Computational Auditory Scene Analysis Based on Loudness/Pitch/Timbre Decomposition

Mototsugu Abe and **Shigeru Ando**
Department of Mathematical Engineering and Information Physics
Faculty of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113, JAPAN
E-Mail: abe@alab.t.u-tokyo.ac.jp

Abstract

For two significant problems: 1) how to catch and trace individual stream attributes, and 2) how to extract an actual sound for a specified stream, of computational auditory scene analysis, we have developed an algorithm for the former one based on loudness/pitch/timbre decomposition of time-varying sound energy[Abe *et al.* , 1996]. By using this, we propose in this paper an algorithm for computational auditory scene analysis, in which we construct a time-varying multimodal distribution of loudness change rate (LCR) and pitch shift rate (PSR), trace modes of it as stream attributes, and extract a sound waveform corresponding to a selected trace of LCR and PSR. We investigate the performance of this algorithm using several musical sounds and voices.

1 Introduction

In a complex auditory scene, it is known[Bregman, 1990] that man recognizes a set of tones as a stream when all their cues such as pitches and amplitudes change synchronously, but man perceives an arrival of different stream if some of them change asynchronously. This suggests that human auditory system is very sensitive to coherency and incoherency involved in complex sounds.

We have proposed a method of loudness/pitch/timbre decomposition[Abe *et al.* , 1995] by which the changing energy of a sound is decomposed into two coherent components (loudness change component and pitch shift component) and incoherent component (timbre change component). Based on this method, we have developed a basic principle for computational auditory scene analysis which utilizes coherency in pitch and amplitude changes as a segregation cue[Abe *et al.* , 1996].

In this paper, we propose an algorithm including both an estimation/tracing method of stream attributes and an extraction method of a sound corresponding to a selected trace of the attributes. We first review the loudness/pitch/timbre decomposition. We next extend it for application to the segregation problem for multiple streams. In it, we construct a multimodal distribution of loudness change rate (LCR) and pitch shift rate (PSR) using Parzen's method[Fukunaga, 1972], integrate the sequence of it using non-parametric Kalman filter[Ando, 1994], and trace its dominant modes as stream attributes. We then construct a reconstruction method (gradient space filtering) which extracts the wavelet components corresponding to the trace of LCR and PSR. Inverse wavelet transform reconstructs an actual sound waveform (Fig.1). We examine the performance of this algorithm using several musical sounds and voices.

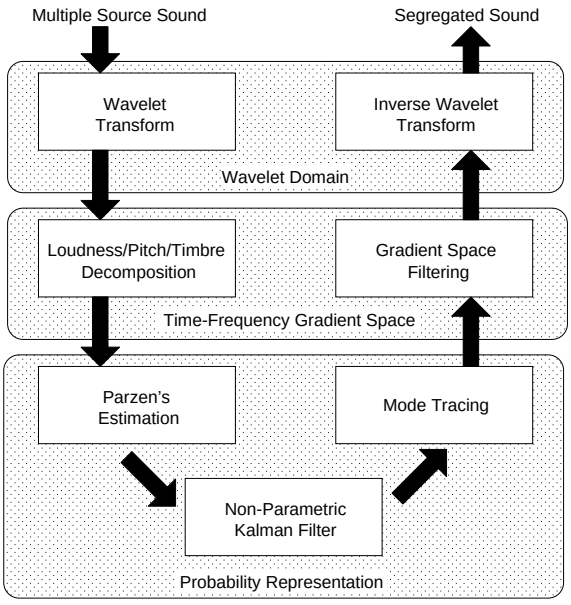

Figure 1: Schematic diagram of our system.

2 Loudness/Pitch/Timbre Decomposition

Let $f(t)$ be a sound comes from a single source. Let

$$F(t,\omega) = |\int_{-\infty}^{\infty} e^{\omega}\psi^*(e^{\omega}(\tau - t))f(\tau)d\tau|^2 \quad (1)$$

$$\psi(t) \equiv A\exp(-\frac{\Delta^2 t^2}{2} + j\Omega_c t) \quad (2)$$

be a wavelet energy distribution(WED)[Daubechies, 1990] and let

$$\Phi(t,\omega) = \arg[\int_{-\infty}^{\infty} e^{\omega}\psi^*(e^{\omega}(\tau - t))f(\tau)d\tau] \quad (3)$$

be a wavelet phase of $f(t)$, where $\psi(t)$ denotes Gabor function, Ω_c denotes the center frequency of it and ω denotes log-frequency. We eliminate undesirable cross-terms in the WED by applying an LPF whose cut off frequency is lower than an expected pitch of the incoming sound. Let

$$F_t(t,\omega) = \frac{\partial}{\partial t}F(t,\omega),\ F_\omega(t,\omega) = \frac{\partial}{\partial \omega}F(t,\omega) \quad (4)$$

be time and log-frequency gradient of $F(t,\omega)$. Then, we can make a scatter diagram of (F, F_t, F_ω) at $t = t_k$ in a small region

$$\Gamma \ = \ [t_k - \Delta t < t < t_k + \Delta t] \times [-\infty < \omega < \infty] \quad (5)$$

to describe instantaneous changes of $F(t,\omega)$ where Δt denotes very short interval (actually, a few sampling interval). We call the space spanned by F, F_t, F_ω the time-frequency gradient space (TFGS).

We have shown[Abe *et al.* , 1996] that an instantaneous change in wavelet domain can be decomposed into three orthogonal components (Fig.2) as

$$F_t(t,\omega) \ = \ \alpha F(t,\omega) + \beta F_\omega(t,\omega) + \xi(t,\omega) \quad \text{for } \forall(t,\omega) \in \Gamma \quad (6)$$

where αF corresponds to a loudness change component, α denotes loudness change rate (LCR), βF_ω corresponds to a pitch shift component and β is pitch shift rate (PSR). We refer to the residual change ξ as a timbre change component (Fig.3)[Abe *et al.* , 1995]. Eq.(6) shows that if a timbre change component is sufficiently small, i.e. the change is almost coherent, the TFGS distribution reduces to a plane. Least square estimates of α and β are estimated as

$$\tilde{\alpha} = \frac{\det[\mathbf{S}_{12}]}{\det[\mathbf{S}_{22}]}, \quad \tilde{\beta} = \frac{\det[\mathbf{S}_{32}]}{\det[\mathbf{S}_{22}]}, \quad (7)$$

where $\mathbf{S}$ is a correlation matrix of $[F, F_t, F_\omega]$ in Γ as

$$\mathbf{S} \ = \ \begin{bmatrix} S_{FF} & S_{Ft} & S_{\omega F} \\ S_{Ft} & S_{tt} & S_{t\omega} \\ S_{\omega F} & S_{t\omega} & S_{\omega\omega} \end{bmatrix} = \int_\Gamma \begin{bmatrix} F \\ F_t \\ F_\omega \end{bmatrix} [\ F \ \ F_t \ \ F_\omega \ ]\, dtd\omega, \quad (8)$$

and $\mathbf{S}_{ij}$ $(i,j = 1,2,3)$ denote 2×2 submatrices of $\mathbf{S}$ in which the ith row and jth column of $\mathbf{S}$ are removed.

Error variances of α and β can be estimated as

$$\sigma_\alpha^2 = \frac{S_{FF}J(\tilde{\alpha},\tilde{\beta})}{L^2\det[\mathbf{S}_{22}]}, \quad \sigma_\beta^2 = \frac{S_{\omega\omega}J(\tilde{\alpha},\tilde{\beta})}{L^2\det[\mathbf{S}_{22}]}, \quad (9)$$

where $L = \int_\Gamma dtd\omega$ and the residual

$$J(\tilde{\alpha},\tilde{\beta}) = \frac{\det[\mathbf{S}]}{\det[\mathbf{S}_{22}]} \quad (10)$$

is an actual energy of timbre change.

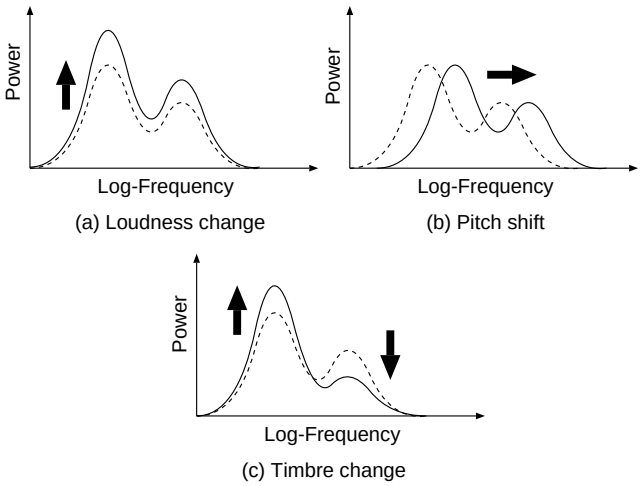

Figure 2: Three orthogonal components of an instantaneous change of wavelet energy. Dashed lines show a WED $F(t,\omega)$ and solid lines show a WED $F(t+dt,\omega)$, where dt denotes very short interval. Loudness change shown in (a) can be modeled as $F(t+dt,\omega) = (1+\alpha dt)F(t,\omega)$ and pitch shift shown in (b) as $F(t,\omega) = F(t,\omega+\beta dt)$, where α and β denote LCR and PSR, respectively. An example of timbre change is shown in (c), which is defined as the residual change of loudness change and pitch shift.

3 Multimodal Stream Tracing Algorithm

3.1 Multiple (α, β) Estimation

For individual sound from single source and their mixture from multiple sources, we assume that:

Assumption 1: A change of sound from a single source is smooth and coherent almost everywhere.

Assumption 2: Overlaps of wavelet energy distributions from different sources are small.

The former assumption assures that TFGS distribution of single source sound is a plane almost always,

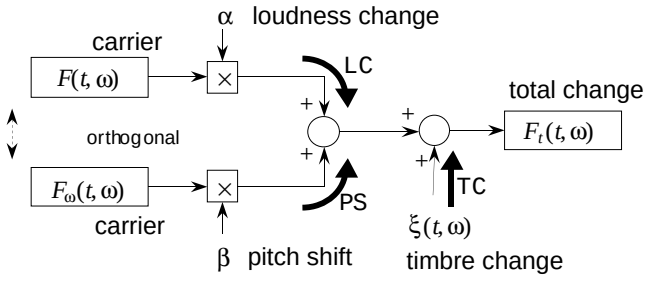

Figure 3: Block diagram of loudness/pitch/timbre decomposition. An arbitrary change $F_t(t,\omega)$ is composed of orthogonal three components as: 1) loudness change which is encoded on a carrier $F(t,\omega)$, 2) pitch shift which is encoded on a carrier $F_\omega(t,\omega)$, and 3) timbre change $\xi(t,\omega)$ which is the residual component of these two change.

and the latter assures that TFGS distribution of multiple source sound is an inclusion of individual planes[Abe *et al.* , 1996]. This is because the second assumption assures that any interference terms between WED of different streams are zero, i.e.,

$$F(t,\omega) = F_1(t,\omega) + F_2(t,\omega) \tag{11}$$

where from the first assumption

$$F_{1t} = \alpha_1 F_1 + \beta_1 F_{1\omega} \tag{12}$$
$$F_{2t} = \alpha_2 F_2 + \beta_2 F_{2\omega}. \tag{13}$$

Again from the second assumption, either F_1 or F_2 equal to 0. This assures that

$$F_t = \begin{cases} \alpha_1 F + \beta_1 F_\omega & \text{in } \Gamma_1 \\ \alpha_2 F + \beta_2 F_\omega & \text{in } \Gamma_2 \end{cases}, \tag{14}$$

i.e. (F, F_t, F_ω) is always on one of two planes. Γ_1 and Γ_2 are regions of (t,ω) where F_1 and F_2, respectively, have significant values.

Based on these properties, we can estimate (α,β) of individual streams as follows:

1. Divide Γ into small regions as
$$\Gamma_m = [t_k - \Delta t < t < t_k + \Delta t] \times [\omega_m - \Delta\omega < \omega < \omega_m + \Delta\omega], \tag{15}$$
where $\Delta\omega$ denotes very short interval in log-frequency axis and $m(=1,...,M)$ denotes an index of a region.
2. Compute LCR and PSR with their error variance $\{\tilde{\alpha}_m, \tilde{\beta}_m, \sigma_{\alpha_m}, \sigma_{\beta_m}\}$ in each region Γ_m using eqs.(7) and (9).
3. Construct a non-parametric probability density function(pdf) for (α,β) using Parzen's method[Fukunaga, 1972] as
$$Q(\alpha,\beta) = \sum_{m=1}^{M} \frac{1}{2\pi\sigma_{\alpha_m}\sigma_{\beta_m} M} \exp\left(-\frac{(\alpha-\tilde{\alpha}_m)^2}{2\sigma_{\alpha_m}^2} - \frac{(\beta-\tilde{\beta}_m)^2}{2\sigma_{\beta_m}^2}\right). \tag{16}$$

3.2 Optimum Tracing using Non-parametric Kalman filter

In order to integrate information which is successively obtained by the above algorithm, we use non-parametric Kalman filter(NPKF)[Ando, 1994] by which we can exclude instances which are incompatible with the assumptions and then we can construct a reliable pdf.

NPKF is a discrete time discrete observation successive filter which is performed by two steps as follows (Fig.4):

a) Diffusion step
Let $A_k = (\alpha_k, \beta_k)$ and y_k, respectively, be stream attributes and an observation at time t_k, and let $Y_k = \{y_k, y_{k-1}, \ldots\}$ be a sequence of observations before t_k. Then, the conditional pdf of A_{k-1} and A_k are related as

$$P(A_k|Y_{k-1}) = \int P(A_k|A_{k-1}) P(A_{k-1}|Y_{k-1}) dA_{k-1} \tag{17}$$

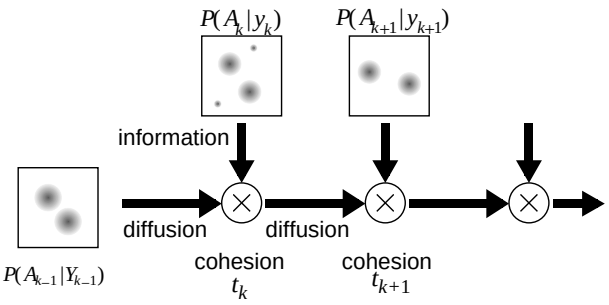

Figure 4: Schematic diagram of non-parametric Kalman filter. In intervals of no information, the conditional pdf $P(A_k|Y_k)$ is driven by a dynamics of (α,β), for which we use a simple Gaussian diffusion. When information is given, the pdf is updated as eq.(18).

where the transition probability $P(A_k|A_{k-1})$ is determined by a dynamics of (α,β). Because they change smoothly from the assumption 1, we can apply a simple Gaussian diffusion for this step.

b) Cohesion step
From the algorithm discussed in last section, we can express information (reduction of ambiguity) obtained at time t_k as

$$P(A_k|y_k) = Q(\alpha,\beta). \tag{18}$$

Using this, we can update the conditional pdf as

$$P(A_k|Y_k) = \frac{P(A_k|y_k) P(A_k|Y_{k-1})}{\int P(A_k|y_k) P(A_k|Y_{k-1}) dA_k}. \tag{19}$$

In general, the conditional pdf $P(A_k|Y_k)$ obtained through these iteration is multimodal. Therefore we can estimate multiple stream parameters by: 1) tracing each peak position $(\tilde{\alpha}_i, \tilde{\beta}_i)$ of the pdf sequence, and 2) determining $(\sigma_{\alpha_i}, \sigma_{\beta_i})$ from variance around the peak.

4 Stream Reconstruction Algorithm

4.1 Wavelet Component Classification

From the assumptions, wavelet components coming from a single source are restricted on a plane in TFGS. Therefore if we determine the plane

$$F_t - \tilde{\alpha}_i F - \tilde{\beta}_i F_\omega = 0 \tag{20}$$

using the above algorithm, extraction of the corresponding wavelet components is done by placing a window near the target plane as shown in Fig.5.

For evaluating a margin of the window, let $\epsilon_\alpha, \epsilon_\beta$ be actual estimation errors of α_i and β_i, i.e.

$$\alpha_i = \tilde{\alpha}_i + \epsilon_\alpha, \quad \beta_i = \tilde{\beta}_i + \epsilon_\beta. \tag{21}$$

Generally, the errors are expected to be smaller than their estimate variances $\{\sigma_{\alpha_i}, \sigma_{\beta_i}\}$. Let $(F_i, F_{it}, F_{i\omega})$ be a point on the true target plane. Then eq.(20) can be written as,

$$F_{it} - \tilde{\alpha}_i F_i - \tilde{\beta}_i F_{i\omega} = \epsilon_\alpha F_i + \epsilon_\beta F_{i\omega}. \tag{22}$$

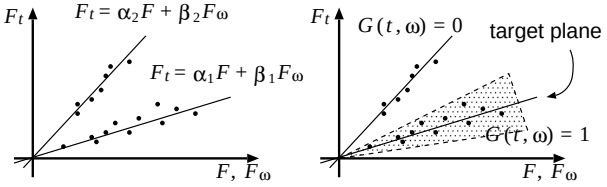

Figure 5: Classification of wavelet components of multiple streams. TFGS components (F, F_t, F_ω) in Γ (denoted by black points) are restricted on planes (solid lines) corresponding to each stream (left figure). Extraction of the corresponding components is done by placing a window $G(t, \omega)$ near to the plane (right figure).

Then distance between the true target plane to the estimated plane is

$$\begin{aligned} D &\equiv |F_{it} - \tilde{\alpha}_i F_i - \tilde{\beta}_i F_{i\omega}| \\ &= |\epsilon_\alpha F_i + \epsilon_\beta F_{i\omega}| \\ &\leq \sigma_{\alpha_i} |F_i| + \sigma_{\beta_i} |F_{i\omega}|. \end{aligned} \tag{23}$$

Therefore we can set the size of margin as

$$D_0 \equiv \sigma_{\alpha_i} |F| + \sigma_{\beta_i} |F_\omega|. \tag{24}$$

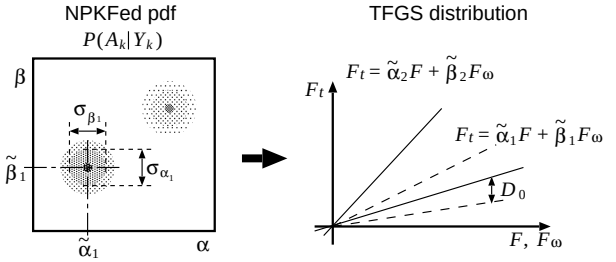

Figure 6: A margin of the extraction window for a specified single stream. Using $\{\tilde{\alpha}_i, \tilde{\beta}_i, \sigma_{\alpha_i}, \sigma_{\beta_i}\}$ estimated from NPKFed pdf (left figure), the target plane and the margin are determined as right figure.

4.2 Gradient Space Filtering

We extract wavelet components from a single source by a following algorithm. We call the algorithm the *gradient space filtering*.

1) Let

$$D(t, \omega) = |F_t(t, \omega) - \tilde{\alpha}_i F(t, \omega) - \tilde{\beta}_i F_\omega(t, \omega)| \tag{25}$$

be a distance between the target plane and a wavelet component (F, F_t, F_ω).

2) Let a compatibility function $G(t, \omega)$ be

$$G(t, \omega) = \frac{1}{(\frac{D(t,\omega)}{D_0})^2 + 1}, \tag{26}$$

where D_0 is calculated by eq.(24). $G(t, \omega)$ has a following property:

$$G(t, \omega) = \begin{cases} 1 & \text{if } D(t, \omega) = 0 \\ 1/2 & \text{if } D(t, \omega) = D_0 \\ 0 & \text{if } D(t, \omega) \to \infty. \end{cases} \tag{27}$$

3) Multiply the original WED $F(t, \omega)$ by the compatibility function $G(t, \omega)$ as

$$\tilde{F}_i(t, \omega) = G(t, \omega) F(t, \omega). \tag{28}$$

4) Reconstruct an actual sound for a target stream by the inverse wavelet transform as

$$\tilde{f}(t) = \int e^{\frac{1}{\omega}} \tilde{F}_i(\tau, \omega)^{\frac{1}{2}} e^{j\Phi(\tau,\omega)} \psi(e^\omega(\tau - t)) dt d\omega \tag{29}$$

where the phase $\Phi(t, \omega)$ is given by eq.(3).

5 Experiments

a) Synthesized Sound

In order to evaluate the above theories and algorithm, we first apply it to an ideally synthesized sound which almost completely satisfies the assumptions except some overlapping components. The waveform of it is shown in Fig.7(a) and the WED is shown in Fig.7(b), in which two streams are clearly observed. In order to satisfy the assumption 2, we set the bandwidth of the Gabor wavelet function to 1/24 octave.

One stream (stream A) is composed of 7 harmonic components whose frequencies are stable (base frequency is 312Hz), and the other (stream B) is composed of 4 harmonic components whose frequencies is slightly increasing at the rate of 2 octave/sec. Amplitude of them are sinusoidally modulated. SN ratio of stream A to B is 3.6dB, and therefore that of stream B to A is -3.6dB.

LCR and PSR estimated by eq.(7) in Γ are shown in Fig.7(c) and Fig.7(d), respectively. LCR and PSR of stream A are correctly estimated in one stream intervals, while the estimates are meaningless in a two stream interval, because they are estimated by treating all log-frequency component together as eq.(5).

TFGS distributions at 0.12 sec. and 0.40 sec. are shown in Fig.8(a) and Fig.8(b), respectively. As shown in eq.(14), it distributes on a plane in one stream interval(a), while it distributes on two planes in two stream interval(b).

A pdf sequence estimated by Parzen's method is shown in Fig.9. Two peaks corresponding to two streams are observed, while some irregular peaks are existing because of uneliminated cross-term components between the 1st harmonic component of stream A and the 2nd component of stream B.

A pdf sequence filtered by NPKF is shown in Fig.10. The irregular peaks observed in Fig.9 are removed. We set an initial distribution $P(A_0)$ as an uniform distribution because we had no information initially, and in order to facilitate the growth of multiple peaks, we used

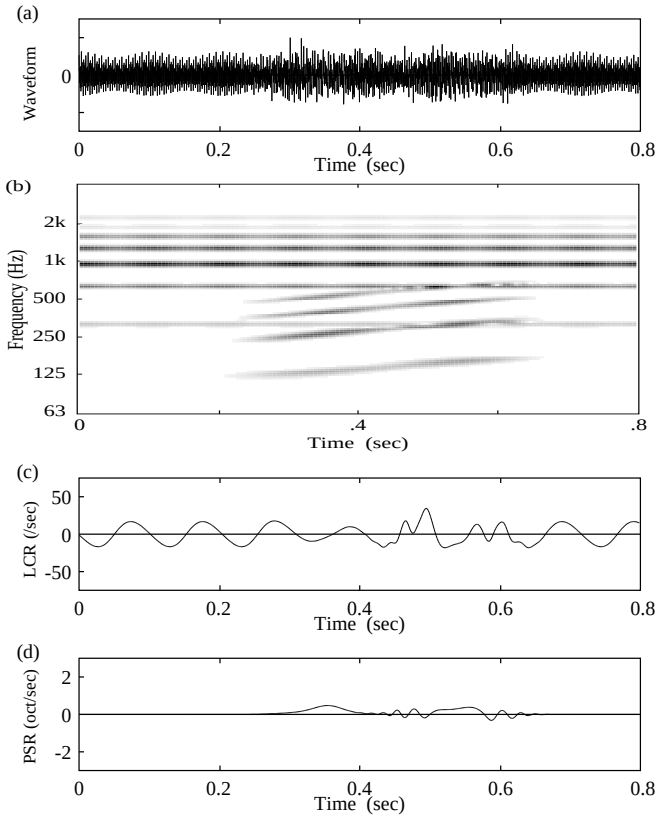

Figure 7: (a) Waveform, (b) WED, (c) LCR and (d) PSR of a synthesized sound. The sound is composed of two streams referred to stream A and B. Stream A: from 0 sec. to 0.8 sec., 7 harmonics, with amplitude change, with no pitch shift (312Hz stable pitch). Stream B: from 0.2 sec to 0.7 sec., 4 harmonics, with amplitude change, with pitch shift from 100Hz to 150Hz. LCR and PSR are properly estimated in the interval of one stream, whereas the estimation is meaningless in the interval of two stream. This is because LCR and PSR are estimated using all components along log-frequency.

large variances of Gaussian diffusion as 100sec^{-1} for α and 10oct./sec. for β.

Traces of LCR and PSR of the strong stream which are the position of the highest peak of NPKFed pdf sequence are shown in Fig.11(a) and Fig.11(b), respectively. Traces of LCR and PSR of the weak stream which are the position of the second highest peak are shown in Fig.11(c) and Fig.11(d), respectively. Comparing them to Fig.7(c) and (d), two streams are independently traced.

The reconstructed WED and waveform of the strong stream by the gradient space filtering are shown in Fig.12(a) and Fig.12(b), respectively. Except overlapping components, components of stream A is correctly extracted. SN ratio of stream A is increased to 10.8dB (i.e. 7.2dB improvement).

The reconstructed WED and waveform of the weak stream are shown in Fig.12(c) and Fig.12(d), respectively. Because amplitude of wavelet components at onset and offset are small and then estimation errors of LCR and PSR become large, some components of the other stream are remaining. However, except overlapping components and at onset and offset, stream B is almost correctly extracted. SN ratio of stream B is increased to 5.0dB (i.e. 8.6dB improvement).

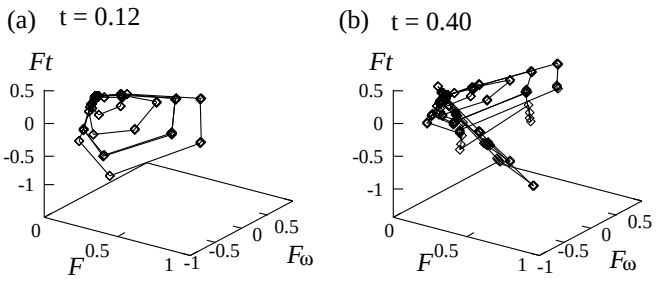

Figure 8: TFGS distributions at t=0.12sec. and 0.40sec. It shows that the distribution is restricted on a plane in one stream interval (left figure), while it forms two plane in two stream interval.

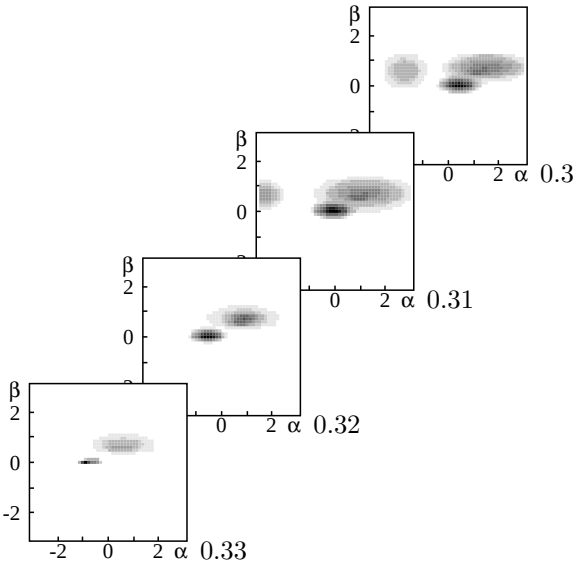

Figure 9: A raw pdf sequence ($Q(\alpha, \beta)$) constructed by Parzen's method (from 0.30 sec. to 0.33 sec.). Two peaks corresponding to the two streams are observed. Some irregular peaks are also exist (at 0.30 and 0.31 sec.).

b) Mixture of Musical Sound and Voice

In order to examine the proposed algorithm in more practical situation, we apply it to a mixture of a musical sound (horn) and a human voice (female: Japanese word 'ASAHI'). This sound is sampled at 10kHz and discretized into 12bit.

The waveform and WED is shown in Fig.13(a) and Fig.13(b), respectively. The sounds are separately

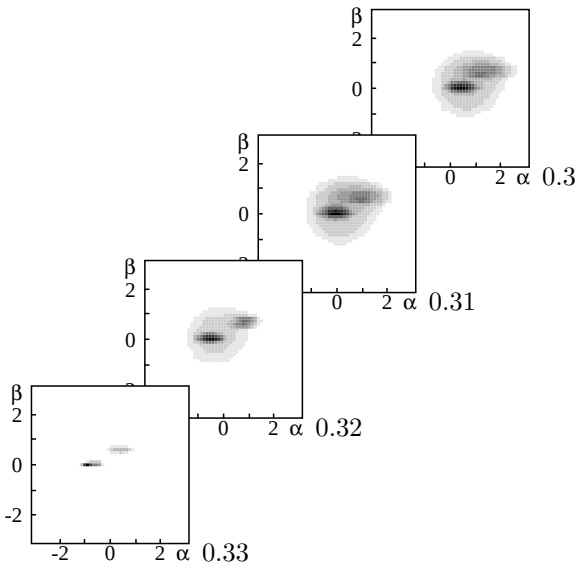

Figure 10: A NPKFed pdf sequence ($P(A_k|Y_k)$). Because we used a simple Gaussian diffusion as a dynamics of (α, β), the irregular peaks are eliminated while dominant peaks are remained.

recorded and mixed in a computer. SN ratio of the musical sound is 3.2dB.

Traces of LCR and PSR of the strong stream are shown in Fig.14(a) and Fig.14(b), respectively. Because the musical sound changes more coherently than the voice, the trace of the largest peak follows the musical sound. Except overlapping components and at offset, the musical sound is almost correctly reconstructed. SN ratio of the musical sound is 11.6dB (i.e. 8.4dB improvement).

6 Discussion and Summary

We proposed a new algorithm for computational auditory scene analysis based on coherency of amplitude and frequency change in wavelet distribution. We estimated LCRs and PSRs using loudness/pitch/timbre decomposition, integrated the sequence of them using NPKF, traced it modes as stream attributes, and reconstructed a single sound stream by gradient space filtering.

We found two problems through the experiments. First, "select the highest" tracing rule fails sometimes because the order of peak heights is rather unstable. Not only the height but also continuity should be introduced to the tracing algorithm. Second, the gradient space filtering often fails at overlapping zones of two streams. It is an interesting subject to consider a reconstruction method which produces an interpolated stream there.

For parameters of NPKF, we used a uniform initial distribution and large diffusion variance in order to facilitate the growth of multiple streams. However, it will be interesting to use, for example, a weighted distribution and small diffusion variance to trace a particular stream. Or, we can set the dynamics so that it can readily jump

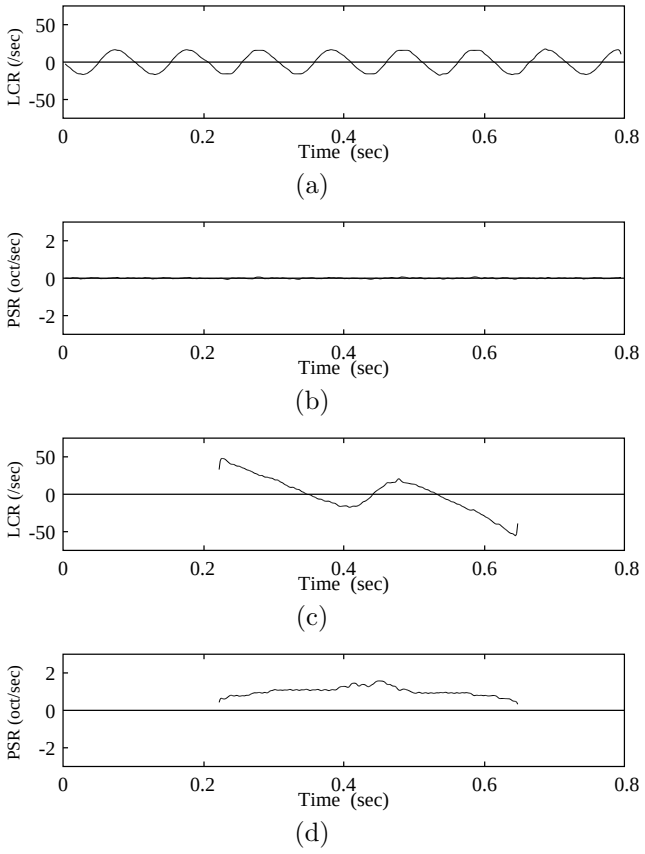

Figure 11: Estimated LCRs and PSRs of multiple streams by tracing peaks of NPKFed pdf. (a) LCR of the strong stream (stream A), (b) PSR of the strong stream, (c) LCR of the weak stream (stream B), (d) PSR of the weak stream. Comparing to Fig.7(c) and (d), It can be confirmed that LCRs and PSRs of each stream are separately estimated.

to suddenly appeared trace by biasing the distribution for large α. This can be a model of a changing focus of attention to incoming streams.

References

[Abe *et al.* , 1996] M.Abe and S.Ando. Application of Loudness/ Pitch/ Timbre Decomposition Operators to Auditory Scene Analysis. *Proc. ICASSP96*, Atlanta, 2646–2649, May 1996.

[Abe *et al.* , 1995] M.Abe and S.Ando. Nonlinear Time-Frequency Domain Operator for Decomposing Sounds into Loudness, Pitch and Timbre. *Proc. ICASSP95*, Detroit, 1368–1371, May 1995.

[Ando, 1994] S.Ando. Density Observation Kalman Filtering. *Proceedings of SICE94*, Tokyo, 225–226, July 1994 (*in Japanese*).

[Bregman, 1990] A.S.Bregman. *Auditory Scene Analysis.* MIT Press, Cambridge, 1990.

[Brown *et al.* , 1994] G.J.Brown et. al. Computational

Auditory Scene Analysis. *Computer Speech & Language*, 8:297–336, 1994.

[Cherry, 1953] E.C.Cherry. Some Experiments on the recognition of speech with one, and two ears. *J. Acoust. Soc. Am.*, 25:975–979, 1953.

[Daubechies, 1990] I.Daubechies. The Wavelet Transform, Time-frequency Localization and Signal Analysis. *IEEE Trans. Information Theory*, 36(5):961–1005, 1990.

[Fukunaga, 1972] K.Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, New York, 1972.

[Jazwinski, 1970] A.H.Jazwinski. *Stochastic Processes and Filtering Theory.* Academic Press, New York, 1970.

[Nakatani, et. al., 1995] T.Nakatani, H.G.Okuno and T.Kawabata. Residue-Driven Architecture for Computational Auditory Scene Analysis. *Proc. IJCAI95*, Montreal, 165–172, 1995.

[Nishi *et al.* , 1996] K.Nishi, S.Ando and S.Aida. Optimum Harmonics Traching Filter for Auditory Scene Analysis. *Proc. ICASSP96*, Atlanta, 573–576, May 1996.

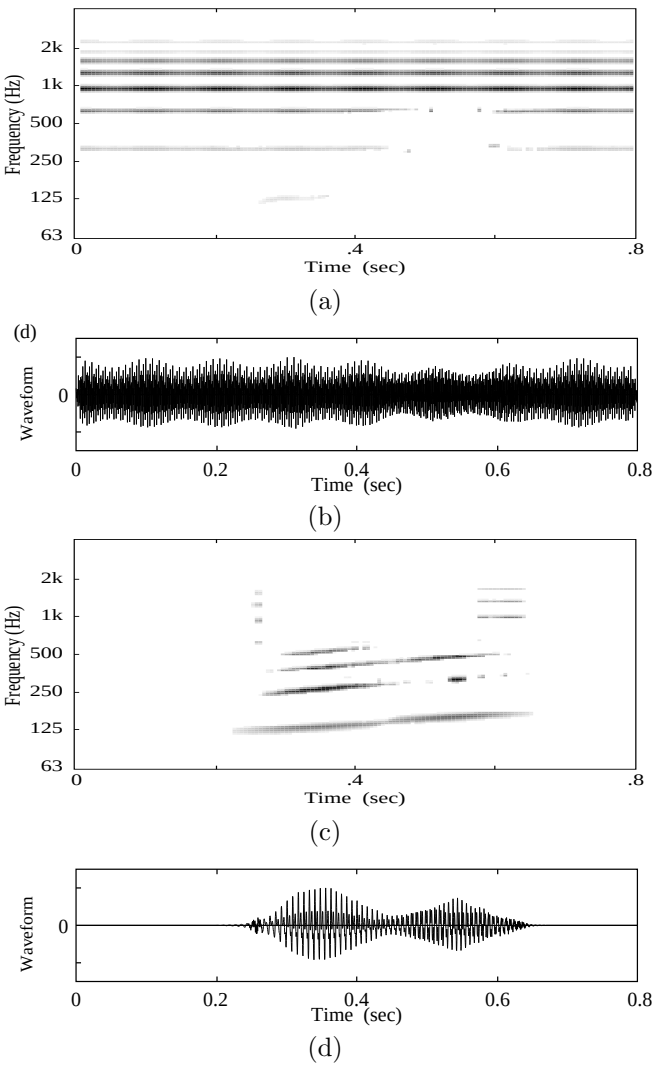

Figure 12: Reconstructed WEDs and waveforms by the gradient space filtering and inverse wavelet transform. (a) reconstructed WED of the strong stream, (b) reconstructed waveform of the strong stream, (c) reconstructed WED of the weak stream, (d) reconstructed waveform of the weak stream. For the strong stream, the WED and the waveform are almost completely reconstructed except overlapping components. For the weak stream, some components of the strong stream are remained at around its onset and offset.

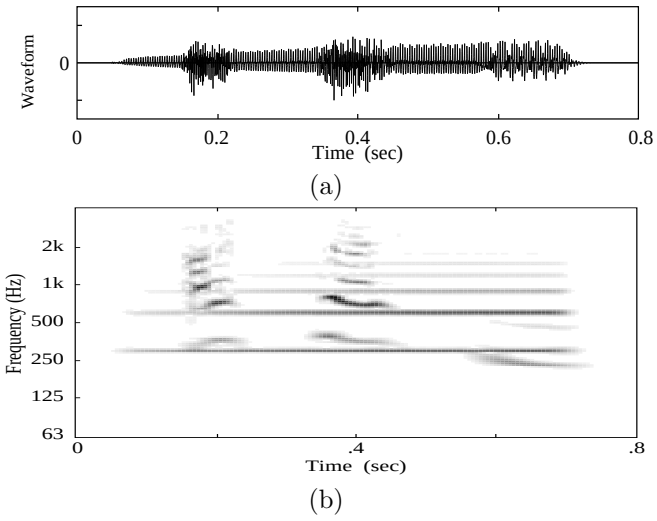

Figure 13: A mixture of a musical sound(horn) and a human voice (female, speaking 'ASAHI'). (a) waveform, (b) WED.

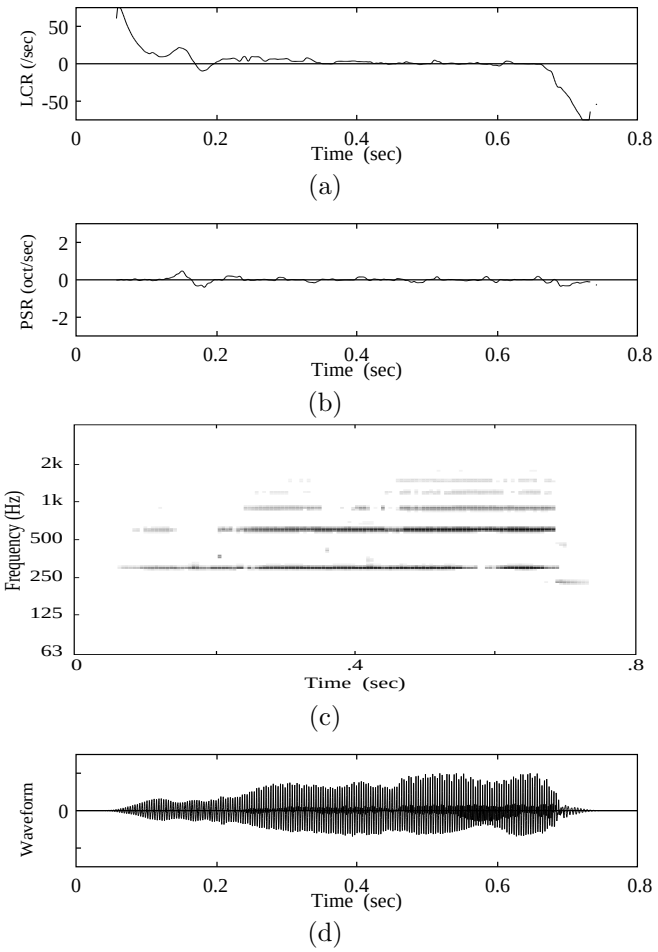

Figure 14: Stream segregation for the sound shown in Fig.13. (a) The trace of the largest peak position (LCR), (b) the trace of the largest peak position (PSR), (c) reconstructed WED, (d) reconstructed waveform. Because the musical sound changes more coherently than the voice, the trace of the largest peak follows the musical sound. Except overlapping components and some small remaining components, the musical sound is almost correctly reconstructed.