APPLICATION OF LOUDNESS/PITCH/TIMBRE DECOMPOSITION OPERATORS TO AUDITORY SCENE ANALYSIS

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

ABSTRACT

We proposed[1] nonlinear operators which decompose a changing energy of sound in wavelet domain into three orthogonal components: i.e., loudness and pitch as coherent changes, and timbre as incoherent change. We showed that they could detect the discontinuity of a single sound stream with excellent temporal resolution and sensitivity. In this paper, we extend the coherency principle so that it can describe and pursue the individual coherency of non-overlapping sound streams in wavelet domain. It is realized by Parzen's non-parametric estimates and Kalman filtering of loudness change rate and pitch shift rate. Using this method, we show some experiments for extraction of the most salient stream from multiple sound streams.

1. INTRODUCTION

For segregating complex sounds from different sources, man utilizes a lot of cues such as periodicity of each sound, synchrony (or coherency) of changes of each frequency component, etc. M. Kubovy showed that man recognizes a complex tone as a stream when all its cues change synchronously, but man perceives an arrival of different stream if some of it change asynchronously[2]. This suggests that human auditory system is very sensitive to incoherency in complex sound which is almost coherent.

We proposed recently some nonlinear operators for decomposing a changing energy of sound into loudness, pitch, and timbre[1]. We showed in the time-frequency gradient space that the loudness change (LC) and pitch shift (PS) distribute on a plane, but the timbre change (TC) does not. According to this difference, the operators could decompose the changing energy of single sound into two coherent components (LC and PS) and incoherent component (TC). In other words, it could detect the incoherency of sound as features of single sound stream. But it could not be applicable to complex sound with multiple streams, because almost everywhere becomes TC due to incoherency in different sources.

To solve this problem and apply these operators to the auditory scene analysis, we extend the coherency principle so that it can describe and pursue the individual coherency of sound streams in wavelet domain. It is realized by Parzen's non-parametric estimates and a novel Kalman filtering method of loudness change rate and pitch shift rate. Using this method, we sequentially extract the most salient stream from multiple sound streams.

2. LOUDNESS/PITCH/TIMBRE DECOMPOSITION OPERATORS

We briefly outline the decomposition operators[1]. Let $f(t)$ be a sound comes from single source. Let

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

$$\psi(t) \equiv A\exp(-\frac{\Delta^2 t^2}{2} - j\Omega_0 t), \quad (2)$$

be a wavelet modulus distribution of $f(t)$. $\psi(t)$ is Gabor function[3] and ω denotes log frequency. Let

$$F_t = \frac{\partial F}{\partial t},\ F_\omega = \frac{\partial F}{\partial \omega} \quad (3)$$

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

$$\Gamma = [t_k - \Delta t < t < t_k + \Delta t] \times [-\infty < \omega < \infty]$$
$$\Delta t : \text{a few sampling intervals}$$

to describe instantaneous changes of $F(t,\omega)$. We call the space spanned by F, F_t, F_ω the time-frequency gradient space (TFGS).

For a distribution of TFGS, it is shown that: 1) it distributes uncorrelatedly for stationary random signals, 2) it reduces to a line or plane where $F(t,\omega)$ is magnifying or shifting coherently (see Figure 1). The second property is because such a sound is expressed in Γ as

$$F(t+dt,\omega) = (1+\alpha dt)F(t,\omega-\beta dt) \tag{4}$$

where α is a loudness change ratio and β is a pitch shift ratio. Therefore

$$F_t(t,\omega) - \alpha F(t,\omega) + \beta F_\omega(t,\omega) = 0 \tag{5}$$

which shows that TFGS is a plane. Taking two basis vectors within the plane and one basis vector orthogonally to the plane, we therefore decompose the TFGS energy (variance) into coherent components (LC and PS) and a miscellaneous incoherent component (TC).

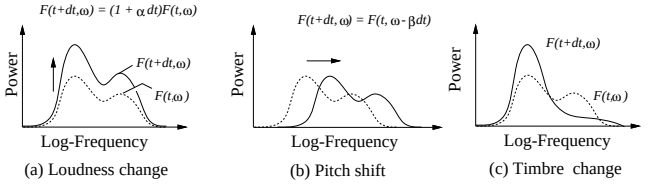

Figure 1: Three fundamental components of instantaneous time-frequency change.

The decomposition is performed by a set of opertors

$$TC = \frac{\det[\mathbf{S}]}{S_{tt}\det[\mathbf{S}_{22}]} \tag{6}$$

$$LC = (1-TC)\cdot\frac{\tilde{\alpha}^2}{\tilde{\alpha}^2+\tilde{\beta}^2} \tag{7}$$

$$PS = (1-TC)\cdot\frac{\tilde{\beta}^2}{\tilde{\alpha}^2+\tilde{\beta}^2} \tag{8}$$

where $\mathbf{S}$ is a correlation matrix of $[F, F_t, F_\omega]$, S_{tt} is its (2,2) component, $\mathbf{S}_{ij}$ is a 2×2 submatrix of $\mathbf{S}$ in which the ith row and jth column of $\mathbf{S}$ is removed, and

$$\tilde{\alpha} = \frac{\det[\mathbf{S}_{12}]}{\det[\mathbf{S}_{22}]}, \quad \tilde{\beta} = -\frac{\det[\mathbf{S}_{32}]}{\det[\mathbf{S}_{22}]}. \tag{9}$$

are least square estimates of α,β. They can be regarded as energy ratio because they are positive and normalized such that

$$LC + PS + TC = 1. \tag{10}$$

An actual sound energy relevant to TC is important for the analysis of following sections. It is computed as

$$J(\tilde{\alpha},\tilde{\beta}) = \frac{\det[\mathbf{S}]}{\det[\mathbf{S}_{22}]}. \tag{11}$$

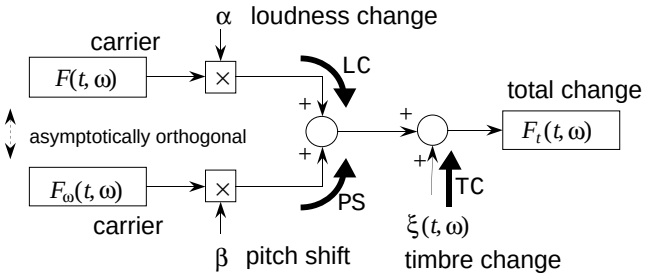

Figure 2: Schema of loudness/pitch/timbre decomposition.

3. COHERENCY PRINCIPLE FOR COMPLEX SOUNDS

We assume a following property of the changing energy of sound from single source.

Assumption 1: *A change of sound from single source is coherent almost everywhere.*

This assumption assures that TFGS of a single source sound is a plane almost always. Under this assumption, we can show the following theorem.

Theorem 1: *1) For non-overlapping sound streams from different sources, TFGS distribution still has coherency such that it is a superposition of individual planes.*
2) For overlapping sound streams from different sources, TFGS distribution becomes as broad as a linear combination of individual distribution.

Proof: For wavelet energy distribution of a sound from two sources, it follows that

$$F(t,\omega) = F_1(t,\omega) + F_2(t,\omega) + \eta(t,\omega) \tag{12}$$

where $\eta(t,\omega)$ is a cross term between each distribution F_1 and F_2[1]. From the assumption, each stream satisfies

$$F_{1t} = \alpha_1 F_1 - \beta_1 F_{1\omega} \tag{13}$$

$$F_{2t} = \alpha_2 F_2 - \beta_2 F_{2\omega}, \tag{14}$$

and from Eq.(12), it follows

$$F_t = (\alpha_1, -\beta_1)\begin{bmatrix} F_1 \\ F_{1\omega}\end{bmatrix} + (\alpha_2, -\beta_2)\begin{bmatrix} F_2 \\ F_{2\omega}\end{bmatrix} + \eta_t. \tag{15}$$

Since the sources are independent, F_1 and F_2 are independent and $(\alpha_1, -\beta_1) \neq (\alpha_2, -\beta_2)$ almost always.

1) If two sound streams are isolated in wavelet domain, (F, F_t, F_ω) is always on one of two planes

$$F_t = \alpha_1 F - \beta_1 F_\omega \tag{16}$$

$$F_t = \alpha_2 F - \beta_2 F_\omega \tag{17}$$

because $F_1F_2 = F_{1\omega}F_{2\omega} = 0$, hence the cross term $\eta(t,\omega)$ is zero.

2) For sound streams with overlaps in wavelet domain, let us suppose the cross terms is ignoreable. Then, the distribution exists in a space is spanned by a linear combination of the two planes (16) and (17). It is indeed a 3-D distribution except F_1, F_2 are coherent. Even if there are cross terms, it generally contributes to raise randomness because they vary more rapidly than F_1, F_2. □

When the power of each source is almost equal and α_i and β_i of each source is independent, a distribution will be homogeneous in TFGS.

4. ESTIMATION OF LOUDNESS CHANGE RATE AND PITCH SHIFT RATE OF MULTIPLE STREAMS

Independently changing streams such as voices from two speakers are usually separated either in time domain or frequency domain. So, we introduce here the next assumption.

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

From this assumption, we can expect that the case of Theorem 1 (1) is a dominant situation. Therefore, in order to describe multiple planes with multiple parameters α and β, we introduce a probability density function (pdf) $P(\alpha,\beta)$ of α,β. By local peaks of it, the multiple plane parameters are expressed. We estimate a pdf $P(\alpha,\beta)$ as follows:

1) Divide Γ into small subregion Γ_n $(n = 0, 1, \ldots)$ as

$$\Gamma_n = [\omega_n < \omega < \omega_n + \Delta\omega] \times [t_0 - \Delta t < t < t_0 + \Delta t] \quad (18)$$

where $\Delta\omega$ is appropriately small.
2) Estimate $\tilde{\alpha}, \tilde{\beta}$ in Γ_n using Eq.(9).
3) Estimate error variances σ_α^2 and σ_β^2 as

$$\sigma_\alpha^2 = \frac{S_{FF} J(\tilde{\alpha},\tilde{\beta})}{L^2 \det[\mathbf{S}_{22}]}, \quad \sigma_\beta^2 = \frac{S_{\omega\omega} J(\tilde{\alpha},\tilde{\beta})}{L^2 \det[\mathbf{S}_{22}]} \quad (19)$$

where $L = \int_{\Gamma_n} dtd\omega$.
4) Estimate a pdf $P(\alpha,\beta)$ using Parzen's method as

$$P(\alpha,\beta) = \sum_{n=0}^{N-1} \frac{1}{2\pi\sigma_\alpha\sigma_\beta N} \exp\left(\frac{(\alpha-\tilde{\alpha})^2}{2\sigma_\alpha^2} + \frac{(\beta-\tilde{\beta})^2}{2\sigma_\beta^2}\right). \quad (20)$$

The reason why we use this algorithm is as follows: 1) because Γ_ns are small, most of them are expected to include only one stream, 2) not informative Γ_n such as no stream, two streams with cross term and locally exponential distribution are are excluded from the estimation of a pdf, and 3) knowledge of a number of sources is not needed.

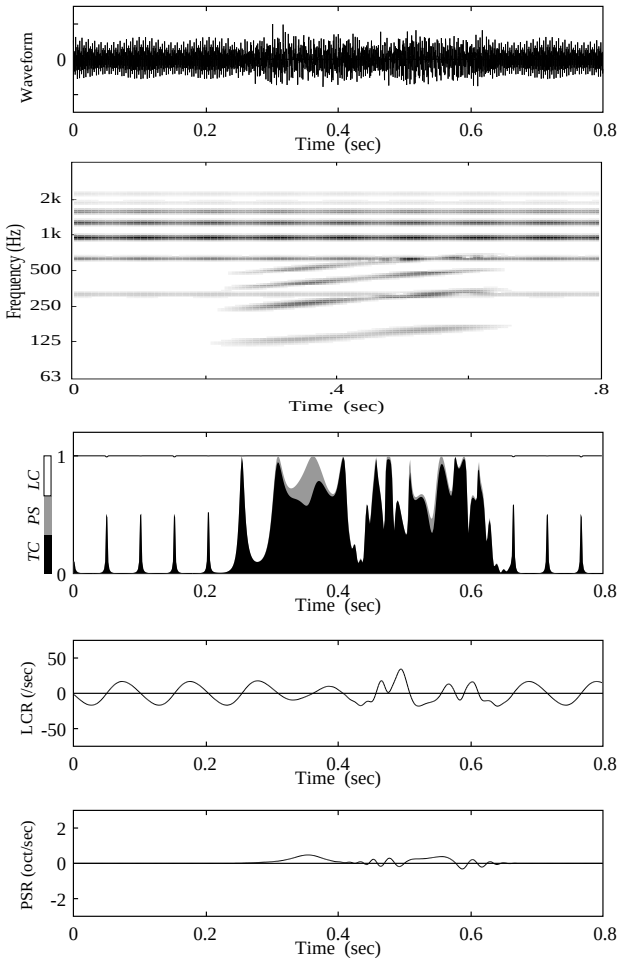

Figure 3: LC/PS/TC decomposition and α,β estimation. (a) waveform, (b) wavelet energy distribution, (c) Loudness/Pitch/Timber decomposition, (d) estimation of α, (e) estimation of β,

5. OPTIMUM SEGREGATION USING NON-PARAMETRIC KALMAN FILTER

In order to exclude an instance which is incompatible with Assumption 1 and 2, and obtain a reliable pdf $P(\alpha,\beta)$, we use non-parametric Kalman filter(NPKF)[4]. NPKF is a discrete time discrete observation successive filter which is performed by three steps as follows:

a) Prediction step
Let $A_k = (\alpha_k, \beta_k)$ be stochastic valuables at time $t = t_k$, let $Y_k = \{y_k, y_{k-1}, \ldots\}$ be observations at time $t = \{t_k, t_{k-1}, \ldots\}$ and let $P(A_k|Y_k)$ be a likelihood function estimated through observations. Then, a prediction of

the pdf at time t_{k+1} is expressed as

$$P(A_{k+1}|Y_k) = \int P(A_{k+1}|A_k)P(A_k|Y_k)dA_k \qquad (21)$$

where a conditional probability $P(A_{k+1}|A_k)$ is determined by a dynamics. Because α and β change smoothly, we can use a simple Gaussian diffusion as a dynamics.

b) Observation step
Using the algorithm discussed in last section, we can obtain a observation probability by $P(A_{k+1}|y_{k+1}) = P(\alpha, \beta)$ in Eq.(20) at time t_{k+1}.

c) Update step
Using a new observation $P(A_{K+1}|y_{k+1})$, we can update the probability function as

$$P(A_{k+1}|Y_{k+1}) = \frac{P(A_{k+1}|y_{k+1})P(A_{k+1}|Y_k)}{\int P(A_{k+1}|y_{k+1})P(A_{k+1}|Y_k)dA_{k+1}} \qquad (22)$$

6. EXPERIMENTS

A) Performance of Loudness/Pitch/Timbre Decomposition Operators

Loudness/pitch/timbre decomposition and estimations of α and β are shown in Figure 3. Since we use whole area along ω axis for Γ, change energy is mainly decomposed into TC and α and β are estimated meaninglessly in the interval of two streams.

B) TFGS Distribution

TFGS distribution at 0.12sec. and 0.4sec. of the sound are shown in Figure 4. The distribution makes a plane for a single stream, while it makes two planes for two different streams.

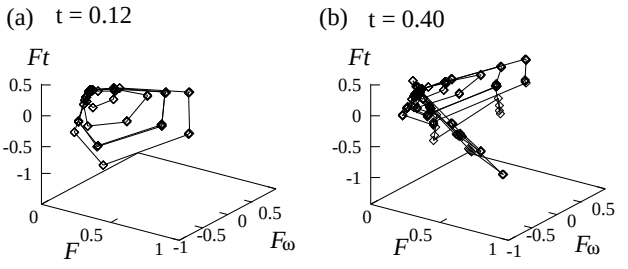

Figure 4: TFGS distributions of the synthesized sound. (a) at 0.12sec. (actually Γ is 0.115sec. to 0.125sec.), (b) at 0.40sec. (0.395sec to 0.405 sec)

C) Estimation of Loudness Change Rate and Pitch Shift Rate of Multiple Streams

By tracking the maxima of NPKF-ed pdf, we can obtain MAP estimation of α and β. They are shown in Figure 5. Change of the sound with larger power and more harmonics is correctly segregated. The observed pdf and the NPKF-ed pdf are shown in Figure 6.

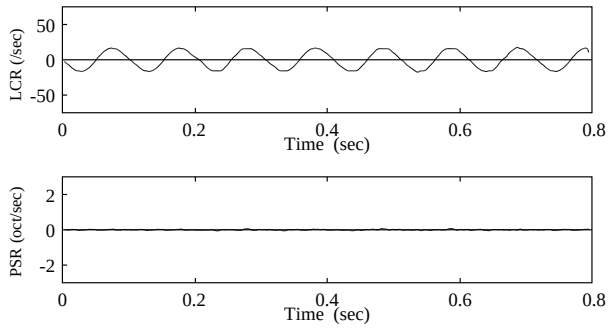

Figure 5: MAP estimated traces of α and β from NPKF-ed pdf. Compare the difference with Figure 3 (d),(e).

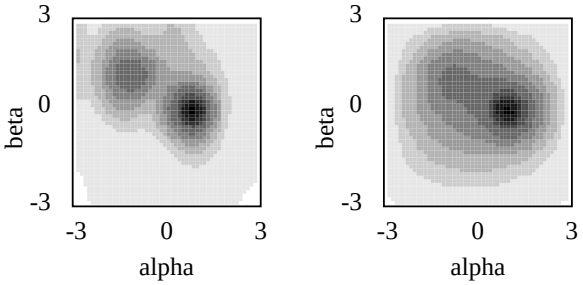

Figure 6: A pdf and an NPKF-ed pdf at 0.40sec.

7. REFERENCES

[1] M.Abe and S.Ando: "Nonlinear Time-Frequency Domain Operator for Decomposing Sounds into Loudness, Pitch and Timbre," Proc. IEEE ICASSP95, 1368/1371 (1995).

[2] A.S.Bregmann: "Auditory Scene Analysis," MIT Press, (1990).

[3] Grossman, et al., "Reading and Understanding Continuous Wavelet Transforms," in *Wavelets*, New York, NY: Springer-Verlag, pp.2-20 (1989).

[4] A.H.Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, (1970).

[5] R.B.Dunn, et al., "On amplitude and frequency demodulation using energy operators," IEEE Trans. Signal Processing, 41, 4, pp.1532 - 1550 (1993).