A music transcription algorithm applying state estimation and parameter identification on the time-frequency plane

Yukio Fukayama[†] and Daisuke Tanaka^{††}

[†] Hiroshima Institute of Technology

2-1-1 Miyake, Saeki-ku, Hiroshima City, Hiroshima 731-5193, Japan

y.fukayama.ik@it-hiroshima.ac.jp

^{††} Nara Institute of Science and Technology

8916-5 Takayama-cho, Ikoma City, Nara, 630-0192, Japan

daisuke-t@is.naist.jp

Abstract

An algorithm for music transcription system that displays notes and rests on the five lines of music sound is proposed. It is able to process multi-voice music which includes chords, superimposed melodies etc. The system first projects the sound onto the time-frequency plane applying Gabor wavelet transform, and then identifies tempo, the length of notes etc. with referring autocorrelation of the projection. At next, it estimates actually played pitch names that correspond to the fundamental frequencies of simultaneously played tones. It applies a state estimation technique to cope with considerable harmonics which depend upon instruments.

1 Introduction

An algorithm for music transcription system that displays notes and rests of music sounds on the staffs, which are the five lines with the G clef etc., shall be discussed. The system should process multi-voice music which includes simultaneously sounding tones of chords, superimposed melodies and so on. In addition, the music is usually played by an instrument with tones which include considerable harmonics. The system identifies the length of notes etc. with referring to the time-frequency plane, and then estimates actually played pitch names under the harmonics applying a state estimation technique.

The system first projects the music sound onto the

time-frequency plane applying Gabor wavelet transform and identifies the length of notes and rests with referring to the projection. Since Fourier transform of autocorrelation of the projection includes the power spectrum of tone envelope of each pitch, the tempo of the music, which is usually measures quarter-note counts in a minute, is sixty times as the principal frequency that gives the amplitude maximum of the spectrum. Thus the note value, e.g. eighth-note etc., the duration of which is a fraction of the integer of or an integer multiple of that of the quarter-note, is identified with referring to the projection on the time-frequency plane considering the tempo.

The system has to extract multiple fundamental frequencies from bunch of harmonics, since actually played pitch names, $\dots$, A, A#, B, etc., correspond only to fundamental frequencies of the tones. Unlike transcription of single-voice music by an instrument where the lowest frequency components of its spectrum is the fundamental frequency, or that of multi-voices music with the sinusoidal waves where its spectrum consists of actually played pitch names only, the system is applied Kalman filter with sixteenth-note duration as the time step to estimate the multiple fundamental frequencies under considerable harmonics which depend upon instruments.

This report describes the projection onto the time-frequency plane in section 2, modeling of music as well as identification of the note value in section 3, and estimation of multiple fundamental frequency in section 4. Finally section 5 shows some example of music transcriptions applying the proposed techniques.

2 Gabor wavelet analysis of music

Music signals $y(t)$ without DC component and their Fourier transform $Y(\omega)$ are modeled as follows:

$$\begin{aligned} y(t) &= \sum_{m=1}^M [y_m(t) e^{j\omega_m t} + y_m^*(t) e^{-j\omega_m t}] \\ &= 2 \sum_{m=1}^M |y_m(t)| \cos(\omega_m t + \angle y_m(t)) \\ &\xleftrightarrow{\text{FT}} Y(\omega) \\ &= \sum_{m=1}^M [Y_m(\omega - \omega_m) + Y_m^*(\omega + \omega_m)], \end{aligned} \quad (1)$$

where $\xleftrightarrow{\text{FT}}$ and superscript $*$ respectively denote Fourier transform pair and complex conjugate. In addition, stochastic process $y_m(t) \xleftrightarrow{\text{FT}} Y_m(\omega)$ contains random phase $\angle y_m(t)$ and information about tempo, rhythm, dynamics etc. on amplitude $|y_m(t)|$ at angle frequency $\omega_m = 2\pi f_m$, $f_m = f_0 2^{\frac{m}{12}}$ of pitch number $m \in \mathbf{N}$ with $f_0 = 8.18[\text{Hz}]$ and time t . For instance, a standard piano has pitch number range $21 \leq m \leq 108$, where frequencies $\dots, f_{69}, f_{70}, f_{71}, \dots$ correspond to pitch names $\dots, A, A\#, B$, and $\dots$, just above middle C on the keyboard.

Then Gabor wavelet $g_m(t)$ and its Fourier transform $G_m(\omega)$ are as follows:

$$\begin{aligned} g_m(t) &\triangleq \frac{1}{\sqrt{|a_m|}} g\left(\frac{t}{a_m}\right) = \frac{1}{\sqrt{|a_m|} \sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \left(\frac{t}{a_m}\right)^2 + j\omega_m t} \\ &\xleftrightarrow{\text{FT}} G_m(\omega) = \sqrt{|a_m|} e^{-\frac{1}{2}\sigma^2 a_m^2 (\omega - \omega_m)^2}, \end{aligned} \quad (2)$$

where the scale factor $a_m = \omega_0 / \omega_m = 2^{-\frac{m}{12}}$ for angle frequency ω_m and mother wavelet is:

$$g(t) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2} + j\omega_0 t}. \quad (3)$$

Thus Gabor wavelet transform $\eta_m(t)$ is the projection of time signal $y(t)$ onto the time-frequency plane (m, t) with angle frequency ω_m as follows:

$$\begin{aligned} \eta_m(t) &\triangleq \int_{-\infty}^{\infty} y(\tau) g_m^*(\tau - t) d\tau = y(t) * g_m(t) \\ &= \frac{1}{\sqrt{2\pi\sigma^2 |a_m|}} \times \int_{-\infty}^{\infty} y(\tau) e^{j\omega_m \tau - \frac{(t-\tau)^2}{2a_m^2 \sigma^2} + j\omega_m (t-\tau)} d\tau \\ &\xleftrightarrow{\text{FT}} \Xi_m(\omega) = G_m(\omega) Y(\omega) \\ &\approx \sqrt{|a_m|} e^{-\frac{1}{2}\sigma^2 a_m^2 (\omega - \omega_m)^2} Y_m(\omega - \omega_m), \end{aligned} \quad (4)$$

where operator $*$ denotes convolution and windowing operation around angle frequency ω_m to $Y(\omega)$ at the last line provides the following approximation:

$$\eta_m(t) \approx \frac{e^{j\omega_m t}}{\sqrt{2\pi\sigma^2 |a_m|}} \int_{-\infty}^{\infty} y_m(\tau) e^{-\frac{(t-\tau)^2}{2a_m^2 \sigma^2}} d\tau. \quad (5)$$

For example, real part of Gabor wavelet $g_m(t)$ is shown in Fig.1 where component number i for η_i , ($0 \leq i \leq 87$) corresponds to standard piano keyboard, or $i = m - 21$.

Furthermore, magnitude of the transform can be regarded as the amplitude $|y_m(t)|$ of the pitch number m component of the music signal through convolution with Gaussian density:

$$\begin{aligned} |\eta_m(t)| &\approx \frac{1}{\sqrt{2\pi\sigma^2 |a_m|}} e^{-\frac{t^2}{2a_m^2 \sigma^2}} * |y_m(t)| \\ &\approx |y_m(t)|, \end{aligned} \quad (6)$$

where properly small variance σ^2 is chosen.

Fig.1 Gabor wavelet (real part)

3 Modeling of music

The tempo of music that is expressed as the counts of a quarter notes per minutes is evaluated as sixty times of the principal frequency f_T of the music signal envelop:

$$f_T = \frac{1}{2\pi} \arg \max_{\omega} \Gamma(\omega) \\ \simeq \frac{1}{2\pi} \arg \max_{\omega} \int_0^{\infty} \rho(t) \cos(\omega t) dt, \quad (7)$$

where power spectrum $\Gamma(\omega)$ of the envelop is Fourier transform of autocorrelation $\rho(t)$ that is approximated by Gabor wavelet transform:

$$\rho(t) \simeq \left\langle \sum_{m=1}^M |\eta_m(\tau)| \sum_{m=1}^M |\eta_m(\tau - t)| \right\rangle. \quad (8)$$

Observation vector $\eta(k) \in \mathbf{R}^p$ is given by magnitude of the Gabor wavelet transform as follows:

$$\eta(k) = \left(|\eta_{l_0}(k\Delta t)| \quad \cdots \quad |\eta_{l_0+p-1}(k\Delta t)| \right)^T, \quad (9)$$

where sampling interval Δt is usually the duration of sixteenth note, a quarter of that of quarter notes:

$$\Delta t = \frac{1}{4f_T}. \quad (10)$$

A piece of music is modeled as a stochastic system with the state equation including vector function $\mathbf{f} \in \mathbf{R}^n$. It is supposed to depend upon time invariant parameter α that represents key and to be driven by i.i.d. (independent identically distributed) noise $\mathbf{w} \in \mathbf{R}^n$ as follows:

$$\xi(k+1) = \mathbf{f}(\xi(k+1), \alpha) + \mathbf{w}(k), \quad (11)$$

where state vector $\xi(k) \in \mathbf{R}^n$ whose component ξ_i is strength of actually played note number $i + l_0$ is set for, typically, $l_0 = 21$ and $n = 88$ in case of piano:

$$\xi(k) = \left(\xi_0(k) \quad \cdots \quad \xi_{n-1}(k) \right)^T. \quad (12)$$

The observation equation is assumed to be contaminated by i.i.d. noise $\mathbf{v} \in \mathbf{R}^p$ as follows:

$$\eta(k) = \mathbf{h}(\xi(k), \gamma) + \mathbf{v}(k), \quad (13)$$

where vector function $\mathbf{h} \in \mathbf{R}^p$ is supposed to be depending upon time invariant parameter γ that represents instrument.

The noises in (11) and (13) are assumed to be zero means $E\{\mathbf{w}(k)\} = \mathbf{0}_n$, $E\{\mathbf{v}(k)\} = \mathbf{0}_p$, non-Gaussian in general, and to have the following white properties for

$\forall k, j \in \mathbf{Z}$:

$$E\left\{ \begin{pmatrix} \mathbf{w}(k) \\ \mathbf{v}(k) \end{pmatrix} \begin{pmatrix} \mathbf{w}^T(j) & \mathbf{v}^T(j) \end{pmatrix} \right\} \\ = \begin{pmatrix} \mathbf{W} & \mathbf{U} \\ \mathbf{U}^T & \mathbf{V} \end{pmatrix} \delta_{k,j}, \quad (14)$$

where $\mathbf{W} \in \mathbf{R}^{n \times n}$, $\mathbf{V} \in \mathbf{R}^{p \times p}$ and $\mathbf{U} \in \mathbf{R}^{n \times p}$ denote covariance matrices.

Then, applying perturbed state vector $\mathbf{x} \triangleq \xi - \bar{\xi} \in \mathbf{R}^n$ and observed vector $\mathbf{y} \triangleq \eta - \bar{\eta} \in \mathbf{R}^p$, around balanced points $\bar{\xi}$ and $\bar{\eta}$ to suffice $\bar{\xi} = \mathbf{f}(\bar{\xi}, \alpha)$ and $\bar{\eta} = \mathbf{h}(\bar{\xi}, \gamma)$ at $k, k+1 \in \mathbf{K}$, vector functions $\mathbf{f}$ and $\mathbf{h}$ shall be respectively linearized as follows:

$$\mathbf{f}(\xi, \alpha) = \mathbf{f}(\bar{\xi} + \mathbf{x}, \alpha) = \bar{\xi} + \mathbf{F}(\alpha)\mathbf{x}, \quad (15)$$

$$\mathbf{h}(\xi, \gamma) = \mathbf{h}(\bar{\xi} + \mathbf{x}, \gamma) = \bar{\eta} + \mathbf{H}(\gamma)\mathbf{x}, \quad (16)$$

where $\mathbf{K}$ denotes a set of integers associated with time when the model can be regarded as locally w.s.s. (wide sense stationary) with constant means $\bar{\xi}$ and $\bar{\eta}$. In addition, the following transition matrix $\mathbf{F} \in \mathbf{R}^{n \times n}$ and observation matrix $\mathbf{H} \in \mathbf{R}^{p \times n}$ are defined.

Applying (15) and (16), the model has been reduced in to the following linear stochastic form:

$$\mathbf{x}(k+1) = \mathbf{F}(\alpha)\mathbf{x}(k) + \mathbf{w}(k), \quad (17)$$

$$\mathbf{y}(k) = \mathbf{H}(\gamma)\mathbf{x}(k) + \mathbf{v}(k), \quad (18)$$

where $\mathbf{F}$ is identified from known music cases in the α -th key and $\mathbf{H}$ from cases using β -th instrument, as follows at $k, k+1 \in \mathbf{K}$:

$$\mathbf{F} = \left\langle \mathbf{x}(k+1)\mathbf{x}^T(k) \right\rangle \left\langle \mathbf{x}(k)\mathbf{x}^T(k) \right\rangle^{-1}, \quad (19)$$

$$\mathbf{H} = \left\langle \mathbf{y}(k)\mathbf{x}^T(k) \right\rangle \left\langle \mathbf{x}(k)\mathbf{x}^T(k) \right\rangle^{-1}. \quad (20)$$

Figure 2 shows an example of observation matrix $\mathbf{H}$ that contains considerable harmonics in case of piano where numbers on axes for state vector $\mathbf{x}$ and observed vector $\mathbf{y}$ respectively denote actually played keys on the keyboard and corresponding magnitude of the spectrum calculated by Gabor wavelet transform.

Fig.2 Observation Matrix (Piano)

4 State estimation

LMMSE (linear minimum mean square estimation) of state vector $\mathbf{x}(k)$ is obtainable as orthogonal projection $\hat{\mathbf{x}}(k) \triangleq \hat{\mathbf{E}}\{\mathbf{x}(k) | \mathbf{Y}^{k-1}\}$ under non-Gaussian assumptions. In case of known parameters α and γ stationary Kalman filter[2] recursively calculates the estimation as follows:

$$\begin{aligned} \hat{\mathbf{x}}(k+1) &= \hat{\mathbf{E}}\{\mathbf{F}(\alpha)\mathbf{x}(k) + \mathbf{w}(k) | \mathbf{Y}^{k-1} \oplus \text{span}\{\tilde{\mathbf{y}}(k)\}\} \\ &= \mathbf{F}(\alpha)\hat{\mathbf{x}}(k) + \mathbf{G}\tilde{\mathbf{y}}(k), \end{aligned} \quad (21)$$

where operation, $\oplus$, denotes orthogonal sum of spaces, and orthogonal projection has been applied to state equation (17) onto data space $\mathbf{Y}^k \triangleq \text{span}\{\mathbf{c}, \tilde{\mathbf{y}}(0), \dots, \tilde{\mathbf{y}}(k)\}$ that consists of proper constant vector, $\mathbf{c} \in \mathbf{R}^p$, and the following innovation[2], $\tilde{\mathbf{y}}(k) \in \mathbf{R}^p$ with the white noise properties:

$$\begin{aligned} \tilde{\mathbf{y}}(k) &\triangleq \mathbf{y}(k) - \hat{\mathbf{E}}\{\mathbf{y}(k) | \mathbf{Y}^{k-1}\} \\ &= \mathbf{y}(k) - \mathbf{H}(\gamma)\hat{\mathbf{x}}(k); \end{aligned} \quad (22)$$

in addition, to suffice the orthogonal property, $(\mathbf{x}(k+1) - \hat{\mathbf{x}}(k+1)) \perp \mathbf{Y}^k$ and Kalman gain, $\mathbf{G} \in \mathbf{R}^{n \times p}$ in (21) is given as follows[2]:

$$\mathbf{G} = (\mathbf{FPH}^T + \mathbf{U})(\mathbf{HPH}^T + \mathbf{V})^{-1}, \quad (23)$$

where $\mathbf{P} \in \mathbf{R}^{n \times n}$ is a solution of the following ARE (algebraic Ricatti equation):

$$\mathbf{P} = \mathbf{FPF}^T + \mathbf{W} - (\mathbf{FPH}^T + \mathbf{U})(\mathbf{HPH}^T + \mathbf{V})^{-1}(\mathbf{FPH}^T + \mathbf{U})^T. \quad (24)$$

5 Example of estimation

Figure 3 shows an example of estimated music on the top stage (a), where the actual music is on the middle stage (b) played with an electronic piano (Casio AP-420BK). According to the foregoing discussion, the acoustic signal was projected on the time-frequency plane and summed all frequency components up to obtain magnitude on the time axis shown in the bottom stage (c). Then tempo and rhythm were identified with referring autocorrelation of the magnitude.

At the next step, the system applied the state estimation in section 4 to obtain note values and actually played pitch names based on the identified tempo adaptively considering harmonics of various instruments.

After passing a threshold detector, the music on the top stage (a), the tempo, rhythm and pitch names of which are practically agree with the actual music (b), has been estimated. Some differences on duration of notes between (a) and (b) imply necessity of an adaptive threshold control that is a subject of future studies.

The first four measures of arabesque 1 (1891) by C. Debussy played by piano are shown in Fig.4 where (c) actually played music, 3-D view of (a) projection on the time-frequency plane, and (b) estimated music are the lower, the upper, and the middle stages, in order. In this example harmonics that are not corresponding to actually played keys in the part A of the upper stage (a) have been reduced to the middle stage (b) where proper threshold suggests actually played music.

Figure 5 shows four measures (the 10th to 13th) of Golliwog's Cakewalk in Children's Corner (1908) by C. Debussy played by piano where upper, middle, and lower stages are arranged in the same order of Fig.3. In this example actually played keys with difficulty to find in (a) appears in the part B of (b) by the state estimation.

(a) Estimated music: (Tempo) = 111.2

(b) Actually played music: (tempo) = 112

(c) Projection on time-frequency plane and its summation on time axis

1st-4th measure, German folk song: Froshgesang
(arranged)

Fig.3 Estimation of actually played music 1

(a) Projection on the time-frequency plane: $\{y(k)\}$

(b) Estimated music: $\{\hat{x}(k)\}$

(c) Actually played music

1st-4th measure, C. Debussy: Arabesque 1, 1891

Fig.4 Estimation of actually played music 2

(a) Projection on the time-frequency plane: $\{y(k)\}$

(b) Estimated music: $\{\hat{x}(k)\}$

(c) Actually played music

10th-13th measure, Golliwog's Cakewalk
C. Debussy: Children's Corner, 1908

Fig.5 Estimation of actually played music 3

6 Conclusions

An algorithm for music transcription systems that has the following features is proposed:

- 1) Suitable for multi-voice music including considerable harmonics which depends upon instruments
- 2) Projects the sound onto the time-frequency plane applying Gabor wavelet transform
- 3) Identifies tempo and rhythm with referring autocorrelation of magnitude of Gabor wavelet transform
- 4) Estimates note values and actually played pitch names applying adaptive state estimation

References

- [1] A. Klapuri and M. Davy, editors: Signal processing methods for music transcription, Springer, New York NY (2006)
- [2] T. Katayama: System identification (in Japanese); Asakura, Tokyo JP (2004)
- [3] R. A. Roberts and C. T. Mullis: Digital Signal Processing; Addison Wesley, Reading MA, 1987.
- [4] M. Kennedy: Oxford dictionary of music; Oxford, New York NY (1985)
- [5] Z. Xie and T. Suzuki: Wavelet and stochastic process (in Japanese); Uchida, Tokyo JP (2002)
- [6] G. Golub and C. Van Loan: Matrix computations; Johns Hopkins, Baltimore MA (1983)