AUDITORY SCENE ANALYSIS BASED ON TIME-FREQUENCY INTEGRATION OF SHARED FM AND AM

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

This paper describes a new method for computational auditory scene analysis which is based on 1) waveform operators to extract instantaneous frequency (IF), frequency change (FM), and amplitude change (AM) from subband signals, and 2) a voting method into a probability distribution to extract coherency (shared fundamental frequency, shared FM, and shared AM) involved in them. We introduce non-parametric Kalman filtering for the time-axis integration. A consistent AM operator which is independent to frequency change is newly defined. Sharpness of the resultant probability distribution is examined with relating to the definition of the operators and subband bandwidth. We evaluate the performance of the algorithm by using several speech sounds.

1. INTRODUCTION

Segregation of individual streams from a mixture of sound is a fundamental subject of auditory scene analysis. To perceive 'stream', it is mentioned[4] that man utilizes mostly the periodicity (harmonic structure) of sound and the synchrony (we use the term coherency instead) of amplitude change and frequency change in the time-frequency domain. For computational implementation of it, algorithm should essentially be composed of: 1) decomposition of mixed sounds into elementary components, 2) giving attributes to the elementary components (labeling), and 3) grouping them into streams according to the attributes. Following this line, we proposed[1] the use of coherency in loudness change and pitch shift extracted by the loudness/pitch/timbre decomposition operators[2].

In this paper, we propose an alternative method in which we extract stream attributes from waveform of subbands. We label each subband signal by three dominant coherency: 1) coherency in instantaneous frequency (shared fundamental frequency), 2) coherency in relative frequency change (shared FM), and 3) coherency in relative amplitude change (shared AM). We introduce a voting method into a probability density function and non-parametric Kalman filter[1] for segregating a complex sound into individual streams. We show several experimental results of finding the most salient stream and extraction of it from complex sounds.

2. SUBBAND DECOMPOSITION

Let $f(t)$ be a sound and let

$$\hat{f}(t,\omega) = e^{\omega} \int \psi^*(e^{\omega}(\tau - t)) f(\tau) d\tau \qquad (1)$$

be wavelet transform of it, where ω denotes log-frequency. In order to assure fine time-frequency resolution and analyticity, we used Gabor analyzing wavelets

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

where A is a normalization constant, Δ is a half bandwidth, and Ω_c is a center frequency. For our application, a narrow bandwidth ($\Delta = 1/24\Omega_c$) is appropriate to reduce interference between components.

Fig.1(a) shows wavelet amplitude ($|\hat{f}(t,\omega)|$) distribution of a single voice (female: utterance 'realize') and (b) shows waveforms of it (100ms from the beginning of (a)). We can observe: 1) in almost all the subbands, vibrating frequencies maintain an integer multiple relation of the fundamental frequency, 2) the pattern often shows a uniform shift across their adjacent subbands, and 3) a uniform increases/decrease of its amplitude.

3. EXTRACTION OF ATTRIBUTES

3.1. Shared Fundamental Frequency

Using instantaneous amplitude and phase

$$A(t,\omega) = |\hat{f}(t,\omega)|, \quad \phi(t,\omega) = \arg[\hat{f}(t,\omega)] \qquad (3)$$

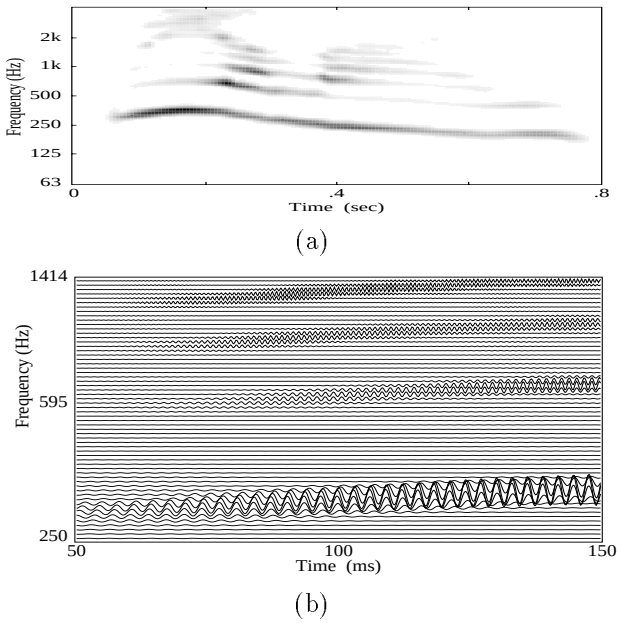

Fig.1: Wavelet distribution of a voice (female, utterance 'realize'). (a) wavelet modulus, (b) waveforms of subbands [50-150ms].

of $\hat{f}(t,\omega)$, respectively, we can calculate instantaneous frequency (IF)[3] as

$$\gamma(t,\omega) = \frac{1}{2\pi}\frac{\partial}{\partial t}\phi(t,\omega). \tag{4}$$

Two scatter diagrams of IFs of subband signals are shown in Fig.2(a) and (b). Although IFs from 1/3 octave filters spread widely, those from 1/24 octave filters concentrate on the fundamental frequency and its integer multiples. These results show the narrow bandwidth (e.g. $\Delta = 1/24\Omega_c$) is better to reduce the interference.

3.2. Shared FM

Let us define a frequency change rate (FCR)

$$\beta(t,\omega) = \frac{\dot{\gamma}(t,\omega)}{\gamma(t,\omega)} \tag{5}$$

of a subband signal. It is a relative measure of frequency change defined in the subband. Fig.2(c) shows a scatter diagram of FCR. Concentration of FCRs is seen as a black line, which is corresponding to the intonation of the utterance.

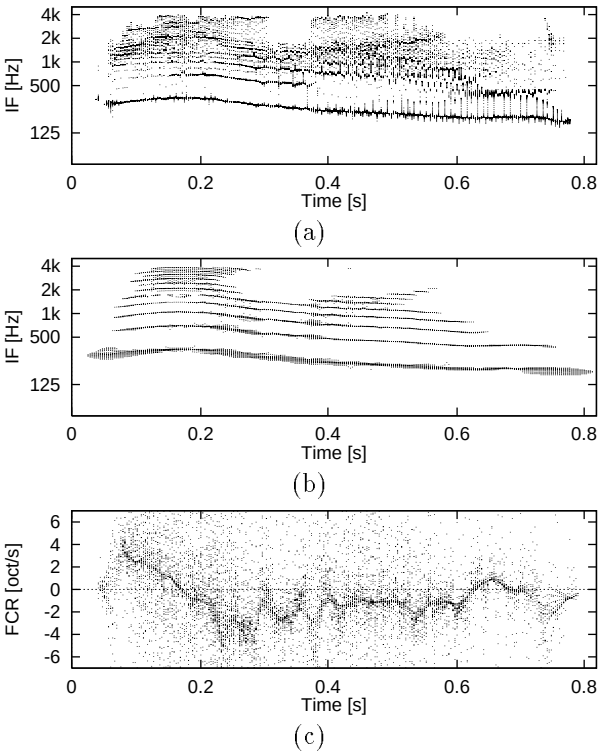

Fig.2: Scatter diagram of (a) instantaneous frequency ($\Delta = 1/3\Omega_c$), (b) instantaneous frequency ($\Delta = 1/24\Omega$), (c) frequency change rates($\Delta = 1/24$). (Utterance 'realize' shown in Fig.1.)

3.3. Shared AM

Let us define, tentatively, an amplitude change rate(ACR) as

$$\alpha(t,\omega) = \frac{\dot{A}(t,\omega)}{A(t,\omega)}. \tag{6}$$

Fig.3(b) shows a scatter diagram of ACR of a synthesized sound shown in Fig.3(a). Although ACRs of subband signals having stable frequency concentrate on one line, those of increasing signals spread widely. This is because instantaneous amplitude of a subband is affected by both amplitude change and frequency change of a stream which shifts across adjacent subbands(Fig.4). To define an ACR independent to frequency change, let

$$\alpha(t,\omega) = \frac{\frac{D}{Dt}A(t,\omega)}{A(t,\omega)}, \tag{7}$$

be a modified ACR, where $\frac{D}{Dt} = \frac{\partial}{\partial t} + v(t)\frac{\partial}{\partial \omega}$ denotes Lagrangian description of a differential[7]. Because velocity of a stream is equal to $\beta(t)$, we can calculate the modified ACR as

$$\alpha(t,\omega) = (\frac{\partial}{\partial t}A(t,\omega) + \beta(t,\omega)\frac{\partial}{\partial \omega}A(t,\omega))/A(t,\omega). \tag{8}$$

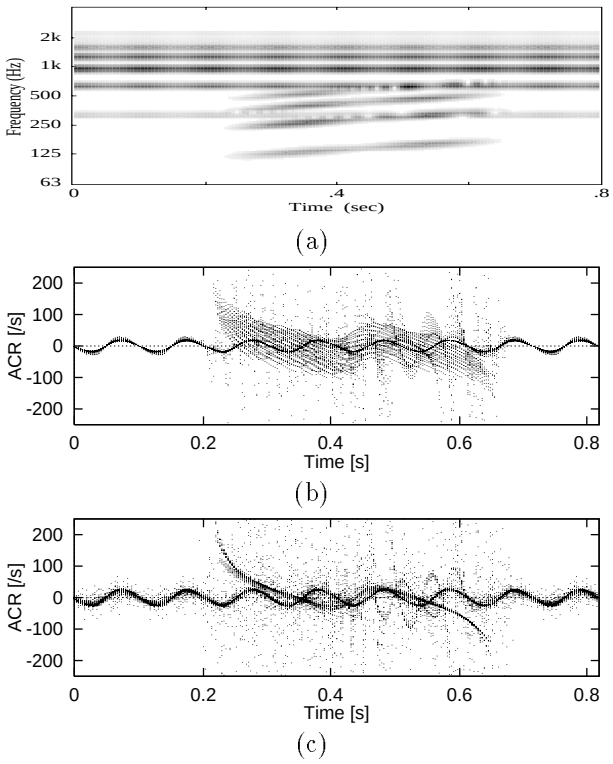

Fig.3: (a) Wavelet modulus of a synthesized sound, (b) scatter diagram of amplitude change rates(ACR), (c) scatter diagram of modified ACRs.

Fig.3(c) shows a scatter diagram of modified ACRs. We can observe that the spreading components in (b) concentrates on one line.

4. FREQUENCY AXIS INTEGRATION: VOTING

We construct a probability density function (pdf) by voting $(\alpha(t,\omega),\beta(t,\omega),\gamma(t,\omega))$ along frequency axis. The highest peak will correspond to the most salient stream.

Let $N(x,\sigma^2)$ be a normal distribution whose mean and variance are x and σ^2, respectively. We construct a pdf at a sampled time t_l as

$$Q_l(\alpha,\beta,\gamma) = \frac{1}{T(\omega_H-\omega_L)}\int_{t_l-T/2}^{t_l+T/2}\int_{\omega_L}^{\omega_H} N(\alpha(t,\omega),\sigma_\alpha^2)N(\beta(t,\omega),\sigma_\beta^2)\Phi(\gamma(t,\omega))dtd\omega, \quad (9)$$

$$\Phi(\gamma(t,\omega)) = (\sum_n \frac{1}{n})^{-1}\sum_n \frac{1}{n}N(\gamma(t,\omega)/n,\sigma_\gamma^2), \quad (10)$$

where $[\omega_L,\ \omega_H]$ and T denote bounds of voting region which we used [65Hz, 4kHz] and 5ms, respectively. In order to integrate integer multiple relations of IFs, we also vote them at the log-frequency $\gamma/n,\ (n=2,3,...)$ as eq.(10). Their weights are set to $1/n$. Variances σ_α, σ_β and σ_γ are small constants which assimilate small differences in subband attributes.

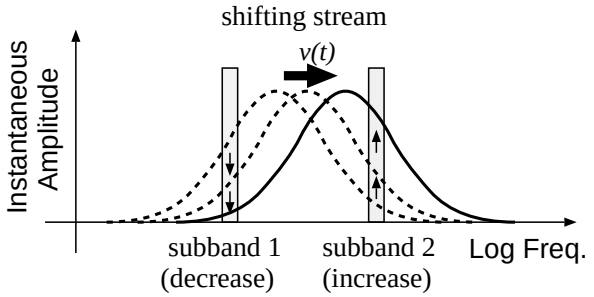

Fig.4: Affection of frequency change of a stream to amplitude change.

5. TIME AXIS INTEGRATION: NON-PARAMETRIC KALMAN FILTER

Because the constructed pdf generally has lots of peaks caused by harmonics, sub-harmonics, noise etc. (see Fig.5(b) in Experiments), we successively integrate the pdf sequence by non-parametric Kalman filter (NPKF)[1].

Let $\mathbf{x}_t=(\alpha,\beta,\gamma)$ be a stochastic variable vector at t, $P(\mathbf{x}_t)$ be a state pdf and $Y_l=\{y_l,...,y_1\}$ be a set of observations before t_l where y_i denotes an observation at t_i. Initially we give an uniform distribution to $P(\mathbf{x}_0)$.

a) Diffusion Step

Because we have no observation $Q_l(\mathbf{x})$ in the interval $(t_{l-1}<t<t_l)$, we simply diffuse $P(\mathbf{x}_{t_{l-1}}|Y_{l-1})$ by convolving a diffusion kernel $P(\mathbf{x}_t|\mathbf{x}_{t_{l-1}})$ which we used a simple Gaussian, as

$$P(\mathbf{x}_t|Y_{l-1}) = \int P(\mathbf{x}_t|\mathbf{x}_{t_{l-1}})P(\mathbf{x}_{t_{l-1}}|Y_{l-1})d\mathbf{x}_{t_{l-1}}. \quad (11)$$

b) Cohesion Step

At $t=t_l$, we integrate (cohese) $Q_l(\mathbf{x})$ and $P(\mathbf{x})$ as

$$P(\mathbf{x}_t|Y_l) = \frac{Q_l(\mathbf{x})P(\mathbf{x}_t|Y_{l-1})}{\int Q_l(\mathbf{x})P(\mathbf{x}_t|Y_{l-1})d\mathbf{x}}. \quad (12)$$

To extract attributes of the most salient stream, we continuously trace the maximum position of P as

$$(\tilde{\alpha}(t),\tilde{\beta}(t),\tilde{\gamma}(t)) = \{\mathbf{x}_t|P(\mathbf{x}_t)\rightarrow \text{maximum}\}. \quad (13)$$

6. RECONSTRUCTION OF SINGLE STREAM

To find subbands consistent with the traced stream attributes, let

$$D_k^2(t,\omega) = \eta_\alpha(\alpha(t,\omega)-\tilde{\alpha}(t))^2 + \eta_\beta(\beta(t,\omega)-\tilde{\beta}(t))^2 + \eta_\gamma(\gamma(t,\omega)-k\tilde{\gamma(t)})^2 \quad (14)$$

be a square distance between the trace and attributes of a subband, where k denotes an integer multiplied to the traced IF and η_α, η_β and η_γ denote weight constants. Let

$$G(t,\omega) = \sum_k \frac{1}{(D_k(t,\omega)/D_0)^N + 1} \quad (15)$$

be a compatibility function, where D_0 determines admissible error size and N determines sharpness of cut-off. Then we can find signals which forms the stream by multiplying G to $\hat{f}$. We reconstruct a sound of the stream by inverse wavelet transform as

$$\tilde{f}(t) = C\int e^{\frac{1}{\omega}} G(\tau,\omega)\hat{f}(\tau,\omega)\psi(e^{\omega}(\tau-t))d\tau d\omega, \quad (16)$$

where C denotes a normalization constant.

7. EXPERIMENTS

Fig.5(a) shows a wavelet modulus of a mixture of two voices, one of which is the voice shown in Fig.1 and the other is a word 'weekday' uttered by a male speaker. The range of wavelet transform is 6 octave from 62.5Hz to 4kHz. SNR of the female's voice is 1.5dB. Fig.5(b) shows a pdf sequence constructed by voting. Since we cannot display a 3-dimensional distribution, we show a result along IF-axis, which is computed by $q(t,\gamma) = \int Q_t(\alpha,\beta,\gamma)d\alpha d\beta$. Fig.5(c) shows a NPKFed pdf sequence (again for only IF-axis). One peak is continuously extracted, which corresponds to fundamental frequency of the female's voice. Fig.5(d) shows a reconstructed wavelet modulus of the traced stream. Comparing to Fig.1, we can confirm that the voice 'realize' is almost perfectly extracted. SNR of the female's voice is improved to 12.7dB.

8. REFERENCES

[1] M.Abe and S.Ando: "Application of Loudness/Pitch/Timbre Decomposition Operators to Auditory Scene Analysis," Proc. IEEE ICASSP96, 2646/2649 (1996).

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

[3] B. Boashash: "Estimating and Interpreting the Instantaneous Frequency of Signal — Part I: Fundamental," Proc. IEEE, Vol. 80, No. 4, 520/538 (1992).

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

[5] G.J.Brown et. al., "Computational Auditory Scene Analysis," Comp. Speech & Lang., 8, 297/336 (1994).

[6] T.Nakatani et.al., "A Computational Model of Sound Stream Segregation with Multi-Agent Paradigm," Proc. IEEE ICASSP95, 2671/2674 (1995).

[7] F.M.White: "Fluid Mechanics," Mc-GrawHill (1979).

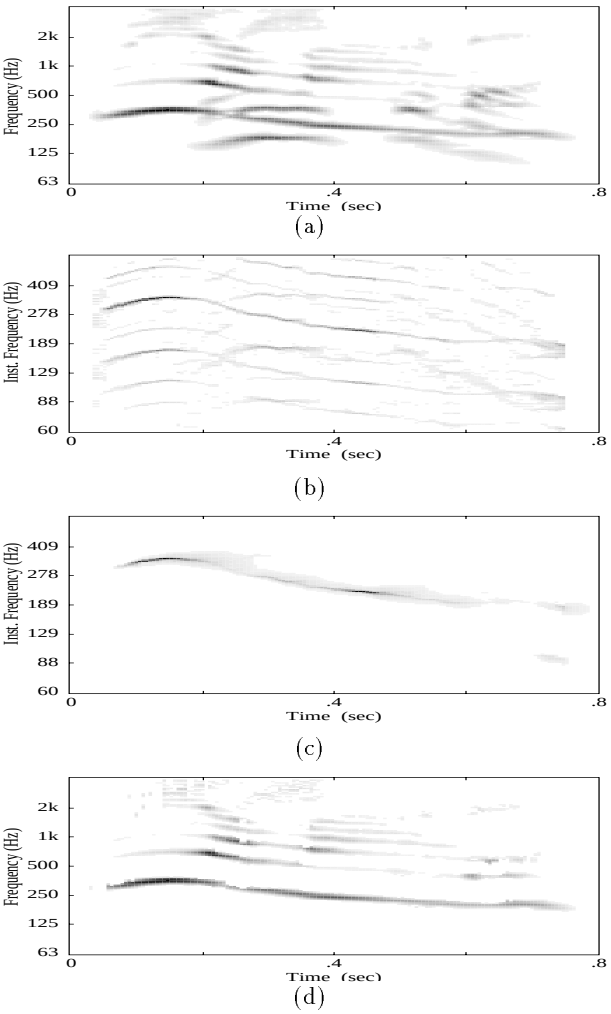

Fig.5: Experimental results for concurrent voices. (a) wavelet modulus of the mixed voice ('realize' (female) and 'weekday' (male)), (b) pdf sequence constructed by voting (showing only along IF axis), (c) NPKFed pdf sequence (IF-axis), (d) reconstructed wavelet modulus.