ABSTRACT
In this paper, we propose a method for decomposing instantaneous changes of sounds into three energy components, i.e., loudness, pitch, and timbre. These operators are derived from an eigenstructure analysis of the time-frequency gradient space (a 3-D space spanned by a modulus and partial derivatives of a wavelet transform). By several experiments, we found that they have superior resolution and sensitivity for segmenting speech into phonemes, characterizing dynamical nature of musical sounds, and so on.

1. INTRODUCTION
As is well known, the first signal processing in human auditory system is short time-frequency analysis in the basilar membrane. Also in signal analysis, the time-frequency domain has been successfully defined by using filterbanks, sound-spectrograms, short time Fourier transforms, Wigner distributions, and so on. Especially, rapidly growing method of the wavelet transforms[1] (WT) provides well-organized bases to describe the time-frequency domain[2] and a log-linear nature of the human auditory processing in cochlea.

The main purpose of our research is to develop simple and informative measures for describing time-varying natures of musical sound and speech. We construct a time-frequency energy distribution by using the wavelet transform. We show an instantaneously changing energy of it can be decomposed into three types: 1) loudness change, 2) pitch shift, and 3) timbre change according to the coherency in power magnification and the coherency in frequency shift. These three types of changes may correspond to three fundamental attributes of a musical sound in acoustics and psychology[3]. To construct nonlinear operators for the decomposition, we relate these three change types with an eigenstructure in the time-frequency gradient space. By several experiments, we found that they have superior resolution and sensitivity for segmenting speech into phonemes, characterizing dynamical nature of musical sounds, and so on.

2. CONSTRUCTION OF LOG-LINEAR TIME-FREQUENCY DOMAIN
Let $f(t)$ be a sound signal, and let $\psi(t)$ be an analyzing wavelet. Then the wavelet distribution $\hat{f}(t, \omega)$ of a signal $f(t)$ is expressed as

$$\hat{f}(t, \omega) = \int_{-\infty}^{\infty} e^{\omega^*(\omega - \tau)} f(\tau) d\tau. \quad (1)$$

‘$t$’ and ‘$\omega$’ denote time and log-frequency respectively. If the wavelet $\psi$ satisfy the QMF conditions[2], it provides a dyadic bases set of the time-frequency space. But we use here a Gabor function

$$\psi(t) = A \exp \left( - \frac{\Delta^2 t^2}{2} - j \Omega_0 t \right) \quad (2)$$

where $\Omega_0$ is the center frequency of highest frequency band, $2\Delta$ is the half bandwidth of $\psi$ in frequency domain. It is because in such an analyzing wavelet, real and imaginary parts of the transform are an approximate Hilbert transform pair, therefore it is suitable for computing instantaneous energy such as

$$F(t, \omega) = |\hat{f}(t, \omega)|^2. \quad (3)$$

3. COOPERATION INSTANTANEOUS CHANGES IN TIME-FREQUENCY GRADIENT SPACE
3.1. Loudness Change and Pitch Shift
By the term ‘coherency’, we mean a kind of orderness in changes of the time-frequency distribution. Let $\Gamma$ be a square region in time frequency space such that $\Gamma = [-\Delta t < t < \Delta t] \times [0 < \omega < \infty]$. The most important two coherency will be:
1. coherency in magnification — wavelet energy is magnifying uniformly in $\Gamma$, i.e. along an entire $\omega$ axis and in a short interval along $t$ axis.

2. coherency in shift — wavelet energy is shifting along $\omega$ axis uniformly in $\Omega$.

We call the former coherency as ‘Loudness Change’ (Fig.1(a)), and the latter one as ‘Pitch Shift’ (Fig.1(b)). Of course, there are many changes other than the above two types of changes. We define here the rest of changes as ‘Timbre change’ which can never be described by these two bases.

$$F(t+dt, \omega) = (1 + \alpha dt)F(t, \omega)$$

where $\alpha$ (loudness change ratio) and $\beta$ (pitch shift ratio) are constants and $dt$ denotes very short time. Letting $dt \to 0$, we get a differential equation

$$F_t(t, \omega) - \alpha F(t, \omega) + \beta F_\omega(t, \omega) = 0. \quad (5)$$

where $F_t$ and $F_\omega$ denote time and log-frequency partial derivatives of $F$ respectively.

Although they distribute three dimensionally generally (Fig.2(a)), eq.(5) shows it reduces to a plane if the magnification and/or shift changes are dominant (Fig.2(b)). That is to say, we can classify the changes by using the TFGS distribution.

The most advantage of the use of TFGS is that it has a complete instantaneous nature in definition and computation, and it can gather coordinated changes from separated region in time-frequency space.

### 4. ANALYSIS ON MULTICOMPONENT AM/FM SIGNALS

A multicomponent AM/FM signal[4] as

$$f(t) = \sum_k a_k(t)e^{j\phi_k(t)} \quad (6)$$

is often used as a basic model for a musical sound or speech. Under following assumptions, it can be shown that the wavelet energy distribution ($F(t, \omega) = |f(t, \omega)|^2$) of $f(t)$ is calculated as

$$F(t, \omega) = \sum_k a_k^2(t)e^{-\frac{1}{2}\sigma^2(c_k\phi_k(t) - \Theta_k)^2}. \quad (7)$$

Assumptions:

1. Each component $a_k(t)e^{j\phi_k(t)}$ can be regarded as locally monochromatic asymptotic signals[5], i.e. $\dot{a}_k/a_k$ and $\dot{\phi}_k/\phi_k$ are much smaller than the carrier frequency $\phi(t)$.

2. Cross terms between each component in wavelet energy distribution can be removed.

The former assumption is general for AM/FM model. For sounds with pitch, the latter can be satisfied after low-pass-filtering.

In this case, we can prove that eq.(7) is the solution of eq.(5) if and only if

$$a_k(t) = c_k a(t), \quad \dot{\phi}_k(t) = d_k \dot{\phi}(t) \quad (8)$$

where $a(t)$ and $\phi(t)$ are a common envelope and a common instantaneous frequency respectively. This implies that $F(t, \omega)$ is a solution of eq.(5) when all envelopes and instantaneous frequencies of the component signals are changing coherently.

### 5. CONSTRUCTION OF NONLINEAR OPERATORS

To characterize the TFGS distribution, we define multidimensional correlation coefficients between $F, F_t, F_\omega$. 

---

**Fig.1** Loudness change and pitch shift

3.2. Time-Frequency Gradient Space

Mathematically, the coherent magnifying and shifting are described as the relation

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

where $\alpha$ (loudness change ratio) and $\beta$ (pitch shift ratio) are constants and $dt$ denotes very short time. Letting $dt \to 0$, we get a differential equation

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

where $F_t$ and $F_\omega$ denote time and log-frequency partial derivatives of $F$ respectively.
First, we define the correlation matrix \( S \) as follows.

\[
S = \begin{bmatrix}
S_{FF} & S_{Ft} & S_{tF} & S_{tFt} & S_{tt} & S_{tFt} & S_{tt}
\end{bmatrix}
\]

\[
= \int \Gamma \begin{bmatrix}
F & F_t & F_\omega \\
F_t & F_t & F_\omega \\
S_{FF} & S_{Ft} & S_{tF} & S_{tFt} & S_{tt} & S_{tFt} & S_{tt}
\end{bmatrix} dt d\omega
\tag{9}
\]

where \( \Gamma \) denotes actually a \( \Delta t \) (one sampling interval) times \( \forall \omega \).

For notational simplicity, we define 2 \( \times \) 2 submatrices \( S_{ij} \) \((i, j = 1, 2, 3)\) which are built by taking out the \( i \)th row and \( j \)th column from \( S \). For example, \( S_{33} \) denotes the matrix as follows.

\[
S_{33} = \begin{bmatrix}
S_{FF} & S_{Ft} & S_{tF} & S_{tFt} & S_{tt} & S_{tFt} & S_{tt}
\end{bmatrix}
\tag{10}
\]

The least square plane can be estimated by minimizing eq.(11).

\[
J(\alpha, \beta) = \int \Gamma (F_t - \alpha F + \beta F_\omega)^2 dt d\omega
\tag{11}
\]

Differentiating \( J \) with respect to \( \alpha \) and \( \beta \), we can estimate \( \hat{\alpha} \) and \( \hat{\beta} \) as

\[
\hat{\alpha} = -\frac{\text{det}[S_{12}]}{\text{det}[S_{22}]}, \quad \hat{\beta} = -\frac{\text{det}[S_{32}]}{\text{det}[S_{22}]}
\tag{12}
\]

Total error from the estimated plane (that is \( J \)) can be thought as quantity of timbre change component. It can be calculated as

\[
J(\hat{\alpha}, \hat{\beta}) = \frac{\text{det}[S]}{\text{det}[S_{22}]}
\tag{13}
\]

Normalizing \( J \) by total change energy \( \int F_t^2 dt d\omega = S_{tt} \) (see Fig.3), we can define timbre change detection operator \( TC \) as

\[
TC = \frac{\text{det}[S]}{S_{tt}\text{det}[S_{22}]} \quad (0 \leq TC \leq 1)
\tag{14}
\]

which becomes 0 when the changes are purely loudness and/or pitch shift, and becomes 1 when they are not loudness or pitch shift at all.

From the above definition, the remainder \((1 - TC)\) expresses a ratio of loudness change and pitch shift. Dividing it according to the gradient \((\alpha, \beta)\) of the distribution plane, we define operators \( LC \) and \( PS \) for detecting loudness change and pitch shift respectively as

\[
LC = (1 - TC) \cdot \frac{\hat{\alpha}^2}{\hat{\alpha}^2 + \hat{\beta}^2} \quad (0 \leq LC \leq 1)
\tag{15}
\]

\[
PS = (1 - TC) \cdot \frac{\hat{\beta}^2}{\hat{\alpha}^2 + \hat{\beta}^2} \quad (0 \leq PS \leq 1).
\tag{16}
\]

All of these operators are normalized in \([0,1]\), and

\[
LC + PS + TC = 1.
\tag{17}
\]

Therefore they can be regarded as an energy ratio of the three types of sound changes.

### 6. EXPERIMENTS

By several experiments, we found that the proposed operators have the following properties for musical sounds and speech.

<table>
<thead>
<tr>
<th>Musical Sound (Horn)</th>
<th>Stable Part</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rising Part</td>
<td>Stable Part</td>
</tr>
<tr>
<td>Pitch Fluctuation</td>
<td>Vibrato</td>
</tr>
<tr>
<td>Higher Harmonics</td>
<td>Vibrato/Dip</td>
</tr>
<tr>
<td>Vibrato</td>
<td>Vibro/Pea/Dip</td>
</tr>
<tr>
<td>Str</td>
<td>Large</td>
</tr>
<tr>
<td>LC</td>
<td>1.0 - 0.8</td>
</tr>
<tr>
<td>PS</td>
<td>0.7 - 0.4</td>
</tr>
<tr>
<td>TC</td>
<td>0.1 - 0.2</td>
</tr>
<tr>
<td>Str</td>
<td>Large</td>
</tr>
</tbody>
</table>

**Speech**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Boundary</th>
<th>Consonant</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>0.8 ~ 0</td>
<td>0 ~ 0.5</td>
</tr>
<tr>
<td>PS</td>
<td>0.2 ~ 0</td>
<td>0 ~ 0</td>
</tr>
<tr>
<td>TC</td>
<td>0 ~ 2</td>
<td>0 ~ 0.3</td>
</tr>
<tr>
<td>Str</td>
<td>Small</td>
<td>Large</td>
</tr>
</tbody>
</table>

**Table 1 Performance of operators**

These results show that possible applications of these operators are phoneme segmentation and characterization of musical sounds.
Fig. 4 Experimental results for horn
(a) Waveform, (b) Wavelet energy (gray scale), (c) Operator outputs (LC: white, PS: gray, TC: black)

Rising part (0.1 – 0.3sec):
The sound is rising and loudness change is dominant. But because of frequency fluctuation at the beginning, pitch shift are seen. It is said that harmonics of brassy sound (ex. trumpet, horn) have some delay in higher harmonics rising. In this case, they are rising at 0.17sec (P) and 0.26sec (Q). At these moments, spectrum structures of the sound are changing and are detected by timber change.

Stable Part (0.3sec – 0.7sec):
Both loudness change and timbre change are dominant because of existence of vibrato. The peaks of timbre change are corresponding to the peaks and dips of vibrato.

7. SUMMARY

We propose a set of non-linear operators which decompose instantaneous changes of sound. The idea depends on coherency/incoherency of magnification and log-frequency shift in wavelet domain. To create decomposition method for another type of coherency (now, it is in timbre change) and integration of information along time axis will be next interest.

Fig. 5 Experimental results for speech ‘my name.’
(a) Waveform, (b) Wavelet energy (c) Operator outputs.

Timbre change is dominant around consonant ‘m’(0.2sec), after and before consonant ‘n’(0.6sec) and between vowels. Pitch shift and loudness change are dominant in a vowel.

8. REFERENCES