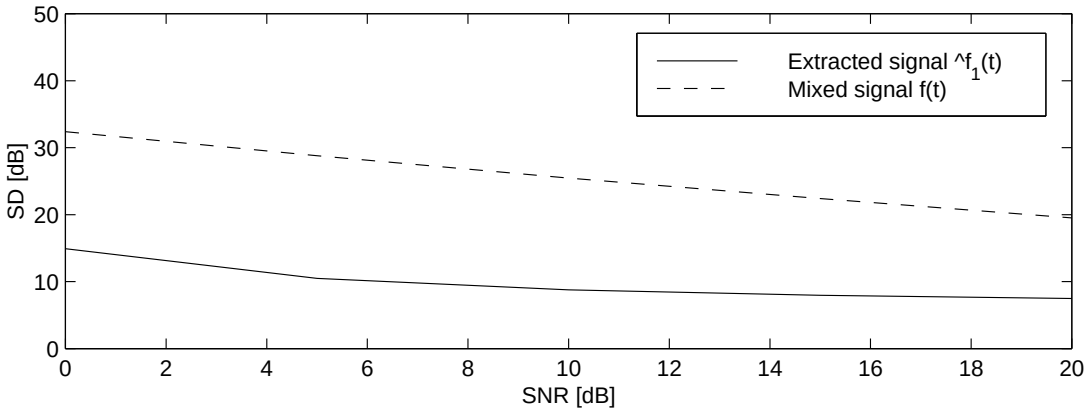

Figure 22: SD for the extracted signal $\hat{f}_1(t)$.

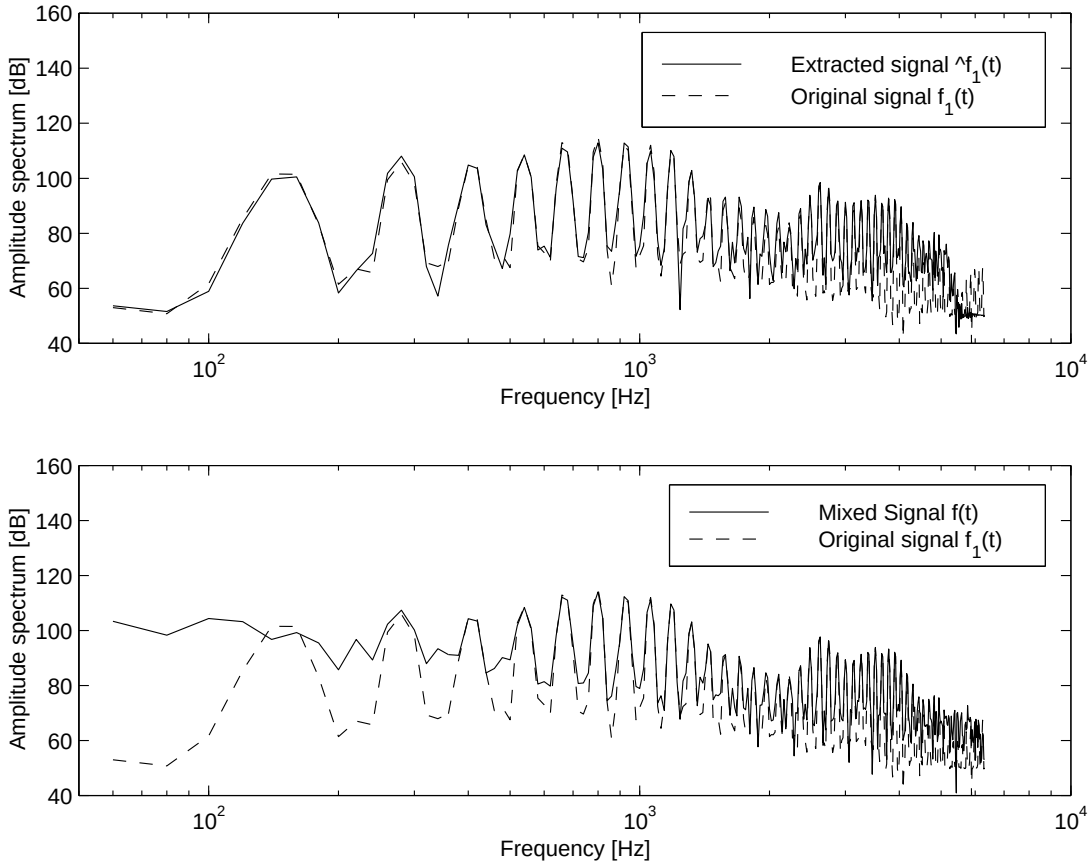

Figure 23: Comparison of the amplitude spectra of $\hat{f}_1(t)$ and $f_1(t)$.

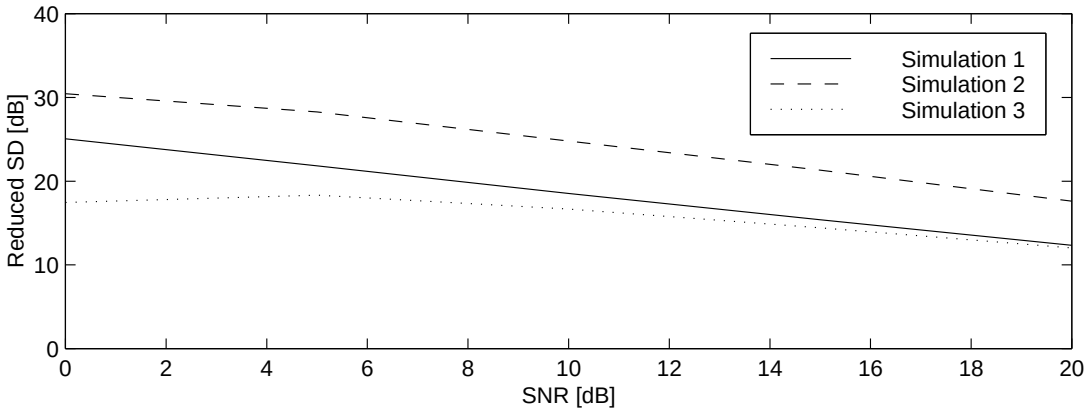

Figure 24: Characteristics of the reduced SD for the three simulations.